SAND2006-6337
Unlimited Release
October 2006
Updated November 6, 2008

DAKOTA, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis

Version 4.2 User's Manual

Michael S. Eldred, Brian M. Adams, David M. Gay, Laura P. Swiler
Optimization and Uncertainty Quantification Department

Karen Haskell
Scientific Applications and User Support Department

William J. Bohnhoff
Radiation Transport Department

John P. Eddy
System Readiness and Sustainment Technologies Department

William E. Hart, Jean-Paul Watson
Discrete Algorithms and Math Department

Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185

Patricia D. Hough, Tammy G. Kolda
Informatics and Decision Sciences Department

Sandia National Laboratories
P.O. Box 969
Livermore, CA 94551

Abstract

The DAKOTA (Design Analysis Kit for Optimization and Terascale Applications) toolkit provides a flexible and extensible interface between simulation codes and iterative analysis methods. DAKOTA contains algorithms for optimization with gradient and nongradient-based methods; uncertainty quantification with sampling, reliability, polynomial chaos, and epistemic methods; parameter estimation with nonlinear least squares methods; and sensitivity/variance analysis with design of experiments and parameter study methods. These capabilities may be used on their own or as components within advanced strategies such as hybrid optimization, surrogate-based optimization, mixed integer nonlinear programming, or optimization under uncertainty. By employing object-oriented design to implement abstractions of the key components required for iterative systems analyses, the DAKOTA toolkit provides a flexible and extensible problem-solving environment for design and performance analysis of computational models on high performance computers.

This report serves as a user's manual for the DAKOTA software and provides capability overviews and procedures for software execution, as well as a variety of example studies.

Contents

Preface **11**

1 Introduction **13**

1.1 Motivation for DAKOTA Development 13
1.2 Capabilities of DAKOTA 14
1.3 How Does DAKOTA Work? 14
1.4 Background and Mathematical Formulations 15
1.5 Using this Manual 19

2 DAKOTA Tutorial **21**

2.1 Installation Guide 21
2.2 Rosenbrock and Textbook Test Problems 24
2.3 DAKOTA Input File Format 25
2.4 Example Problems 28
2.5 Where to Go from Here 49

3 DAKOTA Capability Overview **51**

3.1 Purpose 51
3.2 Parameter Study Methods 51
3.3 Design of Experiments 51
3.4 Uncertainty Quantification 52
3.5 Optimization 54
3.6 Additional Optimization Capabilities 55
3.7 Nonlinear Least Squares for Parameter Estimation 56
3.8 Surrogate-Based Minimization 57
3.9 Optimization Strategies 57

3.10 Surrogate Models . . . 58
3.11 Nested Models . . . 59
3.12 Parallel Computing . . . 60
3.13 Summary . . . 60

4 Parameter Study Capabilities **63**
4.1 Overview . . . 63
4.2 Vector Parameter Study . . . 64
4.3 List Parameter Study . . . 66
4.4 Centered Parameter Study . . . 67
4.5 Multidimensional Parameter Study . . . 68

5 Design of Experiments Capabilities **71**
5.1 Overview . . . 71
5.2 Design of Computer Experiments . . . 71
5.3 DDACE Background . . . 73
5.4 FSUDace Background . . . 76
5.5 PSUADE MOAT Background . . . 77
5.6 Sensitivity Analysis . . . 78

6 Uncertainty Quantification Capabilities **81**
6.1 Overview . . . 81
6.2 Sampling Methods . . . 81
6.3 Reliability Methods . . . 88
6.4 Stochastic Expansion Methods . . . 100
6.5 Epistemic Nondeterministic Methods . . . 111
6.6 Future Nondeterministic Methods . . . 114

7 Optimization Capabilities **115**
7.1 Overview . . . 115
7.2 Optimization Software Packages . . . 116
7.3 Additional Optimization Capabilities . . . 121

8 Nonlinear Least Squares Capabilities **129**
8.1 Overview . . . 129
8.2 Solution Techniques . . . 130

8.3 Examples . . . 131

9 Surrogate-Based Minimization 135

9.1 Overview . . . 135
9.2 Surrogate-Based Local Minimization . . . 135
9.3 Surrogate-Based Global Minimization . . . 145
9.4 Efficient Global Minimization . . . 146

10 Advanced Strategies 149

10.1 Overview . . . 149
10.2 Hybrid Minimization . . . 149
10.3 Multistart Local Minimization . . . 150
10.4 Pareto Optimization . . . 153
10.5 Mixed Integer Nonlinear Programming (MINLP) . . . 153

11 Models 159

11.1 Overview . . . 159
11.2 Single Models . . . 160
11.3 Recast Models . . . 160
11.4 Surrogate Models . . . 160
11.5 Nested Models . . . 168
11.6 Advanced Examples . . . 168

12 Variables 183

12.1 Overview . . . 183
12.2 Design Variables . . . 183
12.3 Uncertain Variables . . . 184
12.4 State Variables . . . 185
12.5 Mixed Variables . . . 186
12.6 DAKOTA Parameters File Data Format . . . 186
12.7 The Active Set Vector . . . 188

13 Interfaces 191

13.1 Overview . . . 191
13.2 Algebraic Mappings . . . 191
13.3 Simulation Interfaces . . . 194

13.4 Simulation Interface Components . . . 197
13.5 Simulation File Management . . . 201
13.6 Parameter to Response Mappings . . . 204

14 Responses **209**
14.1 Overview . . . 209
14.2 DAKOTA Results File Data Format . . . 210
14.3 Active Variables for Derivatives . . . 212

15 Inputs to DAKOTA **213**
15.1 Overview of Inputs . . . 213
15.2 JAGUAR . . . 213
15.3 Data Imports . . . 215

16 Output from DAKOTA **221**
16.1 Overview of Output Formats . . . 221
16.2 Standard Output . . . 221
16.3 Tabular Output Data . . . 227
16.4 Graphics Output . . . 227
16.5 Error Messages Output . . . 230

17 Advanced Simulation Code Interfaces **233**
17.1 Building an Interface to a Engineering Simulation Code . . . 233
17.2 Developing a Direct Simulation Interface . . . 241
17.3 Existing Direct Interfaces to External Simulators . . . 242

18 Parallel Computing **247**
18.1 Overview . . . 247
18.2 Single-level parallelism . . . 251
18.3 Multilevel parallelism . . . 261
18.4 Capability Summary . . . 264
18.5 Running a Parallel DAKOTA Job . . . 265
18.6 Specifying Parallelism . . . 266
18.7 Application Parallelism Use Cases . . . 274

19 DAKOTA Usage Guidelines **279**

19.1 Problem Exploration 279
19.2 Optimization Method Selection 279
19.3 UQ Method Selection 281
19.4 Parameter Study/DOE/DACE/Sampling Method Selection 283

20 Restart Capabilities and Utilities 285
20.1 Restart Management 285
20.2 The DAKOTA Restart Utility 286

21 Simulation Failure Capturing 291
21.1 Failure detection 291
21.2 Failure communication 292
21.3 Failure mitigation 292
21.4 Special values 293

22 Additional Examples 295
22.1 Textbook Example 295
22.2 Rosenbrock Example 297
22.3 Cylinder Head Example 302
22.4 Container Example 304
22.5 Log Ratio Example 307
22.6 Steel Section Example 307
22.7 Portal Frame Example 308
22.8 Short Column Example 308
22.9 Cantilever Example 310
22.10 Steel Column Example 312
22.11 Multiobjective Examples 313
22.12 Morris example 316

Preface

The DAKOTA (Design Analysis Kit for Optimization and Terascale Applications) project started in 1994 as an internal research and development activity at Sandia National Laboratories in Albuquerque, New Mexico. The original goal of this effort was to provide a common set of optimization tools for a group of engineers who were solving structural analysis and design problems. Prior to the start of the DAKOTA project, there was not a focused effort to archive the optimization methods for reuse on other projects. Thus, for each new project the engineers found themselves custom building new interfaces between the engineering analysis software and the optimization software. This was a particular burden when attempts were made to use parallel computing resources, where each project required the development of a unique master program that coordinated concurrent simulations on a network of workstations or a parallel computer. The initial DAKOTA toolkit provided the engineering and analysis community at Sandia Labs with access to a variety of different optimization methods and algorithms, with much of the complexity of the optimization software interfaces hidden from the user. Thus, the engineers were easily able to switch between optimization software packages simply by changing a few lines in the DAKOTA input file. In addition to applications in structural analysis, DAKOTA has been applied to applications in computational fluid dynamics, nonlinear dynamics, shock physics, heat transfer, and many others.

DAKOTA has grown significantly beyond its original focus as a toolkit of optimization methods. In addition to having many state-of-the-art optimization methods, DAKOTA now includes methods for global sensitivity and variance analysis, parameter estimation, and uncertainty quantification, as well as meta-level strategies for surrogate-based optimization, mixed-integer nonlinear programming, hybrid optimization, and optimization under uncertainty. Underlying all of these algorithms is support for parallel computation; ranging from the level of a desktop multiprocessor computer up to massively parallel computers found at national laboratories and supercomputer centers.

This document corresponds to DAKOTA Version 4.2. Release notes for this release, past releases, and current developmental releases are available from http://www.cs.sandia.gov/dakota/licensing/release_notes.html. Starting with Version 3.0, DAKOTA has been publicly released as open source under a GNU General Public License and is available for free download world-wide. See http://www.gnu.org/licenses/gpl.html for more information on the GPL software use agreement. The objective of this public release is to facilitate research and software collaborations among the developers of DAKOTA at Sandia National Laboratories and other institutions, including academic, governmental, and corporate entities. For more information on the objectives of the open source release and how to contribute, refer to the DAKOTA FAQ at http://www.cs.sandia.gov/dakota/faq.html.

The DAKOTA leadership team consists of Mike Eldred (project lead), Brian Adams (product manager), Karen Haskell (support manager), and Jim Stewart (business manager). DAKOTA development team members include Bill Bohnhoff, John Eddy, David Gay, Bill Hart, Patty Hough, Laura Swiler, and Jean-Paul Watson. In addition, contributors to the APPSPACK, COLINY, OPT++, DDACE, APPS, FSUDace, and Surfpack libraries used by DAKOTA include Tammy Kolda from Sandia, John Burkardt from Virginia Tech, Prof. Jonathan Eckstein from Rutgers University, Prof. Roger Ghanem from Johns Hopkins University, Mark Richards from the University of

Illinois, Prof. Virginia Torczon from the College of William and Mary, and Prof. Steve Wojtkiewicz from the University of Minnesota.

Contact Information:

Michael S. Eldred, DAKOTA Project Lead
Sandia National Laboratories
P.O. Box 5800, Mail Stop 1318
Albuquerque, NM 87185-1318

email: dakota-developers@development.sandia.gov
web: `http://www.cs.sandia.gov/dakota`

Chapter 1

Introduction

1.1 Motivation for DAKOTA Development

Computational models are commonly used in engineering design activities for simulating complex physical systems in disciplines such as fluid mechanics, structural dynamics, heat transfer, nonlinear structural mechanics, shock physics, and many others. These simulators can be an enormous aid to engineers who want to develop an understanding and/or predictive capability for the complex behaviors that are often observed in the respective physical systems. Often, these simulators are employed as virtual prototypes, where a set of predefined system parameters, such as size or location dimensions and material properties, are adjusted to improve or optimize the performance of a particular system, as defined by one or more system performance objectives. Optimization of the virtual prototype then requires execution of the simulator, evaluation of the performance objective(s), and adjustment of the system parameters in an iterative and directed way, such that an improved or optimal solution is obtained for the simulation as measured by the performance objective(s). System performance objectives can be formulated, for example, to minimize weight, cost, or defects; to limit a critical temperature, stress, or vibration response; or to maximize performance, reliability, throughput, agility, or design robustness. In addition, one would often like to design computer experiments, run parameter studies, or perform uncertainty quantification. These methods allow one to understand how the system performance changes as a design variable or an uncertain input changes. Sampling strategies are often used in uncertainty quantification to calculate a distribution on system performance measures, and to understand which uncertain inputs are the biggest contributors to the variance of the outputs.

One of the primary motivations for the development of DAKOTA (Design Analysis Kit for Optimization and Terascale Applications) has been to provide engineers with a systematic and rapid means of obtaining improved or optimal designs using their simulator-based models. Making this capability available to engineers generally leads to better designs and improved system performance at earlier stages of the design phase, and eliminates some of the dependence on real prototypes and testing, thereby shortening the design cycle and reducing overall product development costs. In addition to providing this environment for answering systems performance questions, the DAKOTA toolkit also provides an extensible platform for the research and rapid prototyping of customized methods and strategies [29].

1.2 Capabilities of DAKOTA

The DAKOTA toolkit provides a flexible, extensible interface between your simulation code and a variety of iterative methods and strategies. While DAKOTA was originally conceived as an easy-to-use interface between simulation codes and optimization algorithms, recent versions have been expanded to interface with other types of iterative analysis methods such as uncertainty quantification with nondeterministic propagation methods, parameter estimation with nonlinear least squares solution methods, and sensitivity/variance analysis with general-purpose design of experiments and parameter study capabilities. These capabilities may be used on their own or as building blocks within more sophisticated strategies such as hybrid optimization, surrogate-based optimization, mixed integer nonlinear programming, or optimization under uncertainty.

Thus, one of the primary advantages that DAKOTA has to offer is that access to a broad range of iterative capabilities can be obtained through a single, relatively simple interface between DAKOTA and your simulator. Should you want to try a different type of iterative method or strategy with your simulator, it is only necessary to change a few commands in the DAKOTA input and start a new analysis. The need to learn a completely different style of command syntax and the need to construct a new interface each time you want to use a new algorithm are eliminated.

1.3 How Does DAKOTA Work?

Figure 1.1 depicts the loosely-coupled, or "black-box," relationship between DAKOTA and the simulation code(s). This loose coupling is the simplest approach and is the one that most DAKOTA users will employ. Data is exchanged between DAKOTA and the simulation code by reading and writing short data files, and DAKOTA does not require access to the source code of the user's simulation software. DAKOTA is executed using commands that the user supplies in an input file (not shown in Figure 1.1) which specify the type of analysis to be performed (e.g., parameter study, optimization, uncertainty estimation, etc.), along with the file names associated with the user's simulation code. During its operation, DAKOTA automatically executes the user's simulation code by creating a separate process that is external to DAKOTA.

The solid lines in Figure 1.1 denote file input/output (I/O) operations that are part of DAKOTA or the user's simulation code. The dotted lines indicate the passing of information that must be handled by the user. As DAKOTA is running, it writes out a parameters file that contains the values of the current variables. DAKOTA then starts the user's simulation code (or, often, a short driver script), and when the simulation has completed, DAKOTA reads in the response data from a results file. This process is repeated until all of the simulation code runs required by the iterative study have been completed.

In some cases it is advantageous to have a close coupling between DAKOTA and the user's simulation code. This close coupling is an advanced feature of DAKOTA and is accomplished through either a direct interface or a SAND (simultaneous analysis and design) interface. For the direct interface, the user's simulation code is modified to behave as a function or subroutine under DAKOTA. This interface can be considered to be "semi-intrusive" in that it requires relatively minor modifications to the simulation code. Its major advantage is the elimination of the overhead resulting from file I/O and process creation. It can also be a useful tool for parallel processing, by encapsulating everything within a single executable. A SAND interface approach is "fully intrusive" in that it requires further modifications to the simulation code so that an optimizer has access to the internal residual vector and Jacobian matrices computed by the simulation code. In a SAND approach, both the optimization method and a nonlinear simulation code are converged simultaneously. While this approach can greatly reduce the computational expense of optimization, considerable software development effort must be expended to achieve this intrusive coupling between SAND optimization methods and the simulation code. SAND may be supported in future DAKOTA releases.

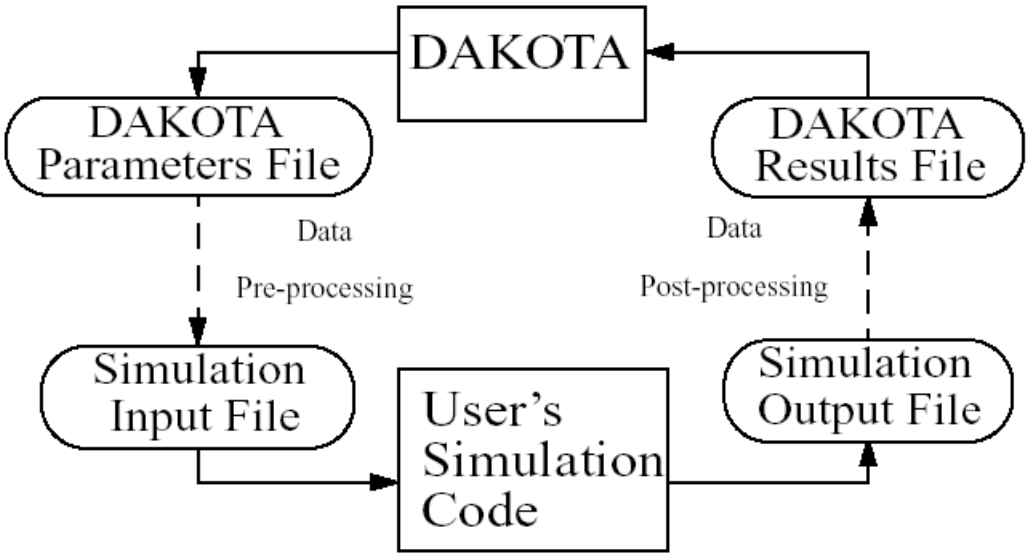


Figure 1.1: The loosely-coupled or "black-box" interface between DAKOTA and a user-supplied simulation code.

1.4 Background and Mathematical Formulations

This section provides a basic introduction to the mathematical formulation of optimization, nonlinear least squares, sensitivity analysis, design of experiments, and uncertainty quantification problems. The primary goal of this section is to introduce terms relating to these topics, and is not intended to be a description of theory or numerical algorithms. There are numerous sources of information on these topics ([5], [56], [69], [70], [98], [131]) and the interested reader is advised to consult one or more of these texts.

1.4.1 Optimization

A general optimization problem is formulated as follows:

$$
\begin{aligned}
\text{minimize:} \quad & f(\mathbf{x}) \\
& \mathbf{x} \in \Re^n \\
\text{subject to:} \quad & \mathbf{g}_L \leq \mathbf{g}(\mathbf{x}) \leq \mathbf{g}_U \\
& \mathbf{h}(\mathbf{x}) = \mathbf{h}_t \\
& \mathbf{a}_L \leq \mathbf{A}_i\mathbf{x} \leq \mathbf{a}_U \\
& \mathbf{A}_e\mathbf{x} = \mathbf{a}_t \\
& \mathbf{x}_L \leq \mathbf{x} \leq \mathbf{x}_U
\end{aligned}
\tag{1.1}
$$

where vector and matrix terms are marked in bold typeface. In this formulation, $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ is an n-dimensional vector of real-valued *design variables* or *design parameters*. The n-dimensional vectors, $\mathbf{x}_L$ and $\mathbf{x}_U$, are the lower and upper bounds, respectively, on the design parameters. These bounds define the allowable values for the elements of $\mathbf{x}$, and the set of all allowable values is termed the *design space* or the *parameter space*. A *design point* or a *sample point* is a particular set of values within the parameter space.

The optimization goal is to minimize the *objective function*, $f(\mathbf{x})$, while satisfying the constraints. Constraints can be categorized as either linear or nonlinear and as either inequality or equality. The *nonlinear inequality constraints*, $\mathbf{g}(\mathbf{x})$, are "2-sided," in that they have both lower and upper bounds, $\mathbf{g}_L$ and $\mathbf{g}_U$, respectively. The *nonlinear equality constraints*, $\mathbf{h}(\mathbf{x})$, have target values specified by $\mathbf{h}_t$. The linear inequality constraints create a linear system $\mathbf{A}_i\mathbf{x}$, where $\mathbf{A}_i$ is the coefficient matrix for the linear system. These constraints are also 2-sided as they have lower and upper bounds, $\mathbf{a}_L$ and $\mathbf{a}_U$, respectively. The linear equality constraints create a linear system $\mathbf{A}_e\mathbf{x}$, where $\mathbf{A}_e$ is the coefficient matrix for the linear system and $\mathbf{a}_t$ are the target values. The constraints partition the parameter space into feasible and infeasible regions. A design point is said to be *feasible* if and only if it satisfies all of the constraints. Correspondingly, a design point is said to be *infeasible* if it violates one or more of the constraints.

Many different methods exist to solve the optimization problem given by Equation 1.1, all of which iterate on $\mathbf{x}$ in some manner. That is, an initial value for each parameter in $\mathbf{x}$ is chosen, the *response quantities*, $f(\mathbf{x})$, $\mathbf{g}(\mathbf{x})$, $\mathbf{h}(\mathbf{x})$, are computed, and some algorithm is applied to generate a new $\mathbf{x}$ that will either reduce the objective function, reduce the amount of infeasibility, or both. To facilitate a general presentation of these methods, three criteria will be used in the following discussion to differentiate them: optimization problem type, search goal, and search method.

The optimization problem type can be characterized both by the types of constraints present in the problem and by the linearity or nonlinearity of the objective and constraint functions. For constraint categorization, a hierarchy of complexity exists for optimization algorithms, ranging from simple bound constraints, through linear constraints, to full nonlinear constraints. By the nature of this increasing complexity, optimization problem categorizations are inclusive of all constraint types up to a particular level of complexity. That is, an *unconstrained problem* has no constraints, a *bound-constrained problem* has only lower and upper bounds on the design parameters, a *linearly-constrained problem* has both linear and bound constraints, and a *nonlinearly-constrained problem* may contain the full range of nonlinear, linear, and bound constraints. If all of the linear and nonlinear constraints are equality constraints, then this is referred to as an *equality-constrained problem*, and if all of the linear and nonlinear constraints are inequality constraints, then this is referred to as an *inequality-constrained problem*. Further categorizations can be made based on the linearity of the objective and constraint functions. A problem where the objective function and all constraints are linear is called a *linear programming (LP) problem*. These types of problems commonly arise in scheduling, logistics, and resource allocation applications. Likewise, a problem where at least some of the objective and constraint functions are nonlinear is called a *nonlinear programming (NLP) problem*. These NLP problems predominate in engineering applications and are the primary focus of DAKOTA.

The search goal refers to the ultimate objective of the optimization algorithm, i.e., either global or local optimization. In *global optimization*, the goal is to find the design point that gives the lowest feasible objective function value over the entire parameter space. In contrast, in *local optimization*, the goal is to find a design point that is lowest relative to a "nearby" region of the parameter space. In almost all cases, global optimization will be more computationally expensive than local optimization. Thus, the user must choose an optimization algorithm with an appropriate search scope that best fits the problem goals and the computational budget.

The search method refers to the approach taken in the optimization algorithm to locate a new design point that has a lower objective function or is more feasible than the current design point. The search method can be classified as either *gradient-based* or *nongradient-based*. In a gradient-based algorithm, gradients of the response functions are computed to find the direction of improvement. Gradient-based optimization is the search method that underlies many efficient local optimization methods. However, a drawback to this approach is that gradients can be computationally expensive, inaccurate, or even nonexistent. In such situations, nongradient-based search methods may be useful. There are numerous approaches to nongradient-based optimization. Some of the more well known of these include pattern search methods (nongradient-based local techniques) and genetic algorithms (nongradient-based global techniques). Because of the computational cost of running simulation models, surrogate-based optimization (SBO) methods are often used to reduce the number of actual simulation runs. In

SBO, a surrogate or approximate model is constructed based on a limited number of simulation runs. The optimization is then performed on the surrogate model. DAKOTA has an extensive framework for managing a variety of local, multipoint, global, and hierarchical surrogates for use in optimization.

The overview of optimization methods presented above underscores that there is no single optimization method or algorithm that works best for all types of optimization problems. Chapter 19 provides some guidelines on choosing which DAKOTA optimization algorithm is best matched to your specific optimization problem.

1.4.2 Nonlinear Least Squares for Parameter Estimation

Specialized least squares solution algorithms can exploit the structure of a sum of the squares objective function for problems of the form:

$$\begin{aligned}
\text{minimize:} \quad & f(\mathbf{x}) = \sum_{i=1}^{n} [T_i(\mathbf{x})]^2 \\
& \mathbf{x} \in \Re^n \\
\text{subject to:} \quad & \mathbf{g}_L \leq \mathbf{g}(\mathbf{x}) \leq \mathbf{g}_U \\
& \mathbf{h}(\mathbf{x}) = \mathbf{h}_t \\
& \mathbf{a}_L \leq \mathbf{A}_i \mathbf{x} \leq \mathbf{a}_U \\
& \mathbf{A}_e \mathbf{x} = \mathbf{a}_t \\
& \mathbf{x}_L \leq \mathbf{x} \leq \mathbf{x}_U
\end{aligned} \tag{1.2}$$

where $f(\mathbf{x})$ is the objective function to be minimized and $T_i(\mathbf{x})$ is the $\mathrm{i^{th}}$ least squares term. The bound, linear, and nonlinear constraints are the same as described previously for (1.1). Specialized least squares algorithms are generally based on the Gauss-Newton approximation. When differentiating $f(\mathbf{x})$ twice, terms of $T_i(\mathbf{x})T_i''(\mathbf{x})$ and $[T_i'(\mathbf{x})]^2$ result. By assuming that the former term tends toward zero near the solution since $T_i(\mathbf{x})$ tends toward zero, then the Hessian matrix of second derivatives of $f(\mathbf{x})$ can be approximated using only first derivatives of $T_i(\mathbf{x})$. As a result, Gauss-Newton algorithms exhibit quadratic convergence rates near the solution for those cases when the Hessian approximation is accurate, i.e. the residuals tend towards zero at the solution. Thus, by exploiting the structure of the problem, the second order convergence characteristics of a full Newton algorithm can be obtained using only first order information from the least squares terms.

A common example for $T_i(\mathbf{x})$ might be the difference between experimental data and model predictions for a response quantity at a particular location and/or time step, i.e.:

$$T_i(\mathbf{x}) = R_i(\mathbf{x}) - \overline{R_i} \tag{1.3}$$

where $R_i(\mathbf{x})$ is the response quantity predicted by the model and $\overline{R_i}$ is the corresponding experimental data. In this case, $\mathbf{x}$ would have the meaning of model parameters which are not precisely known and are being calibrated to match available data. This class of problem is known by the terms parameter estimation, system identification, model calibration, test/analysis reconciliation, etc.

1.4.3 Sensitivity Analysis and Parameter Studies

In many engineering design applications, sensitivity analysis techniques and parameter study methods are useful in identifying which of the design parameters have the most influence on the response quantities. This information is

helpful prior to an optimization study as it can be used to remove design parameters that do not strongly influence the responses. In addition, these techniques can provide assessments as to the behavior of the response functions (smooth or nonsmooth, unimodal or multimodal) which can be invaluable in algorithm selection for optimization, uncertainty quantification, and related methods. In a post-optimization role, sensitivity information is useful is determining whether or not the response functions are robust with respect to small changes in the optimum design point.

In some instances, the term sensitivity analysis is used in a local sense to denote the computation of response derivatives at a point. These derivatives are then used in a simple analysis to make design decisions. DAKOTA supports this type of study through numerical finite-differencing or retrieval of analytic gradients computed within the analysis code. The desired gradient data is specified in the responses section of the DAKOTA input file and the collection of this data at a single point is accomplished through a parameter study method with no steps. This approach to sensitivity analysis should be distinguished from the activity of augmenting analysis codes to internally compute derivatives using techniques such as direct or adjoint differentiation, automatic differentiation (e.g., ADIFOR), or complex step modifications. These sensitivity augmentation activities are completely separate from DAKOTA and are outside the scope of this manual. However, once completed, DAKOTA can utilize these analytic gradients to perform optimization, uncertainty quantification, and related studies more reliably and efficiently.

In other instances, the term sensitivity analysis is used in a more global sense to denote the investigation of variability in the response functions. DAKOTA supports this type of study through computation of response data sets (typically function values only, but all data sets are supported) at a series of points in the parameter space. The series of points is defined using either a vector, list, centered, or multidimensional parameter study method. For example, a set of closely-spaced points in a vector parameter study could be used to assess the smoothness of the response functions in order to select a finite difference step size, and a set of more widely-spaced points in a centered or multidimensional parameter study could be used to determine whether the response function variation is likely to be unimodal or multimodal. See Chapter 4 for additional information on these methods. These more global approaches to sensitivity analysis can be used to obtain trend data even in situations when gradients are unavailable or unreliable, and they are conceptually similar to the design of experiments methods and sampling approaches to uncertainty quantification described in the following sections.

1.4.4 Design of Experiments

Classical design of experiments (DoE) methods and the more modern design and analysis of computer experiments (DACE) methods are both techniques which seek to extract as much trend data from a parameter space as possible using a limited number of sample points. Classical DoE techniques arose from technical disciplines that assumed some randomness and nonrepeatability in field experiments (e.g., agricultural yield, experimental chemistry). DoE approaches such as central composite design, Box-Behnken design, and full and fractional factorial design generally put sample points at the extremes of the parameter space, since these designs offer more reliable trend extraction in the presence of nonrepeatability. DACE methods are distinguished from DoE methods in that the nonrepeatability component can be omitted since computer simulations are involved. In these cases, space filling designs such as orthogonal array sampling and latin hypercube sampling are more commonly employed in order to accurately extract trend information. Quasi-Monte Carlo sampling techniques which are constructed to fill the unit hypercube with good uniformity of coverage can also be used for DACE.

DAKOTA supports both DoE and DACE techniques. In common usage, only parameter bounds are used in selecting the samples within the parameter space. Thus, DoE and DACE can be viewed as special cases of the more general probabilistic sampling for uncertainty quantification (see following section), in which the DoE/DACE parameters are treated as having uniform probability distributions. The DoE/DACE techniques are commonly used for investigation of global response trends, identification of significant parameters (e.g., main effects), and

as data generation methods for building response surface approximations.

1.4.5 Uncertainty Quantification

Uncertainty quantification (UQ) is the process of determining the effect of input uncertainties on response metrics of interest. These input uncertainties may be characterized as either aleatory uncertainties, which are irreducible variabilities inherent in nature, or epistemic uncertainties, which are reducible uncertainties resulting from a lack of knowledge. Since sufficient data is generally available for aleatory uncertainties, probabilistic methods are commonly used for computing response distribution statistics based on input probability distribution specifications. Conversely, for epistemic uncertainties, data is generally sparse, making the use of probability theory questionable and leading to nonprobabilistic methods based on interval specifications.

UQ is related to sensitivity analysis in that the common goal is to gain an understanding of how variations in the parameters affect the response functions of the engineering design problem. However, for UQ, some or all of the components of the parameter vector, $\mathbf{x}$, are considered to be uncertain as specified by particular probability distributions (e.g., normal, exponential, extreme value). By assigning specific distributional structure to the inputs, distributional structure for the outputs (i.e., response statistics) can be inferred.

Current methods for modeling aleatory uncertainty include sampling methods, local and global reliability methods, and polynomial chaos expansions (PCE). Current methods for modeling epistemic and mixed aleatory/epistemic uncertainties include second-order probability and Dempster-Shafer theory of evidence. The sampling, reliability, stochastic expansion, and Dempster-Shafer UQ approaches are described in more detail in Chapter 6. Second-order probability is described in Section 11.6.1.

1.5 Using this Manual

The previous sections in this chapter have provided a brief overview of the capabilities in DAKOTA, and have introduced some of the common terms that are used in the fields of optimization, parameter estimation, sensitivity analysis, design of experiments, and uncertainty quantification. The DAKOTA user that is new to these techniques is advised to consult the references cited earlier in this chapter to obtain more detailed descriptions of methods and algorithms in these disciplines.

Chapter 2 provides information on how to obtain, install, and use DAKOTA. In addition, example problems are presented in this chapter to demonstrate some of DAKOTA's capabilities for parameter studies, optimization, and UQ. Chapter 3 provides a brief overview of all of the different software packages and capabilities in DAKOTA. Chapter 4 through Chapter 8 provide details on the iterative algorithms supported in DAKOTA, and Chapter 10 describes DAKOTA's advanced optimization strategies. Chapter 11 through Chapter 14 provide information on model components which are involved in parameter to response mappings and Chapters 15 and 16 describe the inputs to and outputs from DAKOTA. Chapter 17 provides information on interfacing DAKOTA with engineering simulation codes, Chapter 18 covers DAKOTA's parallel computing capabilities, and Chapter 19 provides some usage guidelines for selecting DAKOTA algorithms. Finally, Chapter 20 through Chapter 22 describe restart utilities, failure capturing facilities, and additional test problems, respectively.

Chapter 2

DAKOTA Tutorial

2.1 Installation Guide

DAKOTA can be compiled for most common computer systems that run Unix and Linux operating systems. The computers and operating systems actively supported by the DAKOTA project include:

- Intel/AMD Redhat Enterprise Linux 4 and 5 (RHEL4, RHEL5) with gcc and Intel compilers
- Sun Solaris 5.10 with CC compilers
- IBM AIX 5.3 with xlC compilers
- Mac OS X 10.4 with gcc compilers
- ASC Red Storm with PGI compilers

In addition, partial support is provided for PC Windows (via Cygwin), using gcc compilers. Additional details are provided in the file `Dakota/README` in the distribution (see the following section for download instructions). Further platform/operating system-specific guidance can be found in Dakota/examples/platforms included with DAKOTA.

For answers to common questions and solutions to common problems in downloading, building, installing, or running DAKOTA, refer to `http://www.cs.sandia.gov/DAKOTA/faq.html` for additional information.

2.1.1 How to Obtain DAKOTA - External to Sandia Labs

If you are outside of Sandia National Laboratories, the DAKOTA binary executable files and source code files are available through the download link in the following web site:

`http://www.cs.sandia.gov/DAKOTA/software.html`

To receive the binary or source code files, you are asked to fill out a short online registration form. The information you provide is used by the DAKOTA development team to collect software usage metrics; the form also lets you sign up for update announcements.

If you are a new DAKOTA user and are using one of the supported or partially supported platforms, we suggest that you download the relevant binary executable distribution rather than the source code distribution. The compilation process can be somewhat involved, and it will be easier for you to first gain an understanding of DAKOTA by running the example problems that are provided with the binary distributions. For more experienced users, DAKOTA can be customized with additional packages and ported to other computer platforms when building from the source code.

2.1.2 How to Obtain DAKOTA - Internal to Sandia Labs

DAKOTA binary executable files have been compiled and distributed to the ESHPC LAN and common compute servers at Sandia, Los Alamos, and Lawrence Livermore. Common full-path names for the executable include `/usr/local/bin/dakota` and `/projects/dakota/bin/<system>/dakota`, where "`<system>`" is `osf`, `irix`, or other. To see if DAKOTA is available on your computer system and accessible in your Unix environment path settings, type the command `which dakota` at the Unix prompt. If the DAKOTA executable file is in your path, its location will be echoed to the terminal. If the DAKOTA executable file is available on your system but not in your path, then you will need to locate it and add its directory to your path (the Unix `whereis` and `find` commands can be useful for locating the executable).

If DAKOTA is not available on your system, the current preferred options are to either get an account on one of the common compute servers where DAKOTA is maintained, or if this is not practical, to contact one of the DAKOTA team members so that we can provide you with DAKOTA executable files that are as complete as possible (i.e., that include Sandia-specific and site-licensed software that is not yet publicly available). Alternatively, you can follow the instructions given in the previous section to obtain the public version of the DAKOTA binary and/or source codes files. In the future, a download facility on Sandia's internal restricted network may be added to simplify internal distributions.

2.1.3 Installing DAKOTA - Binary Executable Files

Once you have downloaded a binary distribution from the web site listed above, you will have a Unix tar file that has a name similar to `Dakota_4_x.OSversion.tar.gz`.

Use the Unix utility `gunzip` to uncompress the tar file and the Unix `tar` utility to extract the files from the archive by executing the following commands:

```
gunzip Dakota_4_x.OSversion.tar.gz
tar -xvf Dakota_4_x.OSversion.tar
```

Slightly faster and less demanding of disk space is to invoke

```
gzip -dc Dakota_4_x.OSversion.tar.gz | tar xf -
```

The tar utility will create a subdirectory named `Dakota` in which the DAKOTA executables and example files will be stored. The executables are in `Dakota/bin`, and the example problems are `Dakota/test` and in subdirectories of `Dakota/examples`. See file `Dakota/examples/README` for more details about these subdirectories.

2.1.4 Installing DAKOTA - Source Code Files

Following the download, decompression, and extraction of the file `Dakota_4_x.src.tar.gz`, the basic steps follow the standard GNU distribution process of:

```
./configure
make
```

to construct Makefiles and build the system, respectively. Additionally, one can

```
make check
make install
```

to exercise regression and unit tests using the new executable and install the executable in a desired location, respectively. Detailed instructions for building DAKOTA are given in the file `Dakota/INSTALL`.

2.1.5 Running DAKOTA

The DAKOTA executable file is named dakota. If this command is entered at the Unix prompt without any arguments, the following usage message appears:

```
usage: dakota [options and <args>]
        -help (Print this summary)
        -version (Print DAKOTA version number)
        -check (Perform input checks)
        -input <$val> (REQUIRED DAKOTA input file $val)
        -output <$val> (Redirect DAKOTA standard output to file $val)
        -error <$val> (Redirect DAKOTA standard error to file $val)
        -parser <$val> (Parsing technology: nidr[strict][:dumpfile])
        -read_restart <$val> (Read an existing DAKOTA restart file $val)
        -stop_restart <$val> (Stop restart file processing at evaluation $val)
        -write_restart <$val> (Write a new DAKOTA restart file $val)
```

Of these available command line inputs, only the "`-input`" option is required, and "`-input`" can be omitted if the input file name is the final item on the command line; all other command-line inputs are optional. The "`-help`" option prints the usage message above. The "`-version`" option prints the version number of the executable. The "`-check`" option invokes a dry-run mode in which the input file is processed and checked for errors, but the study is not performed. The "`-input`" option provides the name of the DAKOTA input file. The "`-output`" and "`-error`" options provide file names for redirection of the DAKOTA standard output (stdout) and standard error (stderr), respectively. The "`-parser`" input is for debugging and will not be further described here. The "`-read_restart`" and "`-write_restart`" command line inputs provide the names of restart databases to read from and write to, respectively. The "`-stop_restart`" command line input limits the number of function evaluations read from the restart database (the default is all the evaluations) for those cases in which some evaluations were erroneous or corrupted. Restart management is an important technique for retaining data from expensive engineering applications. This is an advanced topic that is discussed in detail in Chapter 19. Note that these command line inputs can be abbreviated so long as the abbreviation is unique (the current set of command line options do not have any possibility for abbreviation ambiguity). That is, "`-h`", "`-v`", "`-c`", "`-i`", "`-o`", "`-e`", "`-r`", "`-s`", and "`-w`" are commonly used in place of the longer forms of the command line inputs.

To run DAKOTA with a particular input file, the following syntax can be used:

```
dakota -i dakota.in
```

or more simply

```
dakota dakota.in
```

This will echo the standard output (stdout) and standard error (stderr) messages to the terminal. To redirect stdout and stderr to separate files, the `-o` and `-e` command line options may be used:

```
dakota -i dakota.in -o dakota.out -e dakota.err
```

or

```
dakota -o dakota.out -e dakota.err dakota.in
```

Alternatively, any of a variety of Unix redirection variants can be used. The simplest of these redirects stdout to another file:

```
dakota dakota.in > dakota.out
```

To append to a file rather than overwrite it, "`>>`" is used in place of "`>`". The syntax to redirect stderr as well as stdout to the same file depends on the shell you are using. With csh, simply append "`&`" with no embedded space, i.e., "`>&`" or "`>>&`". With the Bourne shell (sh or bash) use "`>dakota.out 2>&1`" or "`>>dakota.out 2>&1`". With csh, if you have the noclobber environment variable set but wish either to overwrite an existing output file or to append to a file that does not yet exist, append "`!`" to the redirection operators (with no intervening spaces), i.e., "`>!`", "`>&!`", "`>>!`", or "`>>&!`".

To run the dakota process in the background, append an ampersand symbol (&) to the command with an embedded space, e.g.,

```
dakota dakota.in > dakota.out &
```

Refer to [4] for more information on Unix redirection and background commands.

2.2 Rosenbrock and Textbook Test Problems

Many of the example problems in this chapter use the Rosenbrock function [114] (also described in [56], among other places), which has the form:

$$f(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 \tag{2.1}$$

A three-dimensional plot of this function is shown in Figure 2.1(a), where both x_1 and x_2 range in value from -2 to 2. Figure 2.1(b) shows a contour plot for Rosenbrock's function. An optimization problem using Rosenbrock's function is formulated as follows:

$$\begin{array}{rl} \texttt{minimize} & f(x_1, x_2) \\ & \mathbf{x} \in \Re^2 \\ \texttt{subject to} & -2 \leq x_1 \leq 2 \\ & -2 \leq x_2 \leq 2 \end{array} \tag{2.2}$$

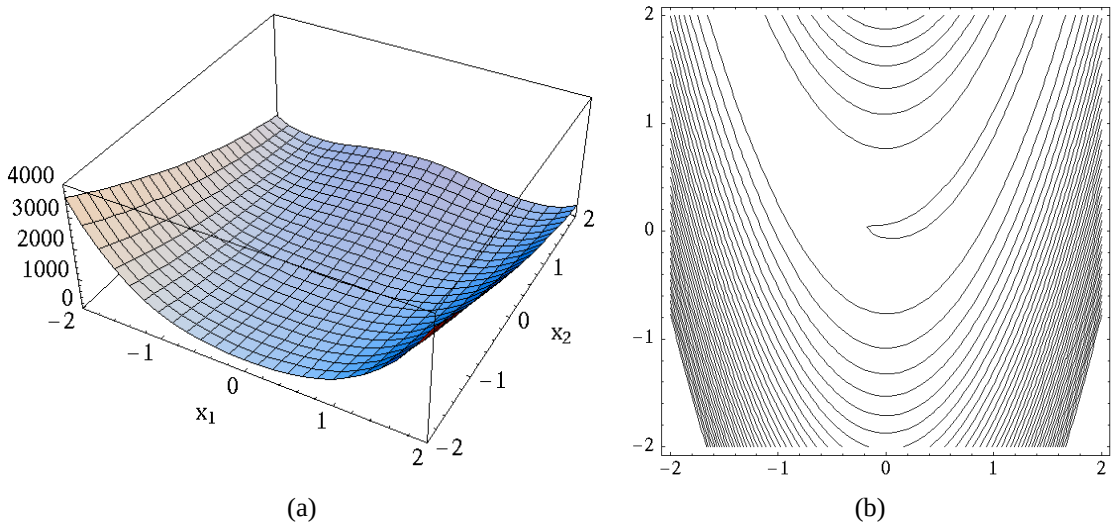


Figure 2.1: Rosenbrock's function: (a) 3-D plot and (b) contours with x_1 on the bottom axis.

Note that there are no linear or nonlinear constraints in this formulation, so this is a bound constrained optimization problem. The unique solution to this problem lies at the point $(x_1, x_2) = (1, 1)$, where the function value is zero.

The two-variable version of the "textbook" example problem provides a nonlinearly constrained optimization test case. It is formulated as:

$$\begin{aligned} \texttt{minimize} \quad & f = (x_1 - 1)^4 + (x_2 - 1)^4 \\ \texttt{subject to} \quad & g_1 = x_1^2 - \frac{x_2}{2} \leq 0 \\ & g_2 = x_2^2 - \frac{x_1}{2} \leq 0 \\ & 0.5 \leq x_1 \leq 5.8 \\ & -2.9 \leq x_2 \leq 2.9 \end{aligned} \tag{2.3}$$

Contours of this example problem are illustrated in Figure 2.2(a), with a close-up view of the feasible region given in Figure 2.2(b).

For the textbook example problem, the unconstrained minimum occurs at $(x_1, x_2) = (1, 1)$. However, the inclusion of the constraints moves the minimum to $(x_1, x_2) = (0.5, 0.5)$.

Several other example problems are available. See Chapter 22 for a description of these example problems as well as further discussion of the Rosenbrock and textbook example problems.

2.3 DAKOTA Input File Format

All of the DAKOTA input files for the simple example problems presented here are included in the distribution tar files within directory `Dakota/examples/tutorial`. A simple DAKOTA input file (that is named `dakota_rosenbrock_2d.in`) for a two-dimensional parameter study on Rosenbrock's function is shown in Figure 2.3. This input file will be used to describe the basic format and syntax used in all DAKOTA input files.

There are six specification blocks that may appear in DAKOTA input files. These are identified in the input

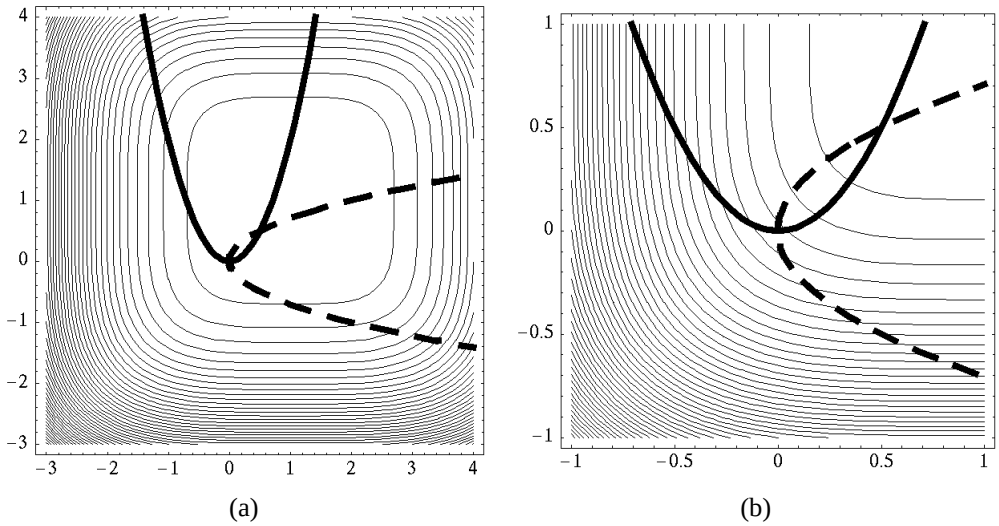


Figure 2.2: Contours of the textbook problem (a) on the $[-3, 4] \times [-3, 4]$ domain and (b) zoomed into an area containing the constrained optimum point $(x_1, x_2) = (0.5, 0.5)$. The feasible region lies at the intersection of the two constraints g_1 (solid) and g_2 (dashed).

file using the following keywords: variables, interface, responses, model, method, and strategy. These keyword blocks can appear in any order in a DAKOTA input file. At least one *variables*, *interface*, *responses*, and *method* specification must appear, and no more than one *strategy* specification should appear. In Figure 2.3, one of each of the keyword blocks is used. Additional syntax features include use of the # symbol to indicate a comment, use of single or double quotes for string inputs (e.g., `'x1'`), the use of commas and/or white space for separation of specifications, and the use of "=" symbols to optionally enhance the association of supplied data. See the DAKOTA Reference Manual [30] for additional details on this input file syntax.

The first section of the input file shown in Figure 2.3 is the *strategy* section. This keyword section is used to specify some of DAKOTA's advanced meta-procedures such as hybrid optimization, surrogate-based optimization, multi-start optimization, and Pareto optimization. See Chapter 10 for more information on these meta-procedures. The *strategy* section also contains the settings for DAKOTA's graphical output (via the `graphics` flag) and the tabular data output (via the `tabular_graphics_data` keyword).

The *method* section of the input file specifies the iterative technique that DAKOTA will employ, such as a parameter study, optimization method, data sampling technique, etc. The keyword `multidim_parameter_study` in Figure 2.3 calls for a multidimensional parameter study, while the keyword `partitions` specifies the number of intervals per variable. In this case, there will be eight intervals (nine data points) evaluated between the lower and upper bounds of both variables (bounds provided subsequently in the *variables* section), for a total of 81 response function evaluations.

The *model* section of the input file specifies the model that DAKOTA will use. A model provides the logical unit for determining how a set of variables is mapped into a set of responses in support of an iterative method. The model allows one to specify a single interface, or to manage more sophisticated mappings involving surrogates or nested iteration. For example, one might want to use an approximate model for optimization or uncertainty quantification, due to the lower computational cost. The `model` keyword allows one to specify if the iterator will be operating on a data fit surrogate (such as a polynomial regression, neural net, etc.), a hierarchical surrogate (which uses the corrected results of a lower fidelity simulation model as an approximation to a higher fidelity simulation), or a nested model. See Chapter 11 for additional model specification details. If these advanced

```
## DAKOTA INPUT FILE - dakota_rosenbrock_2d.in


strategy,
        single_method
          graphics,tabular_graphics_data

method,
        multidim_parameter_study
          partitions = 8 8

model,
        single

variables,
        continuous_design = 2
          lower_bounds    -2.0     -2.0
          upper_bounds     2.0      2.0
          descriptors      'x1'     "x2"

interface,
        direct
          analysis_driver = 'rosenbrock'

responses,
        num_objective_functions = 1
        no_gradients
        no_hessians
```

Figure 2.3: Rosenbrock 2-D parameter study example: the DAKOTA input file.

facilities for surrogate modeling or nested iteration are not required, then it is not necessary to specify the `model` keyword at all, since the default behavior is the use of a "single" model constructed with the last set of responses, variables, and interface specified. In Figure 2.3, the keyword `single` explicitly specifies the use of a single model in the parameter study, even though this is the default.

The *variables* section of the input file specifies the characteristics of the parameters that will be used in the problem formulation. The variables can be continuous or discrete, and can be classified as design variables, uncertain variables, or state variables. See Chapter 12 for more information on the types of variables supported by DAKOTA. The *variables* section shown in Figure 2.3 specifies that there are two continuous design variables. The sub-specifications for continuous design variables provide the descriptors "x1" and "x2" as well as lower and upper bounds for these variables. The information about the variables is organized in column format for readability. So, both variables x_1 and x_2 have a lower bound of -2.0 and an upper bound of 2.0.

The *interface* section of the input file specifies what approach will be used to map variables into responses as well as details on how DAKOTA will pass data to and from a simulation code. In this example, the keyword `direct` is used to indicate the use of a function linked directly into DAKOTA. Alternatively, `fork` or `system` executions can be used to invoke instances of a simulation code that is external to DAKOTA, as explained in Section 2.4.11 and Chapter 17. The `analysis_driver` keyword indicates the name of the test function. With `fork` or `system`, default file names would be used for passing data between DAKOTA and the simulation code.

The *responses* section of the input file specifies the types of data that the interface will return to DAKOTA. For the example shown in Figure 2.3, the assignment `num_objective_functions = 1` indicates that there is only one objective function. Since there are no constraints associated with Rosenbrock's function, the keywords for constraint specifications are omitted. The keywords `no_gradients` and `no_hessians` indicate that no derivatives will be provided to the method; none are needed for a parameter study.

2.4 Example Problems

2.4.1 Two-Dimensional Parameter Study

The 2-D parameter study example problem listed in Figure 2.3 is executed by DAKOTA using the following command:

```
dakota dakota_rosenbrock_2d.in > 2d.out
```

The output of the DAKOTA run is directed to the file named `2d.out`. For comparison, a file named `2d.out.sav` is included in the `Dakota/examples/tutorial` directory. As for many of the examples, DAKOTA provides a report on the best design point located during the study at the end of these output files.

This 2-D parameter study produces the grid of data samples shown in Figure 2.4. Note that the `graphics` flag in the *strategy* section of the input file could be commented out since, for this example, the iteration history plots created by DAKOTA are not particularly instructive. More interesting visualizations can be created by importing DAKOTA's tabular data into an external graphics/plotting package. Common graphics and plotting packages include Mathematica, Matlab, Microsoft Excel, Origin, Tecplot, and many others. (Sandia National Laboratories and the DAKOTA developers do not endorse any of these commercial products.)

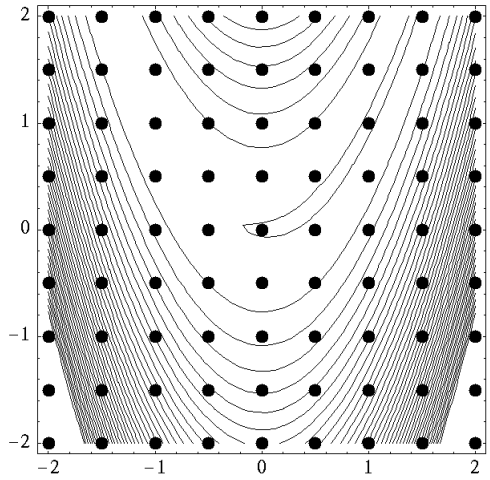


Figure 2.4: Rosenbrock 2-D parameter study example: location of the design points (dots) evaluated.

2.4.2 Vector Parameter Study

In addition to the multidimensional parameter study, DAKOTA can perform a vector parameter study, i.e., a parameter study between any two design points in an *n*-dimensional parameter space.

An input file for the vector parameter study is shown in Figure 2.5. The primary differences between this input file and the previous input file are found in the *variables* and *method* sections. In the variables section, the keywords for the bounds are removed and replaced with the keyword `initial_point` that specifies the starting point for the parameter study. In the method section, the `vector_parameter_study` keyword is used. The `final_point` keyword indicates the stopping point for the parameter study, and `num_steps` specifies the number of steps taken between the initial and final points in the parameter study.

```
# DAKOTA INPUT FILE - dakota_rosenbrock_vector.in


strategy,
        single_method
          graphics,tabular_graphics_data

method,
        vector_parameter_study
          final_point = 1.1  1.3
          num_steps = 10

model,
        single

variables,
        continuous_design = 2
          initial_point   -0.3      0.2
          descriptors      'x1'     "x2"

interface,
        direct
          analysis_driver = 'rosenbrock'

responses,
        num_objective_functions = 1
        no_gradients
        no_hessians
```

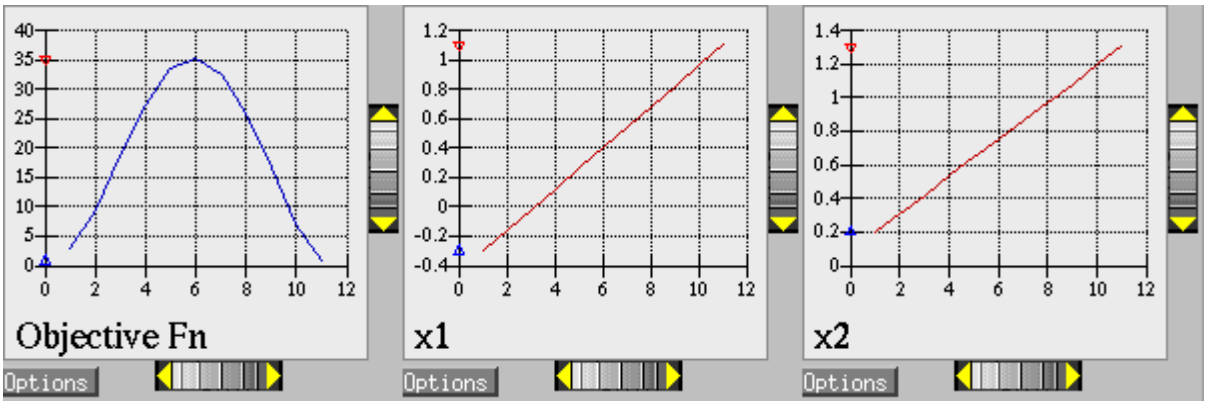
Figure 2.5: Rosenbrock vector parameter study example: the DAKOTA input file.

The vector parameter study example problem is executed using the command

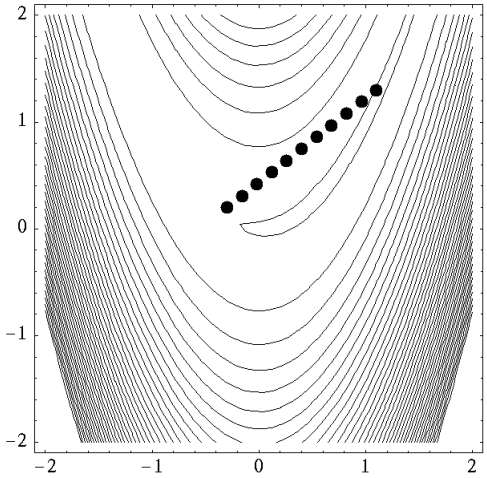
```
dakota dakota_rosenbrock_vector.in > vector.out
```

Figure 2.6(a) shows the graphics output created by DAKOTA. For this study, the simple DAKOTA graphics are more useful for visualizing the results. Figure 2.6(b) shows the locations of the 11 sample points generated in this study. It is evident from these figures that the parameter study starts within the banana-shaped valley, marches up the side of the hill, and then returns to the valley. The output file `vector.out.sav` is provided in the `Dakota/examples/tutorial` directory.

In addition to the vector and multidimensional examples shown, DAKOTA also supports list and centered parameter study methods. Refer to Chapter 4 for additional information.



(a)



(b)

Figure 2.6: Rosenbrock vector parameter study example: (a) screen capture of the DAKOTA graphics and (b) location of the design points (dots) evaluated.

2.4.3 Gradient-based Unconstrained Optimization

A DAKOTA input file for a gradient-based optimization of Rosenbrock's function is listed in Figure 2.7. The format of the input file is similar to that used for the parameter studies, but there are some new keywords in the responses and method sections. First, in the responses section of the input file, the keyword block starting with `numerical_gradients` specifies that a finite difference method will be used to compute gradients for the optimization algorithm. Note that the Rosenbrock function evaluation code inside DAKOTA has the ability to give analytical gradient values. (To switch from finite difference gradient estimates to analytic gradients, uncomment the `analytic_gradients` keyword and comment out the four lines associated with the `numerical_gradients` specification.) Next, in the method section of the input file, several new keywords have been added. In this section, the keyword `conmin_frcg` indicates the use of the Fletcher-Reeves conjugate gradient algorithm in the CONMIN optimization software package [130] for bound-constrained optimization. The keyword `max_iterations` is used to indicate the computational budget for this optimization (in this case, a single iteration includes multiple evaluations of Rosenbrock's function for the gradient computation steps and the line search steps). The keyword `convergence_tolerance` is used to specify one of CONMIN's convergence criteria (under which CONMIN terminates if the objective function value differs by less than the absolute value of the convergence tolerance for three successive iterations).

This DAKOTA input file is executed using the following command:

```
dakota dakota_rosenbrock_grad_opt.in > grad_opt.out
```

The sample file `grad_opt.out.sav` is included in `Dakota/examples/tutorial` for comparison. When this example problem is executed, DAKOTA creates some iteration history graphics similar to the screen capture shown in Figure 2.8(a). These plots show how the objective function and design parameters change in value during the optimization steps. The scaling of the horizontal and vertical axes can be changed by moving the scroll knobs on each plot. Also, the "Options" button allows the user to plot the vertical axes using a logarithmic scale. Note that log-scaling is only allowed if the values on the vertical axis are strictly greater than zero.

Figure 2.8(b) shows the iteration history of the optimization algorithm. The optimization starts at the point $(x_1, x_2) = (-1.2, 1.0)$ as given in the DAKOTA input file. Subsequent iterations follow the banana-shaped valley that curves around toward the minimum point at $(x_1, x_2) = (1.0, 1.0)$. Note that the function evaluations associated with the line search phase of each CONMIN iteration are not shown on the plot. At the end of the DAKOTA run, information is written to the output file to provide data on the optimal design point. These data include the optimum design point parameter values, the optimum objective and constraint function values (if any), plus the number of function evaluations that occurred and the amount of time that elapsed during the optimization study.

2.4.4 Gradient-based Constrained Optimization

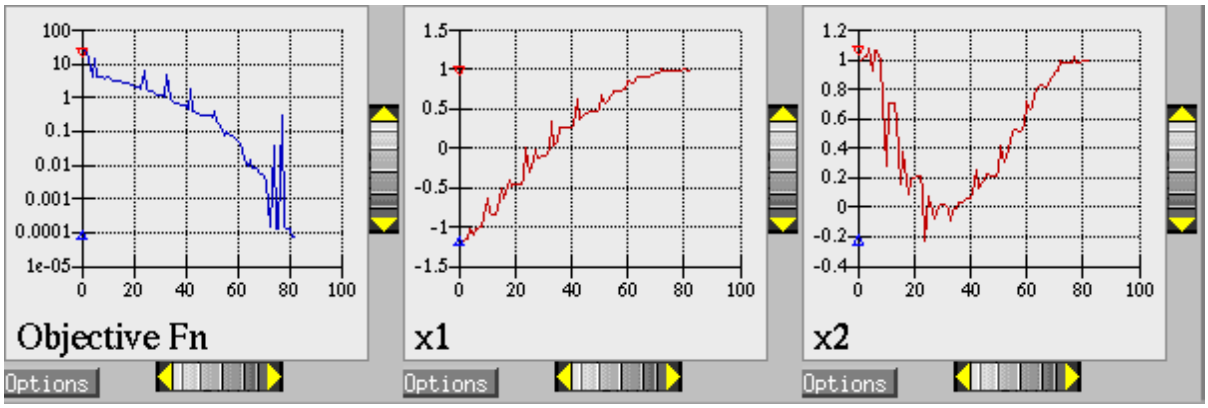
This example demonstrates the use of a gradient-based optimization algorithm on a nonlinearly constrained problem. The "textbook" example problem (see Section 2.2) is used for this purpose and the DAKOTA input file for this example problem is shown in Figure 2.9. This input file is similar to the input file for the unconstrained gradient-based optimization example problem involving the Rosenbrock function. Note the addition of commands in the responses section of the input file that identify the number and type of constraints, along with the upper bounds on these constraints. The commands `direct` and `analysis_driver = 'text_book'` specify that DAKOTA will use its internal version of the textbook problem.

The following command runs this example problem:

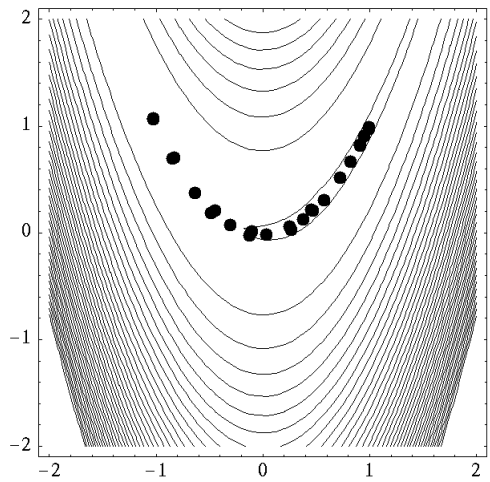
```
dakota dakota_textbook.in > textbook.out
```

```
# DAKOTA INPUT FILE - dakota_rosenbrock_grad_opt.in


strategy,
        single_method
          graphics,tabular_graphics_data

method,
        conmin_frcg
          max_iterations = 100
          convergence_tolerance = 1e-4

model,
        single

variables,
        continuous_design = 2
          initial_point   -1.2  1.0
          lower_bounds    -2.0     -2.0
          upper_bounds     2.0      2.0
          descriptors      'x1'     "x2"

interface,
        direct
          analysis_driver = 'rosenbrock'

responses,
        num_objective_functions = 1
#       analytic_gradients
        numerical_gradients
          method_source dakota
          interval_type forward
          fd_gradient_step_size = 1.e-5
        no_hessians
```

Figure 2.7: Rosenbrock gradient-based unconstrained optimization example: the DAKOTA input file.



(a)



(b)

Figure 2.8: Rosenbrock gradient-based unconstrained optimization example: (a) screen capture of the DAKOTA graphics and (b) sequence of design points (dots) evaluated (line search points omitted).

The file `textbook.out.sav` is included in `Dakota/examples/tutorial` for comparison purposes. The results of the optimization example problem are listed at the end of the `textbook.out` file. This information shows that the optimizer stopped at the point $(x_1, x_2) = (0.5, 0.5)$, where both constraints are approximately satisfied, and where the objective function value is 0.128. The progress of the optimization algorithm is shown in Figure 2.10(a) where the dots correspond to the end points of each iteration in the algorithm. The starting point is $(x_1, x_2) = (0.9, 1.1)$, where both constraints are violated. The optimizer takes a sequence of steps to minimize the objective function while reducing the infeasibility of the constraints. The optimization graphics are also shown in Figure 2.10(b).

```
# DAKOTA INPUT FILE - dakota_textbook.in


strategy,
        single_method
          graphics

method,
        dot_mmfd,
          max_iterations = 50,
          convergence_tolerance = 1e-4

variables,
        continuous_design = 2
          initial_point    0.9    1.1
          upper_bounds     5.8    2.9
          lower_bounds     0.5   -2.9
          descriptors      'x1'   'x2'

interface,
        direct
          analysis_driver =       'text_book'

responses,
        num_objective_functions = 1
        num_nonlinear_inequality_constraints = 2
        numerical_gradients
          method_source dakota
          interval_type central
          fd_gradient_step_size = 1.e-4
        no_hessians
```

Figure 2.9: Textbook gradient-based constrained optimization example: the DAKOTA input file.

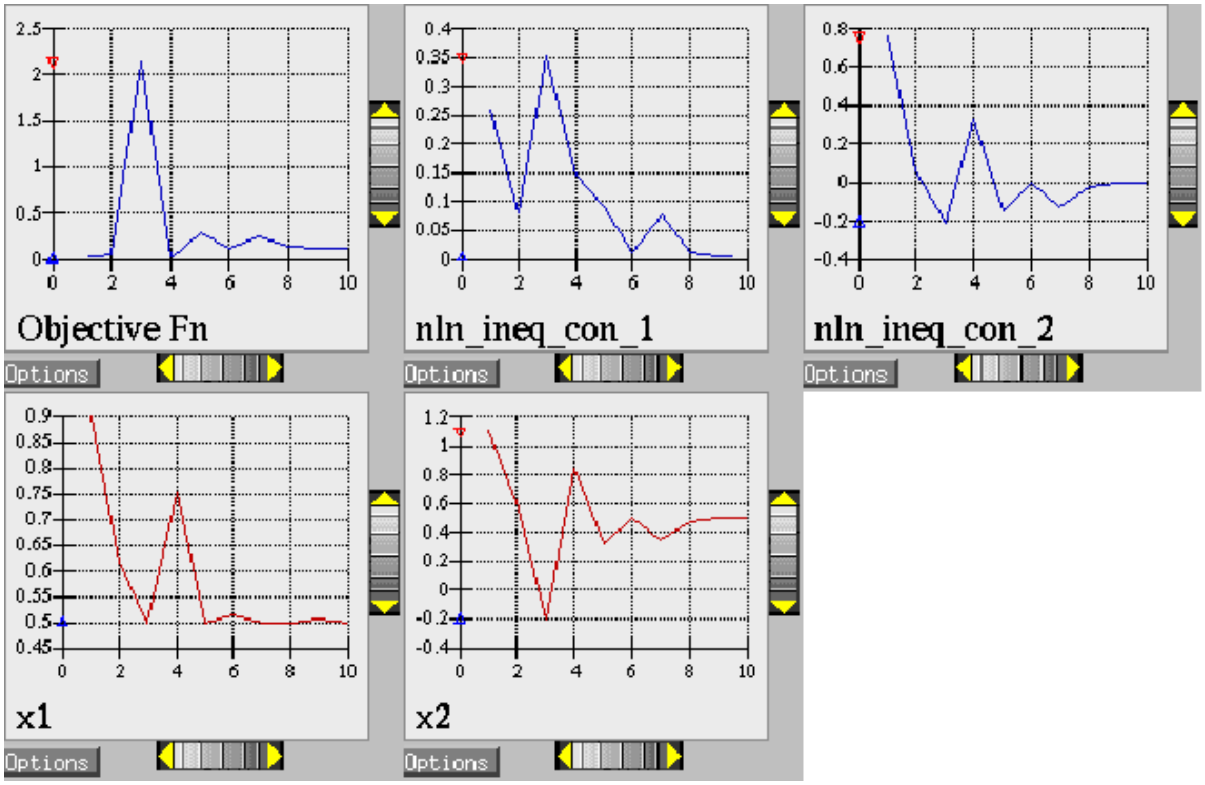
2.4.5 Nonlinear Least Squares Methods for Optimization

Both the Rosenbrock and textbook example problems can be formulated as least-squares minimization problems (see Section 22.1 and Section 22.2). For example, the Rosenbrock problem can be cast as:

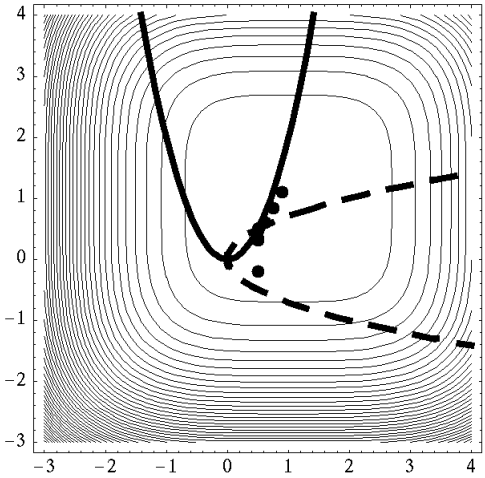
$$\texttt{minimize} \ \ (f_1)^2 + (f_2)^2 \tag{2.4}$$

where $f_1 = 10(x_2 - x_1^2)$ and $f_2 = (1 - x_1)$. When using a least-squares approach to minimize a function, each of the least-squares terms $f_1, f_2, \ldots$ is driven toward zero. This formulation permits the use of specialized algorithms that can be more efficient than general purpose optimization algorithms. See Chapter 8 for more detail on the algorithms used for least-squares minimization, as well as a discussion on the types of engineering design problems (e.g., parameter estimation) that can make use of the least-squares approach.

Figure 2.11 is a listing of the DAKOTA input file `dakota_rosenbrock_ls.in`. This input file differs from the input file shown in Figure 2.7 in several key areas. The responses section of the input file uses the keyword `num_least_squares_terms = 2` instead of the `num_objective_functions = 1`. The method section



(a)



(b)

Figure 2.10: Textbook gradient-based constrained optimization example: (a) screen capture of the DAKOTA graphics shows how the objective function was reduced during the search for a feasible design point and (b) iteration history (iterations marked by solid dots).

of the input file shows that the NL2SOL algorithm [20] (`nl2sol`) is used in this example. (The Gauss-Newton, NL2SOL, and NLSSOL SQP algorithms are currently available for exploiting the special mathematical structure of least squares minimization problems).

```
# DAKOTA INPUT FILE - dakota_rosenbrock_ls.in


strategy,
        single_method
          graphics,tabular_graphics_data

method,
        nl2sol
          max_iterations = 100
          convergence_tolerance = 1e-4

model,
        single

variables,
        continuous_design = 2
          initial_point   -1.2  1.0
          lower_bounds    -2.0     -2.0
          upper_bounds     2.0      2.0
          descriptors      'x1'     "x2"

interface,
        direct
          analysis_driver = 'rosenbrock'

responses,
        num_least_squares_terms = 2
        analytic_gradients
        no_hessians
```

Figure 2.11: Rosenbrock nonlinear least squares example: the DAKOTA input file.

The input file listed in Figure 2.11 is executed using the command:

```
dakota dakota_rosenbrock_ls.in > leastsquares.out
```

The file `leastsquares.out.sav` is included `Dakota/examples/tutorial` for comparison purposes. The optimization results at the end of this file show that the least squares minimization approach has found the same optimum design point, $(x1, x2) = (1.0, 1.0)$, as was found using the conventional gradient-based optimization approach. The iteration history of the least squares minimization is given in Figure 2.12, and shows that 14 function evaluations were needed for convergence. In this example the least squares approach required about half the number of function evaluations as did conventional gradient-based optimization. In many cases a good least squares algorithm will converge more rapidly in the vicinity of the solution.

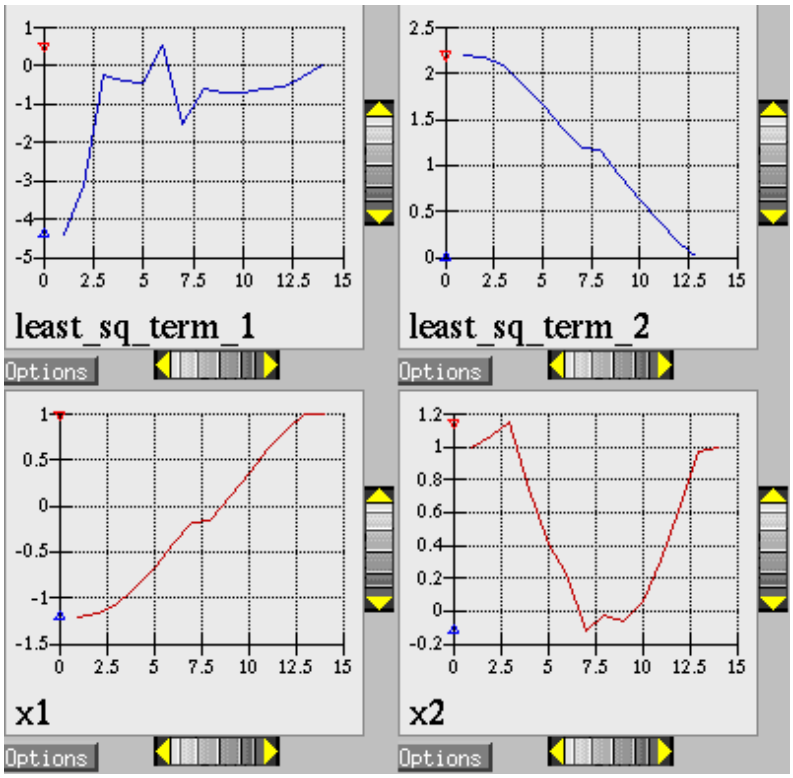


Figure 2.12: Rosenbrock nonlinear least squares example: iteration history for least squares terms f_1 and f_2.

2.4.6 Nongradient-based Optimization via Pattern Search

In addition to gradient-based optimization algorithms, DAKOTA also contains a variety of nongradient-based algorithms. One particular nongradient-based algorithm for local optimization is known as pattern search (see Chapter 1 for a discussion of local versus global optimization). The DAKOTA input file shown in Figure 2.13 applies a pattern search method to minimize the Rosenbrock function. While this provides for an interesting comparison to the previous example problems in this chapter, the Rosenbrock function is not the best test case for a pattern search method. That is, pattern search methods are better suited to problems where the gradients are too expensive to evaluate, inaccurate, or nonexistent — situations common among many engineering optimization problems. It also should be noted that nongradient-based algorithms generally are applicable only to unconstrained or bound-constrained optimization problems, although the inclusion of general linear and nonlinear constraints in nongradient-based algorithms is an active area of research in the optimization community. For most users who wish to use nongradient-based algorithms on constrained optimization problems, the easiest route is to create a penalty function, i.e., a composite function that contains the objective function and the constraints, external to DAKOTA and then optimize on this penalty function. Most optimization textbooks will provide guidance on selecting and using penalty functions.

The DAKOTA input file shown in Figure 2.13 is similar to the input file for the gradient-based optimization, except it has a different set of keywords in the method section of the input file, and the gradient specification in the responses section has been changed to `no_gradients`. The pattern search optimization algorithm used is part of the COLINY library [71]. See the DAKOTA Reference Manual [30] for more information on the *methods* section commands that can be used with COLINY algorithms.

```
# DAKOTA INPUT FILE - dakota_rosenbrock_ps_opt.in


strategy,
        single_method
          graphics,tabular_graphics_data

method,
        coliny_pattern_search
          max_iterations = 1000
          max_function_evaluations = 2000
          solution_accuracy = 1e-4
          initial_delta = 0.5
          threshold_delta = 1e-4
          exploratory_moves basic_pattern
          contraction_factor = 0.75

model,
        single

variables,
        continuous_design = 2
          initial_point    0.0  0.0
          lower_bounds    -2.0     -2.0
          upper_bounds     2.0      2.0
          descriptors      'x1'     "x2"

interface,
        direct
          analysis_driver = 'rosenbrock'

responses,
        num_objective_functions = 1
        no_gradients
        no_hessians
```

Figure 2.13: Rosenbrock pattern search optimization example: the DAKOTA input file.

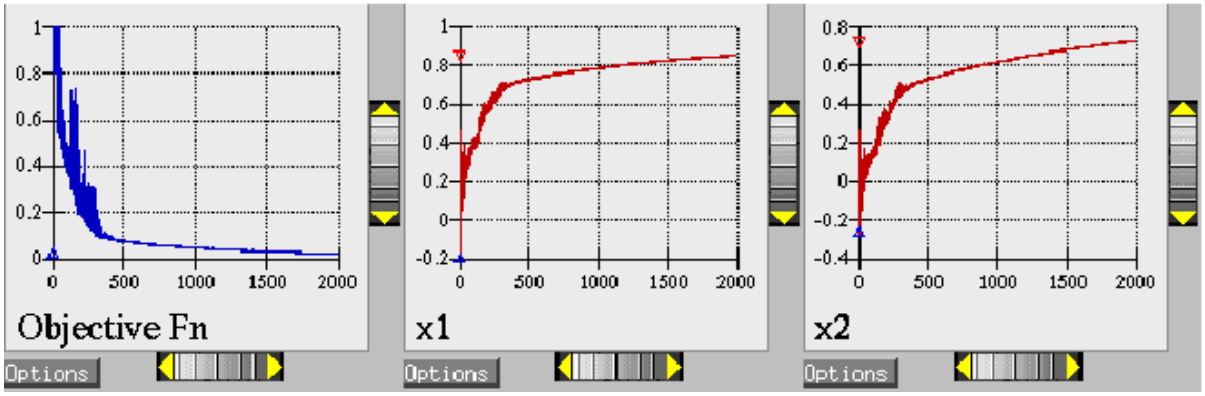
This DAKOTA input file is executed using the following command:

```
dakota dakota_rosenbrock_ps_opt.in > ps_opt.out
```

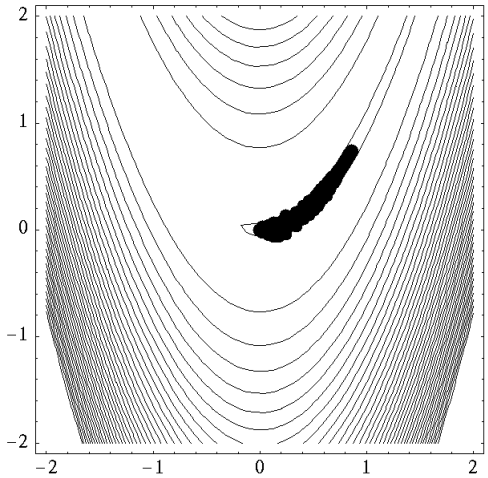
The file `ps_opt.out.sav` is included in the `Dakota/examples/tutorial` directory. For this run, the optimizer was given an initial design point of $(x_1, x_2) = (0.0, 0.0)$ and was limited to 2000 function evaluations. In this case, the pattern search algorithm stopped short of the optimum at $(x_1, x_2) = (1.0, 1, 0)$, although it was making progress in that direction when it was terminated. (It would have reached the minimum point eventually.)

The iteration history is provided in Figures 2.14(a) and (b), which show the locations of the function evaluations used in the pattern search algorithm. Figure 2.14(c) provides a close-up view of the pattern search function evaluations used at the start of the algorithm. The coordinate pattern is clearly visible at the start of the iteration

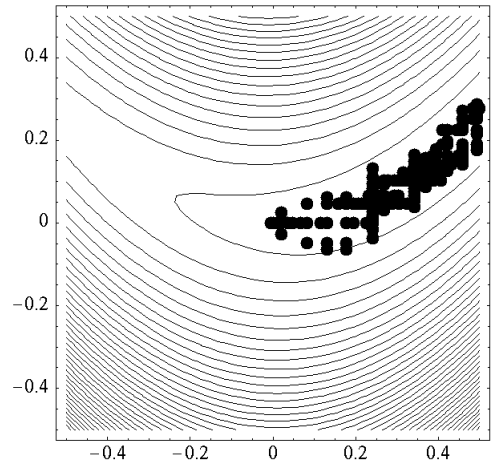
history, and the decreasing size of the coordinate pattern is evident at the design points move toward $(x_1, x_2) = (1.0, 1.0)$.



(a)



(b)



(c)

Figure 2.14: Rosenbrock pattern search optimization example: (a) screen capture of the DAKOTA graphics, (b) sequence of design points (dots) evaluated and (c) close-up view illustrating the shape of the coordinate pattern used.

While pattern search algorithms are useful in many optimization problems, this example shows some of the drawbacks to this algorithm. While a pattern search method may make good initial progress towards an optimum, it is often slow to converge. On a smooth, differentiable function such as Rosenbrock's function, a nongradient-based method will not be as efficient as a gradient-based method. However, there are many engineering design applications where gradient information is inaccurate or unavailable, which renders gradient-based optimizers ineffective. Thus, pattern search algorithms (and other nongradient-based algorithms such as genetic algorithms as discussed in the next section) are often good choices in complex engineering applications when the quality of gradient data is suspect.

2.4.7 Nongradient-based Optimization via Evolutionary Algorithm

In contrast to pattern search algorithms, which are local optimization methods, evolutionary algorithms (EA) are global optimization methods. As was described above for the pattern search algorithm, the Rosenbrock function is not an ideal test problem for showcasing the capabilities of evolutionary algorithms. Rather, EAs are best suited to optimization problems that have multiple local optima, and where gradients are either too expensive to compute or are not readily available.

Evolutionary algorithms are based on Darwin's theory of survival of the fittest. The EA algorithm starts with a randomly selected population of design points in the parameter space, where the values of the design parameters form a "genetic string," analogous to DNA in a biological system, that uniquely represents each design point in the population. The EA then follows a sequence of generations, where the best design points in the population (i.e., those having low objective function values) are considered to be the most "fit" and are allowed to survive and reproduce. The EA simulates the evolutionary process by employing the mathematical analogs of processes such as natural selection, breeding, and mutation. Ultimately, the EA identifies a design point (or a family of design points) that minimizes the objective function of the optimization problem. An extensive discussion of EAs is beyond the scope of this text, but may be found in a variety of sources (cf., [69] pp. 149-158; [63]). Currently, the EAs available in DAKOTA include a genetic algorithm for problems involving discrete variables and an evolution strategy with self-adaptation for problems with continuous variables. Details of these algorithms are given in the DAKOTA Reference Manual [30]. The COLINY library, which provides the EA software that has been linked into DAKOTA, is described in [71].

Figure 2.15 shows a DAKOTA input file that uses an EA to minimize the Rosenbrock function. For this example the EA has a population size of 50. At the start of the first generation, a random number generator is used to select 50 design points that will comprise the initial population. *[A specific seed value is used in this example to generate repeatable results, although, in general, one should use the default setting which allows the EA to choose a random seed.]* A two-point crossover technique is used to exchange genetic string values between the members of the population during the EA breeding process. The result of the breeding process is a population comprised of the 10 best "parent" design points (elitist strategy) plus 40 new "child" design points. The EA optimization process will be terminated after either 100 iterations (generations of the EA) or 2,000 function evaluations. The EA software available in DAKOTA provides the user with much flexibility in choosing the settings used in the optimization process. See [30] and [71] for details on these settings.

The following command runs DAKOTA on the input file:

```
dakota dakota_rosenbrock_ea_opt.in > ea_opt.out
```

A corresponding output file named `ea_opt.out.sav` appears in `Dakota/examples/tutorial`. The EA optimization results printed at the end of this file show that the best design point found was $(x_1, x_2) = (0.98, 0.95)$. The file `ea_tabular.dat.sav` provides a listing of the design parameter values and objective function values for all 2,000 design points evaluated during the running of the EA. Figure 2.16(a) shows the population of 50 randomly selected design points that comprise the first generation of the EA, and Figure 2.16(b) shows the final population of 50 design points, where most of the 50 points are clustered near $(x_1, x_2) = (0.98, 0.95)$.

As described above, an EA is not well-suited to an optimization problem involving a smooth, differentiable objective such as the Rosenbrock function. Rather, EAs are better suited to optimization problems where conventional gradient-based optimization fails, such as situations where there are multiple local optima and/or gradients are not available. In such cases, the computational expense of an EA is warranted since other optimization methods are not applicable or impractical. In many optimization problems, EAs often quickly identify promising regions of the design space where the global minimum may be located. However, an EA can be slow to converge to the optimum. For this reason, it can be an effective approach to combine the global search capabilities of a EA with the efficient local search of a gradient-based algorithm in a *hybrid optimization* strategy. In this approach, the op-

```
# DAKOTA INPUT FILE - dakota_rosenbrock_ea_opt.in


strategy,
        single_method
          graphics,tabular_graphics_data

method,
        coliny_ea
          max_iterations = 100
          max_function_evaluations = 2000
          seed = 11011011
          population_size = 50
          fitness_type merit_function
          mutation_type offset_normal
          mutation_rate 1.0
          crossover_type two_point
          crossover_rate 0.0
          replacement_type chc = 10

model,
        single

variables,
        continuous_design = 2
          lower_bounds    -2.0     -2.0
          upper_bounds     2.0      2.0
          descriptors      'x1'     "x2"

interface,
        direct
          analysis_driver = 'rosenbrock'

responses,
        num_objective_functions = 1
        no_gradients
        no_hessians
```

Figure 2.15: Rosenbrock evolutionary algorithm optimization example: the DAKOTA input file.

timization starts by using a few iterations of a EA to provide the initial search for a good region of the parameter space (low objective function and/or feasible constraints), and then it switches to a gradient-based algorithm (using the best design point found by the EA as its starting point) to perform an efficient local search for an optimum design point. More information on this hybrid approach is provided in Chapter 10.

In addition to the evolutionary algorithm capabilities in the `coliny_ea` method, there is a single-objective genetic algorithm method called `soga`. For more information on `soga`, see Chapter 7.

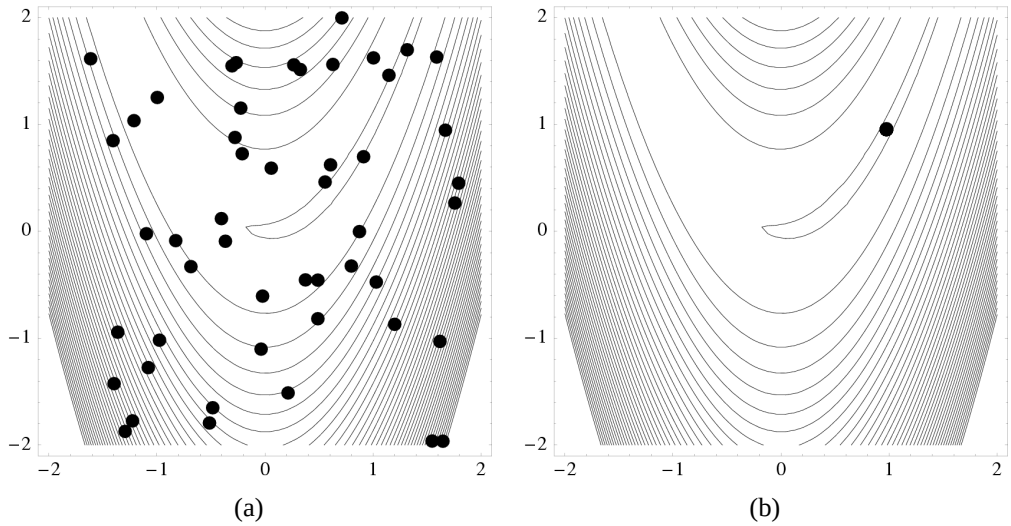


Figure 2.16: Rosenbrock evolutionary algorithm optimization example: 50 design points in the (a) initial and (b) final populations selected by the evolutionary algorithm.

2.4.8 Multiobjective Optimization

Multiobjective optimization means that there are two or more objective functions that you wish to optimize simultaneously. Often these are conflicting objectives, such as cost and performance. The answer to a multi-objective problem is usually not a single point. Rather, it is a set of points called the Pareto front. Each point on the Pareto front satisfies the Pareto optimality criterion, i.e., locally there exists no other feasible vector that would improve some objective without causing a simultaneous worsening in at least one other objective. Thus a feasible point X' from which small moves improve one or more objectives without worsening any others is not Pareto optimal: it is said to be "dominated" and the points along the Pareto front are said to be "non-dominated".

Often multi-objective problems are addressed by simply assigning weights to the individual objectives, summing the weighted objectives, and turning the problem into a single-objective one which can be solved with a variety of optimization techniques. While this approach provides a useful "first cut" analysis (and is supported within DAKOTA—see Section 7.3), this approach has many limitations. The major limitation is that a local solver with a weighted sum objective will only find one point on the Pareto front; if one wants to understand the effects of changing weights, this method can be computationally expensive. Since each optimization of a single weighted objective will find only one point on the Pareto front, many optimizations must be performed to get a good parametric understanding of the influence of the weights and to achieve a good sampling of the entire Pareto frontier.

Starting with version 3.2 of DAKOTA, a capability to perform multi-objective optimization based on a genetic algorithm method has been available. This method is called `moga`. It is based on the idea that as the population evolves in a GA, solutions that are non-dominated are chosen to remain in the population. Until version 4.0 of DAKOTA, there was a selection_type choice of domination_count that performed a custom fitness assessment and selection operation together. As of version 4.0 of DAKOTA, that functionality has been broken into separate, more generally usable fitness assessment and selection operators called the domination_count fitness assessor and below_limit selector respectively. The effect of using these two operators is the same as the previous behavior of the domination_count selector. This means of selection works especially well on multi-objective problems because it has been specifically designed to avoid problems with aggregating and scaling objective function values and transforming them into a single objective. Instead, the fitness assessor works by ranking population members

such that their resulting fitness is a function of the number of other designs that dominate them. The below_limit selector then chooses designs by considering the fitness of each. If the fitness of a design is above a certain limit, which in this case corresponds to a design being dominated by more than a specified number of other designs, then it is discarded. Otherwise it is kept and selected to go to the next generation. The one catch is that this selector will require that a minimum number of selections take place. The `shrinkage_percentage` determines the minimum number of selections that will take place if enough designs are available. It is interpreted as a percentage of the population size that must go on to the subsequent generation. To enforce this, the below_limit selector makes all the selections it would make anyway and if that is not enough, it relaxes its limit and makes selections from the remaining designs. It continues to do this until it has made enough selections. The moga method has many other important features. Complete descriptions can be found in the DAKOTA Reference Manual [30].

Figure 2.17 shows an example input file that demonstrates some of the multi-objective capabilities available with the moga method.

This example has three input variables and two objectives. The example uses objectives different from the Rosenbrock function because we wanted to demonstrate the capability on a problem with two conflicting objectives. This example is taken from a testbed of multi-objective problems [17]. The final results from moga are output to a file called `finaldata.dat` in the directory in which you are running. This `finaldata.dat` file is simply a list of inputs and outputs. Plotting the output columns against each other allows one to see the Pareto front generated by `moga`. Figure 2.18 shows an example of the Pareto front for this problem. Note that a Pareto front easily shows the tradeoffs between Pareto optimal solutions. For example, look at the point with f1 and f2 values equal to (0.9, 0.25). One cannot improve (minimize) the value of objective function f1 without increasing the value of f2: another point on the Pareto front, (0.6, 0.6) represents a better value of objective f1 but a worse value of objective f2.

Sections 7.2 and 7.3 provide more information on multiobjective optimization. There are three detailed examples provided in Section 22.11.

2.4.9 Monte Carlo Sampling

Figure 2.19 shows the DAKOTA input file for an example problem that demonstrates some of the random sampling capabilities available in DAKOTA. In this example, the design parameters, x1 and x2, will be treated as uncertain parameters that have uniform distributions over the interval [-2, 2]. This is specified in the variables section of the input file, beginning with the keyword `uniform_uncertain`. Another change from earlier input files, such as Figure 2.5, occurs in the responses section, where the keyword `num_response_functions` is used in place of `num_objective_functions`. The final changes to the input file occur in the method section, where the keyword `nond_sampling` is used; “nond” is an abbreviation for nondeterministic. The other keywords in the methods section of the input file specify the number of samples (200), the seed for the random number generator (17), the sampling method (random), and the response threshold (100.0). The `seed` specification allows a user to obtain repeatable results from multiple runs. If a seed value is not specified, then DAKOTA’s sampling methods are designed to generate nonrepeatable behavior (by initializing the seed using a system clock). The keyword `response_thresholds` allows the user to specify threshold values for which DAKOTA will compute statistics on the response function output. Note that a unique threshold value can be specified for each response function.

In this example, DAKOTA will select 200 design points from within the parameter space, evaluate the value of Rosenbrock’s function at all 200 points, and then perform some basic statistical calculations on the 200 response values.

This DAKOTA input file is executed using the following command:

```
dakota dakota_rosenbrock_nond.in > nond.out
```

```
##  DAKOTA INPUT FILE - dakota_mogatest1.in



strategy,
        single
        graphics tabular_graphics_data

method,
        moga
        output silent
        seed = 10983
        max_function_evaluations = 2500
        initialization_type unique_random
        crossover_type shuffle_random
          num_offspring = 2 num_parents = 2
          crossover_rate = 0.8
        mutation_type replace_uniform
          mutation_rate = 0.1
        fitness_type domination_count
        replacement_type below_limit = 6
          shrinkage_percentage = 0.9
        convergence_type metric_tracker
          percent_change = 0.05 num_generations = 40

variables,
        continuous_design = 3
          initial_point     0            0       0
          upper_bounds      4    4       4
          lower_bounds     -4   -4      -4
          descriptors      'x1'         'x2'    'x3'

interface,
        system
          analysis_driver = 'mogatest1'

responses,
        num_objective_functions = 2
        no_gradients
        no_hessians
```

Figure 2.17: Multiple objective genetic algorithm (MOGA) example: the DAKOTA input file.

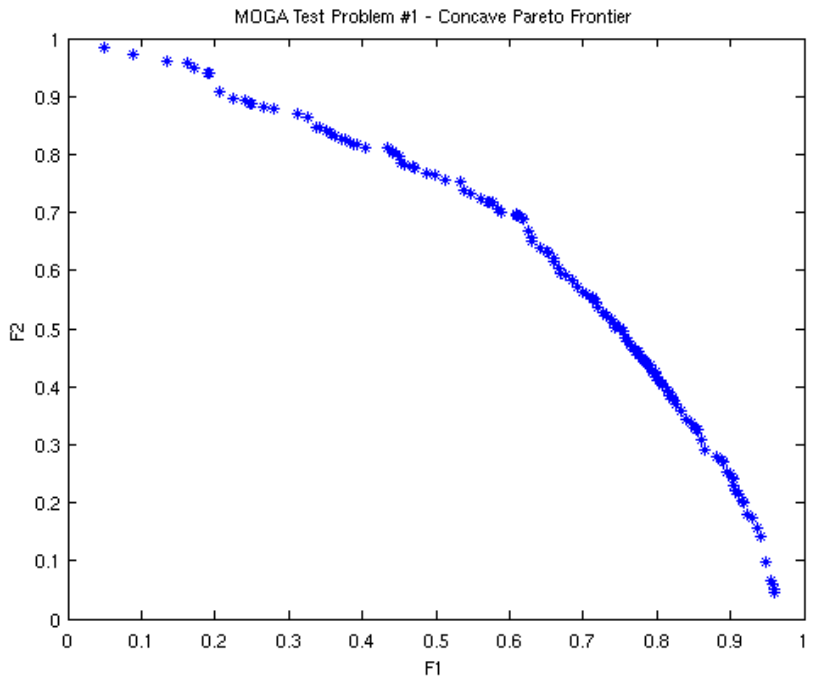


Figure 2.18: Multiple objective genetic algorithm (MOGA) example: Pareto front showing tradeoffs between functions f1 and f2.

See the file `nond.out.sav` in `Dakota/examples/tutorial` for comparison with results produced by DAKOTA. Note that your results will differ from those in this file if your `seed` value differs or if no `seed` is specified.

The statistical data on the 200 Monte Carlo samples is printed at the end of the output file in the section that starts with "Statistics based on 200 samples." In this section, DAKOTA outputs the mean, standard deviation, coefficient of variation, and 95% confidence intervals for each of the response functions, followed by the fractions ("Probability Level") of the response function values that are below the response threshold values specified in the input file, as well as several correlation matrices. Figure 2.20 shows the locations of the 200 sample sites within the parameter space of the Rosenbrock function.

2.4.10 Optimization with a User-Supplied Simulation Code - Case 1

Many of the previous examples made use of the direct interface to access the Rosenbrock and textbook test functions that are compiled into DAKOTA. In engineering applications, it is much more common to use the `system` or `fork` interface approaches within DAKOTA to manage external simulation codes. In both of these cases, the communication between DAKOTA and the external code is conducted through the reading and writing of short text files. For this example, the C++ program `rosenbrock.C` in `Dakota/test` is used as the simulation code. This file is compiled to create the stand-alone `rosenbrock` executable that is referenced as the `analysis_driver` in Figure 2.21. This stand-alone program performs the same function evaluations as DAKOTA's internal Rosenbrock test function.

Figure 2.21 shows the text of the DAKOTA input file named `dakota_rosenbrock_syscall.in` that is

```
# DAKOTA INPUT FILE - dakota_rosenbrock_nond.in


strategy,
        single_method
          graphics,tabular_graphics_data

method,
        nond_sampling
          samples = 200 seed = 17
          sample_type random
          response_levels = 100.0

model,
        single

variables,
        uniform_uncertain = 2
          lower_bounds -2.0  -2.0
          upper_bounds  2.0   2.0
          descriptors        'x1'  'x2'

interface,
        direct
          analysis_driver = 'rosenbrock'

responses,
        num_response_functions = 1
        no_gradients
        no_hessians
```

Figure 2.19: Monte Carlo sampling example: the DAKOTA input file.

provided in the directory `Dakota/examples/tutorial`. The only differences between this input file and the one in Figure 2.7 occur in the *interface* keyword section. The keyword `system` indicates that DAKOTA will use system calls to create separate Unix processes for executions of the user-supplied simulation code. The name of the simulation code, and the names for DAKOTA's parameters and results file are specified using the `analysis_driver`, `parameters_file`, and `results_file` keywords, respectively.

This example problem is executed using the command:

```
dakota dakota_rosenbrock_syscall.in > syscall.out
```

This run of DAKOTA takes longer to complete than the previous gradient-based optimization example since the `system` interface method has additional process creation and file I/O overhead, as compared to the internal communication that occurs when the `direct` interface method is used. File `syscall.out.sav` in the `Dakota/examples/tutorial` directory permits comparison with output results you get by executing the command given above.

To gain a better understanding of what exactly DAKOTA is doing with the `system` interface approach, add the keywords `file_tag` and `file_save` to the interface specification and re-run DAKOTA. Check the listing of

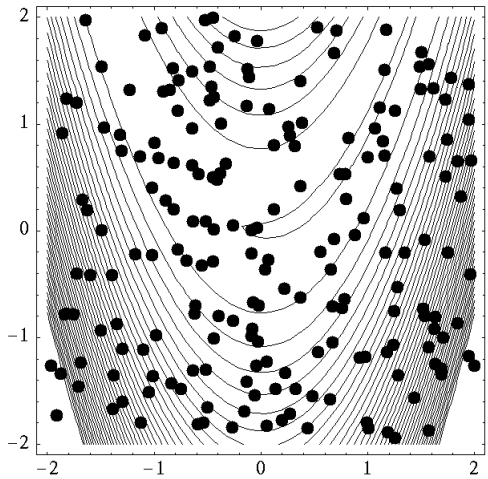


Figure 2.20: Monte Carlo sampling example: locations in the parameter space of the 200 Monte Carlo samples using a uniform distribution for both x_1 and x_2.

the local directory and you will see many new files with names such as `params.in.1`, `params.in.2`, etc., and `results.out.1`, `results.out.2`, etc. There is one `params.in.X` file and one `results.out.X` file for each of the function evaluations performed by DAKOTA. This is the file listing for `params.in.1`:

```
                        2 variables
-1.200000000000000e+00 x1
 1.000000000000000e+00 x2
                        1 functions
                        1 ASV_1
                        2 derivative_variables
                        1 DVV_1
                        2 DVV_2
                        0 analysis_components
```

The basic pattern is that of array lengths and string identifiers followed by listings of the array entries, where the arrays consist of the variables, the active set vector (ASV), the derivative values vector (DVV), and the analysis components (AC). For the variables array, the first line gives the total number of variables (2) and the "variables" string identifier, and the subsequent two lines provide the array listing for the two variable values (-1.2 and 1.0) and descriptor tags ("x1" and "x2" from the DAKOTA input file). The next array conveys the ASV, which indicates what simulator outputs are needed. The first line of the array gives the total number of response functions (1) and the "functions" string identifier, followed by one ASV code and descriptor tag ("ASV_1") for each function. In this case, the ASV value of 1 indicates that DAKOTA is requesting that the simulation code return the response function value in the file `results.out.X`. (Possible ASV values: 1 = value of response function value, 2 = response function gradient, 4 = response function Hessian, and any sum of these for combinations up to 7 = response function value, gradient, and Hessian; see 12.7 for more detail.) The next array provides the DVV, which defines the variable identifiers used in computing derivatives. The first line of the array gives the number of derivative variables (2) and the "derivative_variables" string identifier, followed by the listing of the two DVV variable identifiers (the first and second variables) and descriptor tags ("DVV_1" and "DVV_2"). The final array provides the AC array used to provide additional strings for use by the simulator (e.g., to provide the name of a particular mesh file). The first line of the array gives the total number of analysis components (0) and the "analysis_components" string identifier, followed by the listing of the array, which is empty in this case.

The executable program rosenbrock reads in the `params.in.X` file and evaluates the objective function at the given values for x_1 and x_2. Then, rosenbrock writes out the objective function data to the `results.out.X` file. Here is the listing for the file `results.out.1`:

```
2.420000000000000e+01 f
```

The value shown above is the value of the objective function, and the descriptor 'f' is an optional tag returned by the simulation code. When the system call has completed, DAKOTA reads in the data from the `results.in.X` file and processes the results. DAKOTA then continues with additoinal executions of the rosenbrock program until the optimization process is complete.

```
# DAKOTA INPUT FILE - dakota_rosenbrock_syscall.in


strategy,
        single_method
          graphics,tabular_graphics_data

method,
        conmin_frcg
          max_iterations = 100
          convergence_tolerance = 1e-4

model,
        single

variables,
        continuous_design = 2
          initial_point   -1.2  1.0
          lower_bounds    -2.0     -2.0
          upper_bounds     2.0      2.0
          descriptors      'x1'     "x2"

interface,
        system
          analysis_driver = 'rosenbrock'
          parameters_file = 'params.in'
          results_file    = 'results.out'

responses,
        num_objective_functions = 1
        numerical_gradients
          method_source dakota
          interval_type forward
          fd_gradient_step_size = 1.e-5
        no_hessians
```

Figure 2.21: DAKOTA input file for gradient-based optimization using the system call interface to an external rosenbrock simulator.

2.4.11 Optimization with a User-Supplied Simulation Code - Case 2

In many situations the user-supplied simulation code cannot be modified to read and write the `params.in.X` `file` and the `results.out.X` file, as described above. Typically, this occurs when the simulation code is a commercial or proprietary software product that has specific input file and output file formats. In such cases, it is common to replace the executable program name in the DAKOTA input file with the name of a Unix shell script containing a sequence of commands that read and write the necessary files and run the simulation code. For example, the executable program named `rosenbrock` listed in Figure 2.21 could be replaced by a Unix C-shell script named `simulator_script`, with the script containing a sequence of commands to perform the following steps: insert the data from the `parameters.in.X` file into the input file of the simulation code, execute the simulation code, post-process the files generated by the simulation code to compute response data, and return the response data to DAKOTA in the `results.out.X` file. The steps that are typically used in constructing and using a Unix shell script are described in Section 17.1.

2.5 Where to Go from Here

This chapter has provided an introduction to the basic capabilities of DAKOTA including parameter studies, various types of optimization, and uncertainty quantification sampling. More information on the DAKOTA input file syntax is provided in the remaining chapters in this manual and in the DAKOTA Reference Manual [30]. Additional example problems that demonstrate some of DAKOTA's advanced capabilities are provided in Chapter 6, Chapter 9, Chapter 10, Chapter 17, and Chapter 22.

Here are a few pointers to sections of this manual that many new users find useful:

- Chapter 16 describes the different DAKOTA output file formats, including commonly encountered error messages.
- Chapter 17 demonstrates how to employ DAKOTA with a user-supplied simulation code.
  Most DAKOTA users will follow the approach described in Chapter 17.
- Chapter 19 provides guidelines on how to choose an appropriate optimization, uncertainty quantification, or parameter study method based on the characteristics of your application.
- Chapter 20 describes the file restart and data re-use capabilities of DAKOTA.

Chapter 3

DAKOTA Capability Overview

3.1 Purpose

This chapter provides a brief, but comprehensive, overview of DAKOTA's capabilities. Additional details and example problems are provided in subsequent chapters in this manual.

3.2 Parameter Study Methods

Parameter studies are often performed to explore the effect of parametric changes within simulation models. DAKOTA users may select from four parameter study methods.

Multidimensional: Forms a regular lattice or grid in an n-dimensional parameter space, where the user specifies the number of intervals used for each parameter.

Vector: Performs a parameter study along a line between any two points in an n-dimensional parameter space, where the user specifies the number of steps used in the study.

Centered: Given a point in an n-dimensional parameter space, this method evaluates nearby points along the coordinate axes of the parameter space. The user selects the number of steps and the step size.

List: The user supplies a list of points in an n-dimensional space where DAKOTA will evaluate response data from the simulation code.

Additional information on these methods is provided in Chapter 4.

3.3 Design of Experiments

Design of experiments are often used to explore the parameter space of an engineering design problem, for example to perform global sensitivity analysis. In design of experiments, especially design of computer experiments, one wants to generate input points that provide good coverage of the input parameter space. There is significant overlap between design of experiments and sampling methods, and both techniques can yield similar results about response function behavior and the relative importance of the input variables. We consider design of experiment methods to generate sets of uniform random variables on the interval $[0, 1]$, with the goal of characterizing the

behavior of the response functions over the input parameter ranges of interest. Uncertainty quantification, in contrast, involves characterizing the uncertain input variables with probability distributions such as normal, Weibull, triangular, etc., sampling from the input distributions, and propagating the input uncertainties to obtain a cumulative distribution function on the output or system response. We typically use the Latin Hypercube Sampling software (also developed at Sandia) for generating samples on input distributions used in uncertainty quantification. LHS is explained in more detail in the subsequent section 3.4. Two software packages are available in DAKOTA for design of computer experiments, DDACE (developed at Sandia Labs) and FSUDACE (developed at Florida State University).

DDACE (Distributed Design and Analysis of Computer Experiments): The DACE package includes both stochastic sampling methods and classical design of experiments methods [127]. The stochastic methods are Monte Carlo (random) sampling, Latin Hypercube sampling, orthogonal array sampling, and orthogonal array-latin hypercube sampling. The orthogonal array sampling allows for the calculation of main effects. The DDACE package currently supports variables that have either normal or uniform distributions. However, only the uniform distribution is available in the DAKOTA interface to DDACE. The classical design of experiments methods in DDACE are central composite design (CCD) and Box-Behnken (BB) sampling. A grid-based sampling method also is available. DDACE is available under a GNU Lesser General Public License and is distributed with DAKOTA.

FSUDace (Florida State University Design and Analysis of Computer Experiments): The FSUDace package provides quasi-Monte Carlo sampling (Halton and Hammersley) and Centroidal Voronio Tesselation (CVT) methods. The quasi-Monte Carlo and CVT methods are designed with the goal of low discrepancy. Discrepancy refers to the nonuniformity of the sample points within the unit hypercube. Low discrepancy sequences tend to cover the unit hypercube reasonably uniformly. Quasi-Monte Carlo methods produce low discrepancy sequences, especially if one is interested in the uniformity of projections of the point sets onto lower dimensional faces of the hypercube. CVT does very well volumetrically: it spaces the points fairly equally throughout the space, so that the points cover the region and are isotropically distributed with no directional bias in the point placement. FSUDace is available under a GNU Lesser General Public License and is distributed with DAKOTA.

PSUADE (Problem Solving Environment for Uncertainty Analysis and Design Exploration): PSUADE is a Lawrence Livermore National Laboratory tool for metamodeling, sensitivity analysis, uncertainty quantification, and optimization. Its features include non-intrusive and parallel function evaluations, sampling and analysis methods, an integrated design and analysis framework, global optimization, numerical integration, response surfaces (MARS and higher order regressions), graphical output with Pgplot or Matlab, and fault tolerance [126]. *While PSUADE is only available internally at LLNL,* DAKOTA includes a prototype interface to its MOAT sampling method, a valuable tool for global sensitivity analysis.

Additional information on these methods is provided in Chapter 5.

3.4 Uncertainty Quantification

Uncertainty quantification methods (also referred to as nondeterministic analysis methods) involve the computation of probabilistic information about response functions based on sets of simulations taken from the specified probability distributions for uncertain input parameters. Put another way, these methods perform a forward uncertainty propagation in which probability information for input parameters is mapped to probability information for output response functions. The UQ methods in DAKOTA include various sampling-based approaches (e.g., Monte Carlo and Latin Hypercube sampling), local and global reliability methods, stochastic expansion approaches, and epistemic methods.

LHS (Latin Hypercube Sampling): This package provides both Monte Carlo (random) sampling and Latin

Hypercube sampling methods, which can be used with probabilistic variables in DAKOTA that have the following distributions: normal, lognormal, uniform, loguniform, triangular, exponential, beta, gamma, gumbel, frechet, weibull, and user-supplied histograms. In addition, LHS accounts for correlations among the variables [80], which can be used to accommodate a user-supplied correlation matrix or to minimize correlation when a correlation matrix is not supplied. The LHS package currently serves two purposes: (1) it can be used for uncertainty quantification by sampling over uncertain variables characterized by probability distributions, or (2) it can be used in a DACE mode in which any design and state variables are treated as having uniform distributions (see the `all_variables` flag in the DAKOTA Reference Manual [30]). The LHS package historically came in two versions: "old" (circa 1980) and "new" (circa 1998), but presently only the latter is supported in DAKOTA, requiring a Fortran 90 compiler. This "new" LHS is available under a separate GNU General Public License and is distributed with DAKOTA.

Reliability Methods: This suite of methods includes both local and global reliability methods. Local methods include first- and second-order versions of the Mean Value method (MVFOSM and MVSOSM) and a variety of most probable point (MPP) search methods, including the Advanced Mean Value method (AMV and AMV^2), the iterated Advanced Mean Value method (AMV+ and AMV^2+), the Two-point Adaptive Nonlinearity Approximation method (TANA-3), and the traditional First Order and Second Order Reliability Methods (FORM and SORM) [70]. Each of the MPP search techniques solve local optimization problems in order to locate the MPP, which is then used as the point about which approximate probabilities are integrated (using first- or second-order integrations in combination with refinements based on importance sampling). Reliability mappings may involve computing reliability and probability levels for prescribed response levels (forward reliability analysis, commonly known as the reliability index approach or RIA) or computing response levels for prescribed reliability and probability levels (inverse reliability analysis, commonly known as the performance measure approach or PMA). Approximation-based MPP search methods (AMV, AMV^2, AMV+, AMV^2+, and TANA) may be applied in either x-space or u-space, and mappings may involve either cumulative or complementary cumulative distribution functions. Global reliability methods are designed to handle nonsmooth and multimodal failure surfaces, by creating global approximations based on Gaussian process models. They accurately resolve a particular contour of a response function and then estimate probabilities using multimodal adaptive importance sampling.

Stochastic Expansion Methods: The objective of these techniques is to characterize the response of systems whose governing equations involve stochastic coefficients. The development of these techniques mirrors that of deterministic finite element analysis utilizing the notions of projection, orthogonality, and weak convergence [53], [54]. Rather than estimating point probabilities, they form an approximation to the functional relationship between response functions and their random inputs, which provides a more complete uncertainty representation for use in multi-code simulations. Expansion methods include the Wiener-Askey generalized polynomial chaos expansion (PCE), which employs a family of multivariate orthogonal polynomials that are well matched to particular input probability distributions, and stochastic collocation (SC), which employs multivariate Lagrange interpolation polynomials. For PCE, expansion coefficients may be evaluated using a spectral projection approach (based on sampling, quadrature, or sparse grid methods for integration) or a point collocation approach (based on linear regression). For SC, interpolants may be formed over tensor-product quadrature grids or Smolyak sparse grids. Both methods provide analytic response moments; however, CDF/CCDF probabilities are evaluated by sampling on the expansion.

Dempster-Shafer Theory of Evidence: The objective of Evidence theory is to model the effects of epistemic uncertainties. Epistemic uncertainty refers to the situation where one does not know enough to specify a probability distribution on a variable. Sometimes epistemic uncertainty is referred to as subjective, reducible, or lack of knowledge uncertainty. In contrast, aleatory uncertainty refers to the situation where one does have enough information to specify a probability distribution. In Dempster-Shafer theory of evidence, the uncertain input variables are modeled as sets of intervals. The user assigns a basic probability assignment (BPA) to each interval, indicating how likely it is that the uncertain input falls within the interval. The intervals may be overlapping, contiguous, or have gaps. The intervals and their associated BPAs are then propagated through the simulation

to obtain cumulative distribution functions on belief and plausibility. Belief is the lower bound on a probability estimate that is consistent with the evidence, and plausibility is the uppder bound on a probability estimate that is consistent with the evidence.

Additional information on these methods is provided in Chapter 6.

3.5 Optimization

Several optimization software packages have been integrated with DAKOTA. These include freely-available software packages developed by research groups external to Sandia Labs, Sandia-developed software that has been released to the public under GNU licenses, and commercially-developed software. These optimization software packages provide the DAKOTA user with access to well-tested, proven methods for use in engineering design applications, as well as access to some of the newest developments in optimization algorithm research.

APPSPACK: is a derivative-free optimization library [64]. More specifically, it is an asynchronous implementation of generating set search. APPSPACK can handle unconstrained problems as well as those with bound constraints, linear constraints [66], and general nonlinear constraints [65]. APPSPACK was previously integrated with DAKOTA via COLINY, but is now directly integrated. APPSPACK is available to the public under the GNU LGPL and the source code is included with DAKOTA (web page: `http://software.sandia.gov/appspack`).

COLINY: Methods for nongradient-based local and global optimization which utilize the Common Optimization Library INterface (COLIN). COLINY currently includes evolutionary algorithms (including several genetic algorithms and Evolutionary Pattern Search), simple pattern search, Monte Carlo sampling, and the DIRECT and Solis-Wets algorithms. COLINY also include interfaces to third-party optimizer COBYLA2. This software is available to the public under a GNU Lesser General Public License (LGPL) through ACRO (A Common Repository for Optimizers) and the source code for COLINY is included with DAKOTA (web page: `http://www.cs.sandia.gov/Acro`).

CONMIN (CONstrained MINimization): Methods for gradient-based constrained and unconstrained optimization [130]. The constrained optimization algorithm is the method of feasible directions (MFD) and the unconstrained optimization algorithm is the Fletcher-Reeves conjugate gradient (CG) method. This software is freely available to the public from NASA, and the CONMIN source code is included with DAKOTA.

DOT (Design Optimization Tools): Methods for gradient-based optimization for constrained and unconstrained optimization problems [132]. The algorithms available for constrained optimization are modified-MFD, SQP, and sequential linear programming (SLP). The algorithms available for unconstrained optimization are the Fletcher-Reeves CG method and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton technique. DOT is a commercial software product of Vanderplaats Research and Development, Inc. (web page: `http://www.vrand.com`). Sandia National Laboratories and Los Alamos National Laboratory have limited seats for DOT. *Other users may obtain their own copy of DOT and compile it with the DAKOTA source code by following the steps given in the file* `Dakota/INSTALL`.

JEGA: provides SOGA and MOGA (single- and multi-objective genetic algorithms) optimization methods. The SOGA method provides a basic GA optimization capability that uses many of the same software elements as the MOGA method. The MOGA package allows for the formulation of multiobjective optimization problems without the need to specify weights on the various objective function values. The MOGA method directly identifies nondominated design points that lie on the Pareto front through tailoring of its genetic search operators. The advantage of the MOGA method versus conventional multiobjective optimization with weight factors (see Section 3.6), is that MOGA finds points along the entire Pareto front whereas the multiobjective optimization method produces only a single point on the Pareto front. The advantage of the MOGA method versus the Pareto-set optimization

strategy (see Section 3.9) is that MOGA is better able to find points on the Pareto front when the Pareto front is nonconvex. However, the use of a GA search method in MOGA causes the MOGA method to be much more computationally expensive than conventional multiobjective optimization using weight factors.

NCSUOpt: Nongradient-based optimizers from North Carolina State University, including DIRECT and, eventually, impicit filtering (web site: `http://www4.ncsu.edu/~ctk/matlab_darts.html`). We currently incorporate only an implementation of the DIRECT (DIviding RECTangles) algorithm [50]. While this is somewhat redundant with DIRECT supplied by Coliny, we have found that NCCSU DIRECT performs better in some cases, and presently we maintain both versions in DAKOTA.

NLPQLP: Methods for gradient-based constrained and unconstrained optimization problems using a sequential quadratic programming (SQP) algorithm [117]. NLPQLP is a commercial software product of Prof. Klaus Schittkowski (web site: `http://www.uni-bayreuth.de/departments/math/~kschittkowski/nlpqlp20.htm`). *Users may obtain their own copy of NLPQLP and compile it with the DAKOTA source code by following the steps given in the file* `Dakota/INSTALL`.

NPSOL: Methods for gradient-based constrained and unconstrained optimization problems using a sequential quadratic programming (SQP) algorithm [55]. NPSOL is a commercial software product of Stanford University (web site: www.sbsi-sol-optimize.com). Sandia National Laboratories, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory all have site licenses for NPSOL. *Other users may obtain their own copy of NPSOL and compile it with the DAKOTA source code by following the steps given in the file* `Dakota/INSTALL`.

OPT++: Methods for gradient-based and nongradient-based optimization of unconstrained, bound-constrained, and nonlinearly constrained optimization problems [92]. OPT++ includes a variety of Newton-based methods (quasi-Newton, finite-difference Newton, Gauss-Newton, and full-Newton), as well as the Polak-Ribeire CG method and the parallel direct search (PDS) method. OPT++ now contains a nonlinear interior point algorithm for handling general constraints. OPT++ is available to the public under the GNU LGPL and the source code is included with DAKOTA (web page: `http://csmr.ca.sandia.gov/opt++`).

PICO (Parallel Integer Combinatorial Optimization): PICO's branch-and-bound algorithm can be applied to nonlinear optimization problems involving discrete variables or a combination of continuous and discrete variables [28]. The discrete variables must be noncategorical (see Section 12.2.2). PICO is available to the public under the GNU LGPL (web page: `http://www.cs.sandia.gov/PICO`) and the source code is included with DAKOTA as part of the Acro package. Notes: (1) PICO's linear programming solvers are not included with DAKOTA, (2) PICO is being migrated into COLINY and is not operational in DAKOTA 4.2.

Additional information on these methods is provided in Chapter 7, as is a facility for using other solvers made available to DAKOTA via shared libraries.

3.6 Additional Optimization Capabilities

The optimization software packages described above provide algorithms to handle a wide variety of optimization problems. This includes algorithms for constrained and unconstrained optimization, as well as algorithms for gradient-based and nongradient-based optimization. Listed below are additional optimization capabilities that are available in DAKOTA.

Multiobjective Optimization: There are three capabilities for multiobjective optimization in DAKOTA. First, there is the MOGA capability described previously in Section 3.5. This is a specialized algorithm capability. The second capability involves the use of response data transformations to recast a multiobjective problem as a single-objective problem. Currently, DAKOTA supports the weighting factor approach for this transformation, in which a composite objective function is constructed from a set of individual objective functions using a user-

specified set of weighting factors. This approach is optimization algorithm independent, in that it works with any of the optimization methods listed in Section 3.5. Constraints are not affected by the weighting factor mapping; therefore, both constrained and unconstrained multiobjective optimization problems can be formulated and solved with DAKOTA, assuming selection of an appropriate constrained or unconstrained single-objective optimization algorithm. Future multiobjective response data transformations for goal programming, normal boundary intersection, etc. are planned. The third capability is the Pareto-set optimization strategy described in Section 3.9. This capability also utilizes the multiobjective response data transformations to allow optimization algorithm independence; however, it builds upon the basic approach by computing sets of optima in order to generate a Pareto trade-off surface.

User-Specified or Automatic Scaling: Some optimization algorithms are sensitive to the relative scaling of problem inputs and outputs. With any optimizer or least squares solver, user-specified (and in some cases automatic or logarithmic) scaling may be applied to continuous design variables, responses (objectives or residuals), nonlinear inequality and equality constraints, and/or linear inequality and equality constraints.

Additional information on these capabilities is provided in Chapter 7.

3.7 Nonlinear Least Squares for Parameter Estimation

Nonlinear least squares methods are optimization algorithms which exploit the special structure of a least squares objective function (see Section 1.4.2). These problems commonly arise in parameter estimation and test/analysis reconciliation. In practice, least squares solvers will tend to converge more rapidly than general-purpose optimization algorithms when the residual terms in the least squares formulation tend towards zero at the solution. Least squares solvers may experience difficulty when the residuals at the solution are significant, although experience has shown that the NL2SOL method can handle some problems that are highly nonlinear and have nonzero residuals at the solution.

NL2SOL: The NL2SOL algorithm [20] uses a secant-based algorithm to solve least-squares problems. In practice, it is more robust to nonlinear functions and nonzero residuals than conventional Gauss-Newton algorithms.

Gauss-Newton: DAKOTA's Gauss-Newton algorithm utilizes the Hessian approximation described in Section 1.4.2. The exact objective function value, exact objective function gradient, and the approximate objective function Hessian are defined from the least squares term values and gradients and are passed to the full-Newton optimizer from the OPT++ software package. As for all of the Newton-based optimization algorithms in OPT++, unconstrained, bound-constrained, and generally-constrained problems are supported. However, for the generally-constrained case, a derivative order mismatch exists in that the nonlinear interior point full Newton algorithm will require second-order information for the nonlinear constraints whereas the Gauss-Newton approximation only requires first order information for the least squares terms.

NLSSOL: The NLSSOL algorithm is a commercial software product of Stanford University (web site: `http://www.sbsi-sol-optimize.com`) that is bundled with current versions of the NPSOL library. It uses an SQP-based approach to solve generally-constrained nonlinear least squares problems. It periodically employs the Gauss-Newton Hessian approximation to accelerate the search. It requires only first-order information for the least squares terms and nonlinear constraints. Sandia National Laboratories, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory all have site licenses for NLSSOL. *Other users may obtain their own copy of NLSSOL and compile it with the DAKOTA source code by following the NPSOL installation steps given in the file* `Dakota/INSTALL`.

Additional information on these methods is provided in Chapter 8.

3.8 Surrogate-Based Minimization

Surrogate-Based Local Minimization: This method combines the design of experiments methods, surrogate models, and optimization capabilities of DAKOTA. In SBO, the optimization algorithm operates on a surrogate model instead of directly operating on the computationally expensive simulation model. The surrogate model can be formed from data fitting methods (local, multipoint, or global), from a lower fidelity version of the computational model, or from a mathematically-generated reduced-order model (see Section 3.10). For each of these surrogate model types, the SBO algorithm periodically validates the progress using the surrogate model against the original high-fidelity model. The SBO strategy in DAKOTA can be configured to employ heuristic rules (less expensive) or to be provably convergent to the optimum of the original model (more expensive). The development of SBO strategies is an area of active research in the DAKOTA project.

Surrogate-Based Global Minimization: Similar to surrogate-based local minimization, this method combines design of experiments and surrogate modeling with optimization. However, rather than employing trust region model management to localize and control the extent of the approximation in order to ensure convergence to a local minimum, the surrogate-based global method sequentially refines the full range of a global approximation using global optimizers.

Efficient Global Minimization: Methods for nongradient-based constrained and unconstrained optimization and nonlinear least squares based on Gaussian process models. This approach uses an expected improvement function derived from the expected value and variance estimators in Gaussian process models, and is designed to balance exploitation of regions with good solutions and exploration of regions with limited data. [81]

Additional information on these methods is provided in Chapter 9.

3.9 Optimization Strategies

Due to the flexibility of DAKOTA's object-oriented design, it is relatively easy to create algorithms that combine several of DAKOTA's capabilities. These algorithms are referred to as *strategies*:

Multilevel Hybrid Optimization: This strategy allows the user to specify a sequence of optimization methods, with the results from one method providing the starting point for the next method in the sequence. An example which is useful in many engineering design problems involves the use of a nongradient-based global optimization method (e.g., genetic algorithm) to identify a promising region of the parameter space, which feeds its results into a gradient-based method (quasi-Newton, SQP, etc.) to perform an efficient local search for the optimum point.

Multistart Local Optimization: This strategy uses many local optimization runs (often gradient-based), each of which is started from a different initial point in the parameter space. This is an attractive strategy in situations where multiple local optima are known to exist or may potentially exist in the parameter space. This approach combines the efficiency of local optimization methods with the parameter space coverage of a global stratification technique.

Pareto-Set Optimization: The Pareto-set optimization strategy allows the user to specify different sets of weights for the individual objective functions in a multiobjective optimization problem. DAKOTA executes each of these weighting sets as a separate optimization problem, serially or in parallel, and then outputs the set of optimal designs which define the Pareto set. Pareto set information can be useful in making trade-off decisions in engineering design problems. *[Refer to 3.6 for additional information on multiobjective optimization methods.]*

Mixed Integer Nonlinear Programming (MINLP): This strategy uses the branch and bound capabilities of the PICO package to perform optimization on problems that have both discrete and continuous design variables. PICO provides a branch and bound engine targeted at mixed integer linear programs (MILP), which when combined with

DAKOTA's nonlinear optimization methods, results in a MINLP capability. In addition, the multiple NLPs solved within MINLP provide an opportunity for concurrent execution of multiple optimizations.

These strategies are covered in more detail in Chapter 10.

3.10 Surrogate Models

Surrogate models are inexpensive approximate models that are intended to capture the salient features of an expensive high-fidelity model. They can be used to explore the variations in response quantities over regions of the parameter space, or they can serve as inexpensive stand-ins for optimization or uncertainty quantification studies (see, for example, the surrogate-based optimization strategy in Section 3.9). The surrogate models supported in DAKOTA can be categorized into three types: data fits, multifidelity, and reduced-order model surrogates.

Data fitting methods involve construction of an approximation or surrogate model using data (response values, gradients, and Hessians) generated from the original truth model. Data fit methods can be further categorized as local, multipoint, and global approximation techniques, based on the number of points used in generating the data fit. Local methods involve response data from a single point in parameter space. Available techniques currently include:

Taylor Series Expansion: This is a local first-order or second-order expansion centered at a single point in the parameter space.

Multipoint approximations involve response data from two or more points in parameter space, often involving the current and previous iterates of a minimization algorithm. Available techniques currently include:

TANA-3: This multipoint approximation uses a two-point exponential approximation [146, 43] built with response value and gradient information from the current and previous iterates.

Global methods, often referred to as *response surface methods*, involve many points spread over the parameter ranges of interest. These surface fitting methods work in conjunction with the sampling methods and design of experiments methods described in Section 3.3.

Polynomial Regression: First-order (linear), second-order (quadratic), and third-order (cubic) polynomial response surfaces computed using linear least squares regression methods. Note: there is currently no use of forward- or backward-stepping regression methods to eliminate unnecessary terms from the polynomial model.

Kriging Interpolation: An implementation of spatial interpolation using kriging methods and Gaussian correlation functions [62]. The algorithm used in the kriging process generates a C^2-continuous surface that exactly interpolates the data values.

Gaussian Process (GP): Closely related to kriging, this technique is a spatial interpolation method that assumes the outputs of the simulation model follow a multivariate normal distribution. The implementation of a Gaussian process currently in DAKOTA assumes a constant mean function. The hyperparameters governing the covariance matrix are obtained through Maximum Likelihood Estimation (MLE). We also use a jitter term to better condition the covariance matrix, so the Gaussian process may not exactly interpolate the data values.

Artificial Neural Networks: An implementation of the stochastic layered perceptron neural network developed by Prof. D. C. Zimmerman of the University of Houston [147]. This neural network method is intended to have a lower training (fitting) cost than typical back-propagation neural networks.

Multivariate Adaptive Regression Splines (MARS): Software developed by Prof. J. H. Friedman of Stanford University [48]. The MARS method creates a C^2-continuous patchwork of splines in the parameter space.

DAKOTA also offers preliminary support for radial basis function (RBF) and moving least squares (MLS) ap-

proximations. In addition to data fit surrogates, DAKOTA supports multifidelity and reduced-order model approximations:

Multifidelity Surrogates: Multifidelity modeling involves the use of a low-fidelity physics-based model as a surrogate for the original high-fidelity model. The low-fidelity model typically involves a coarsher mesh, looser convergence tolerances, reduced element order, or omitted physics. It is a separate model in its own right and does not require data from the high-fidelity model for construction. Rather, the primary need for high-fidelity evaluations is for defining correction functions that are applied to the low-fidelity results.

Reduced Order Models: A reduced-order model (ROM) is mathematically derived from a high-fidelity model using the technique of Galerkin projection. By computing a set of basis functions (e.g., eigenmodes, left singular vectors) that capture the principal dynamics of a system, the original high-order system can be projected to a much smaller system, of the size of the number of retained basis functions.

Additional information on these surrogate methods is provided in Sections 11.4.1 through 11.4.3.

3.11 Nested Models

Nested models utilize a sub-iterator and a sub-model to perform a complete iterative study as part of every evaluation of the model. This sub-iteration accepts variables from the outer level, performs the sub-level analysis, and computes a set of sub-level responses which are passed back up to the outer level. The nested model constructs admit a wide variety of multi-iterator, multi-model solution approaches. For example, optimization within optimization (for hierarchical multidisciplinary optimization), uncertainty quantification within uncertainty quantification (for second-order probability), uncertainty quantification within optimization (for optimization under uncertainty), and optimization within uncertainty quantification (for uncertainty of optima) are all supported, with and without surrogate model indirection. Three important examples are highlighted: second-order probability, optimization under uncertainty, and surrogate-based uncertainty quantification.

Second-Order Probability: Second-order probability approaches employ nested models to embed one uncertainty quantification (UQ) within another. The outer level UQ is commonly linked to epistemic uncertainties (also known as reducible uncertainties) resulting from a lack of knowledge, and the inner UQ is commonly linked to aleatory uncertainties (also known as irreducible uncertainties) that are inherent in nature. The outer level generates sets of realizations, typically from sampling within interval distributions. These realizations define values for distribution parameters used in a probabilistic analysis for the inner level UQ. The term "second-order" derives from this use of distributions on distributions and the generation of statistics on statistics.

Optimization Under Uncertainty (OUU): Many real-world engineering design problems contain stochastic features and must be treated using OUU methods such as robust design and reliability-based design. For OUU, the uncertainty quantification methods of DAKOTA are combined with optimization algorithms. This allows the user to formulate problems where one or more of the objective and constraints are stochastic. Due to the computational expense of both optimization and UQ, the simple nesting of these methods in OUU can be computationally prohibitive for real-world design problems. For this reason, surrogate-based optimization under uncertainty (SBOUU), reliability-based design optimization (RBDO), polynomial chaos-based design optimization (PCBDO), and stochastic collocation-based design optimization (SCBDO) methods have been developed which can reduce the overall expense by orders of magnitude. OUU methods are an active research area.

Surrogate-Based Uncertainty Quantification (SBUQ): Since many uncertainty quantification (UQ) methods are computationally costly, requiring many function evaluations to obtain accurate estimates of moments or percentile values of an output distribution, one may wish to embed surrogate models within the UQ process in order to reduce expense. By evaluating the true function on a fixed, small set of samples and using these sample evaluations to create a response surface approximation (e.g. a surrogate model or meta-model) of the underlying "true" function,

the subsequent evaluation of the UQ results (using thousands or millions of samples) based on the approximation can obtain estimates of the mean, variance, and percentiles of the response at much lower overall cost.

Additional information on these nested approaches is provided in Sections 11.5-11.6.

3.12 Parallel Computing

The methods and strategies in DAKOTA are designed to exploit parallel computing resources such as those found in a desktop multiprocessor workstation, a network of workstations, or a massively parallel computing platform. This parallel computing capability is a critical technology for rendering real-world engineering design problems computationally tractable. DAKOTA employs the concept of *multilevel parallelism*, which takes simultaneous advantage of opportunities for parallel execution from multiple sources:

Parallel Simulation Codes: DAKOTA works equally well with both serial and parallel simulation codes.

Concurrent Execution of Analyses within a Function Evaluation: Some engineering design applications call for the use of multiple simulation code executions (different disciplinary codes, the same code for different load cases or environments, etc.) in order to evaluate a single response data set (e.g., abjective functions and constraints) for a single set of parameters. If these simulation code executions are independent (or if coupling is enforced at a higher level), DAKOTA can perform them in parallel.

Concurrent Execution of Function Evaluations within an Iterator: With very few exceptions, the iterative algorithms described in Section 3.2 through Section 3.7 all provide opportunities for the concurrent evaluation of response data sets for different parameter sets. Whenever there exists a set of design point evaluations that are independent, DAKOTA can perform them in parallel.

Concurrent Execution of Iterators within a Strategy: Some of the DAKOTA strategies described in Section 3.9 generate a sequence of iterator subproblems. For example, the MINLP, Pareto-set, and multi-start strategies generate sets of optimization subproblems, and the optimization under uncertainty strategy generates sets of uncertainty quantification subproblems. Whenever these subproblems are independent, DAKOTA can perform them in parallel.

It is important to recognize that these four parallelism levels are nested, in that a strategy can schedule and manage concurrent iterators, each of which may manage concurrent function evaluations, each of which may manage concurrent analyses, each of which may execute on multiple processors. Additional information on parallel computing with DAKOTA is provided in Chapter 18.

3.13 Summary

DAKOTA is both a production tool for engineering design and analysis activities and a research tool for the development of new algorithms in optimization, uncertainty quantification, and related areas. Because of the extensible, object-oriented design of DAKOTA, it is relatively easy to add new iterative algorithms, strategies, simulation interfacing approaches, surface fitting methods, etc. In addition, DAKOTA can serve as a rapid prototyping tool for algorithm development. That is, by having a broad range of building blocks available (i.e., parallel computing, surrogate models, simulation interfaces, fundamental algorithms, etc.), new capabilities can be assembled rapidly which leverage the previous software investments. For additional discussion on framework extensibility, refer to the DAKOTA Developers Manual [31].

The capabilities of DAKOTA have been used to solve engineering design and optimization problems at Sandia Labs, at other Department of Energy labs, and by our industrial and academic collaborators. Often, this real-world

experience has provided motivation for research into new areas of optimization. The DAKOTA development team welcomes feedback on the capabilities of this software toolkit, as well as suggestions for new areas of research.

Chapter 4

Parameter Study Capabilities

4.1 Overview

Parameter study methods in the DAKOTA toolkit involve the computation of response data sets at a selection of points in the parameter space. These response data sets are not linked to any specific interpretation, so they may consist of any allowable specification from the responses keyword block, i.e., objective and constraint functions, least squares terms and constraints, or generic response functions. This allows the use of parameter studies in direct coordination with optimization, least squares, and uncertainty quantification studies without significant modification to the input file. In addition, response data sets are not restricted to function values only; gradients and Hessians of the response functions can also be catalogued by the parameter study. This allows for several different approaches to "sensitivity analysis": (1) the variation of function values over parameter ranges provides a global assessment as to the sensitivity of the functions to the parameters, (2) derivative information can be computed numerically, provided analytically by the simulator, or both (mixed gradients) in directly determining local sensitivity information at a point in parameter space, and (3) the global and local assessments can be combined to investigate the variation of derivative quantities through the parameter space by computing sensitivity information at multiple points.

In addition to sensitivity analysis applications, parameter studies can be used for investigating nonsmoothness in simulation response variations (so that models can be refined or finite difference step sizes can be selected for computing numerical gradients), interrogating problem areas in the parameter space, or performing simulation code verification (verifying simulation robustness) through parameter ranges of interest. A parameter study can also be used in coordination with minimization methods as either a pre-processor (to identify a good starting point) or a post-processor (for post-optimality analysis).

Parameter study methods will iterate any combination of *continuous* design, uncertain, and state variables into any set of responses (any function, gradient, and Hessian definition). Parameter studies draw no distinction between the different types of continuous variables (design, uncertain, or state) and the different types of response functions. They simply pass all of the variables defined in the variables specification into the interface, from which they expect to retrieve all of the responses defined in the responses specification. As described in Section 14.3, when gradient and/or Hessian information is being catalogued in the parameter study, it is assumed that derivative components will be computed with respect to all of the *continuous* variables (continuous design, uncertain, and continuous state variables) specified. Parameter studies over discrete variables will be supported in the future, although response derivatives with respect to these variables are not defined.

DAKOTA currently supports four types of parameter studies. Vector parameter studies compute response data

sets at selected intervals along an n-dimensional vector in parameter space. List parameter studies compute response data sets at a list of points in parameter space, defined by the user. A centered parameter study computes multiple coordinate-based parameter studies, one per parameter, centered about the initial parameter values. A multidimensional parameter study computes response data sets for an n-dimensional hypergrid of points. More detail on these parameter studies is found in Sections 4.2 through 4.5 below.

4.1.1 Initial Values

The vector and centered parameter studies use the initial values of the variables from the `variables` keyword block as the starting point and the central point of the parameter studies, respectively. In the case of design variables, the `initial_point` is used. In the case of state variables, the `initial_state` is used. In the case of uncertain variables, initial values for variables with normal, lognormal, uniform, loguniform, triangular, beta, gamma, gumbel, frechet, and weibull probability distributions are the means of the distributions, and for the histogram and interval distribution, are the bin/point/interval lower bounds. These parameter study starting values for design, uncertain, and state variables are referenced in the following sections using the identifier "Initial Values."

4.1.2 Bounds

The multidimensional parameter study uses the bounds of the variables from the `variables` keyword block to define the range of parameter values to study. In the case of design and state variables, the `lower_bounds` and `upper_bounds` specifications are used. In the case of uncertain variables, bounds for variables with normal and lognormal distributions are the optional distribution bounds (user-specified or default), for uniform, loguniform, triangular, and beta distributions, are the required distribution bounds (user-specified), and for the histogram and interval distributions, are the bin/point/interval lower and upper bounds. For the remaining distributions, parameter study bounds are inferred using $[0, \mu + 3\sigma]$ for gamma, frechet, and weibull, and $[\mu - 3\sigma, \mu + 3\sigma]$ for gumbel.

4.2 Vector Parameter Study

The vector parameter study computes response data sets at selected intervals along an n-dimensional vector in parameter space. This capability encompasses both single-coordinate parameter studies (to study the effect of a single variable on a response set) as well as multiple coordinate vector studies (to investigate the response variations along some arbitrary vector; e.g., to investigate a search direction failure). In addition to these uses, this capability is used recursively within the implementation of the multidimensional parameter study.

DAKOTA's vector parameter study includes three possible specification formulations which are used in conjunction with the Initial Values (see Section 4.1.1) to define the vector and steps of the parameter study:

```
final_point (vector of reals) and step_length (real)
final_point (vector of reals) and num_steps (integer)
step_vector (vector of reals) and num_steps (integer)
```

In each of these three cases, the Initial Values are used as the parameter study starting point and the specification selected from the three above defines the orientation and length of the vector as well as the increments to be evaluated along the vector. Several examples starting from Initial Values of `1.0, 1.0, 1.0` are included below:

`final_point = 1.0, 2.0, 1.0` and `step_length = .4`:

```
    Parameters for function evaluation 1:
                         1.0000000000e+00 d1
                         1.0000000000e+00 d2
                         1.0000000000e+00 d3
    Parameters for function evaluation 2:
                         1.0000000000e+00 d1
                         1.4000000000e+00 d2
                         1.0000000000e+00 d3
    Parameters for function evaluation 3:
                         1.0000000000e+00 d1
                         1.8000000000e+00 d2
                         1.0000000000e+00 d3
```

`final_point = 2.0, 2.0, 2.0` and `step_length = .4` (note that `step_length` defines Cartesian distance of the step and the steps continue up to but not past the `final_point`):

```
    Parameters for function evaluation 1:
                         1.0000000000e+00 d1
                         1.0000000000e+00 d2
                         1.0000000000e+00 d3
    Parameters for function evaluation 2:
                         1.2309401077e+00 d1
                         1.2309401077e+00 d2
                         1.2309401077e+00 d3
    Parameters for function evaluation 3:
                         1.4618802154e+00 d1
                         1.4618802154e+00 d2
                         1.4618802154e+00 d3
    Parameters for function evaluation 4:
                         1.6928203230e+00 d1
                         1.6928203230e+00 d2
                         1.6928203230e+00 d3
    Parameters for function evaluation 5:
                         1.9237604307e+00 d1
                         1.9237604307e+00 d2
                         1.9237604307e+00 d3
```

`final_point = 2.0, 2.0, 2.0` and `num_steps = 4`:

```
    Parameters for function evaluation 1:
                         1.0000000000e+00 d1
                         1.0000000000e+00 d2
                         1.0000000000e+00 d3
    Parameters for function evaluation 2:
                         1.2500000000e+00 d1
                         1.2500000000e+00 d2
                         1.2500000000e+00 d3
    Parameters for function evaluation 3:
                         1.5000000000e+00 d1
                         1.5000000000e+00 d2
                         1.5000000000e+00 d3
```

```
    Parameters for function evaluation 4:
                         1.7500000000e+00 d1
                         1.7500000000e+00 d2
                         1.7500000000e+00 d3
    Parameters for function evaluation 5:
                         2.0000000000e+00 d1
                         2.0000000000e+00 d2
                         2.0000000000e+00 d3
```

`step_vector = .1, .1, .1` and `num_steps = 4`:

```
    Parameters for function evaluation 1:
                         1.0000000000e+00 d1
                         1.0000000000e+00 d2
                         1.0000000000e+00 d3
    Parameters for function evaluation 2:
                         1.1000000000e+00 d1
                         1.1000000000e+00 d2
                         1.1000000000e+00 d3
    Parameters for function evaluation 3:
                         1.2000000000e+00 d1
                         1.2000000000e+00 d2
                         1.2000000000e+00 d3
    Parameters for function evaluation 4:
                         1.3000000000e+00 d1
                         1.3000000000e+00 d2
                         1.3000000000e+00 d3
    Parameters for function evaluation 5:
                         1.4000000000e+00 d1
                         1.4000000000e+00 d2
                         1.4000000000e+00 d3
```

4.3 List Parameter Study

The list parameter study computes response data sets at selected points in parameter space. These points are explicitly specified by the user and are not confined to lie on any line or surface. Thus, this parameter study provides a general facility that supports the case where the desired set of points to evaluate does not fit the prescribed structure of the vector, centered, or multidimensional parameter studies.

The user input consists of a `list_of_points` specification which lists the requested parameter sets in succession. The list parameter study simply performs a simulation for the first parameter set (the first n entries in the list), followed by a simulation for the next parameter set (the next n entries), and so on, until the list of points has been exhausted. Since the Initial Values will not be used, they need not be specified.

An example specification which would result in the same parameter sets as in the first example in Section 4.2 would be:

```
    list_of_points = 1.0, 1.0, 1.0, 1.0, 1.4, 1.0, 1.0, 1.8, 1.0
```

4.4 Centered Parameter Study

The centered parameter study executes multiple coordinate-based parameter studies, one per parameter, centered about the specified Initial Values. This is useful for investigation of function contours in the vicinity of a specific point. For example, after computing an optimum design, this capability could be used for post-optimality analysis in verifying that the computed solution is actually at a minimum or constraint boundary and in investigating the shape of this minimum or constraint boundary.

This method requires `percent_delta` (real) and `deltas_per_variable` (integer) specifications, where the former specifies the size of the increments in percent and the latter specifies the number of increments per variable in each of the plus and minus directions.

For example, with Initial Values of `1.0, 1.0`, a `percent_delta` of `10.0`, and a `deltas_per_variable` of `2`, the center point is evaluated followed by four function evaluations (two minus deltas and two plus deltas) per variable:

```
Parameters for function evaluation 1:
                     1.0000000000e+00 cdv_1
                     1.0000000000e+00 cdv_2
Parameters for function evaluation 2:
                     8.0000000000e-01 cdv_1
                     1.0000000000e+00 cdv_2
Parameters for function evaluation 3:
                     9.0000000000e-01 cdv_1
                     1.0000000000e+00 cdv_2
Parameters for function evaluation 4:
                     1.1000000000e+00 cdv_1
                     1.0000000000e+00 cdv_2
Parameters for function evaluation 5:
                     1.2000000000e+00 cdv_1
                     1.0000000000e+00 cdv_2
Parameters for function evaluation 6:
                     1.0000000000e+00 cdv_1
                     8.0000000000e-01 cdv_2
Parameters for function evaluation 7:
                     1.0000000000e+00 cdv_1
                     9.0000000000e-01 cdv_2
Parameters for function evaluation 8:
                     1.0000000000e+00 cdv_1
                     1.1000000000e+00 cdv_2
Parameters for function evaluation 9:
                     1.0000000000e+00 cdv_1
                     1.2000000000e+00 cdv_2
```

This set of points in parameter space is depicted in Figure 4.1.

If the Initial Values for the centered parameter study are very small or equal to zero, the study will substitute a default step size. This is necessary due to the relative nature of the `percent_delta` specification.

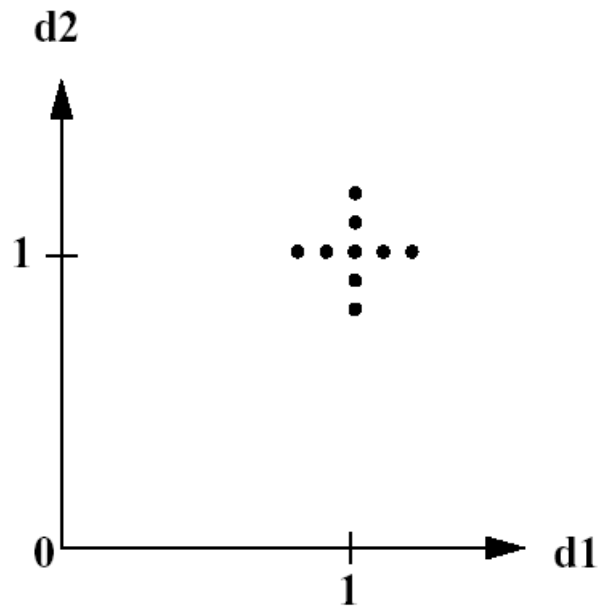


Figure 4.1: Example centered parameter study.

4.5 Multidimensional Parameter Study

The multidimensional parameter study computes response data sets for an n-dimensional hypergrid of points. Each continuous variable is partitioned into equally spaced intervals between its upper and lower bounds (see Section 4.1.2), and each combination of the values defined by these partitions is evaluated. The number of function evaluations performed in the study is:

$$\prod_{i=1}^{n}(\texttt{partitions}_i + 1) \tag{4.1}$$

The partitions information is specified using the `partitions` specification, which provides an integer list of the number of partitions for each continuous variable (i.e., $\texttt{partitions}_i$). Since the Initial Values will not be used, they need not be specified.

In a two variable example problem with `d1` $\in$ `[0,2]` and `d2` $\in$ `[0,3]` (as defined by the upper and lower bounds from the variables specification) and with `partitions = 2,3`, the interval `[0,2]` is divided into two equal-sized partitions and the interval `[0,3]` is divided into three equal-sized partitions. This two-dimensional grid, shown in Figure 4.2, would result in the following twelve function evaluations:

```
Parameters for function evaluation 1:
                     0.0000000000e+00 d1
                     0.0000000000e+00 d2
Parameters for function evaluation 2:
                     1.0000000000e+00 d1
                     0.0000000000e+00 d2
Parameters for function evaluation 3:
```

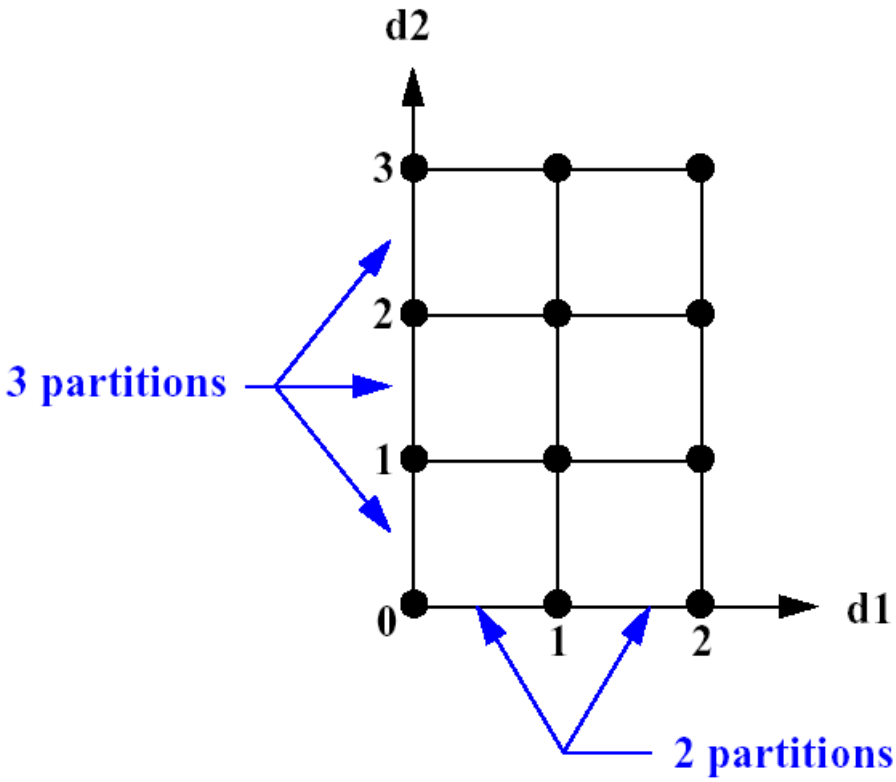


Figure 4.2: Example multidimensional parameter study

```
                      2.0000000000e+00 d1
                      0.0000000000e+00 d2
Parameters for function evaluation 4:
                      0.0000000000e+00 d1
                      1.0000000000e+00 d2
Parameters for function evaluation 5:
                      1.0000000000e+00 d1
                      1.0000000000e+00 d2
Parameters for function evaluation 6:
                      2.0000000000e+00 d1
                      1.0000000000e+00 d2
Parameters for function evaluation 7:
                      0.0000000000e+00 d1
                      2.0000000000e+00 d2
Parameters for function evaluation 8:
                      1.0000000000e+00 d1
                      2.0000000000e+00 d2
Parameters for function evaluation 9:
                      2.0000000000e+00 d1
                      2.0000000000e+00 d2
Parameters for function evaluation 10:
                      0.0000000000e+00 d1
                      3.0000000000e+00 d2
Parameters for function evaluation 11:
                      1.0000000000e+00 d1
                      3.0000000000e+00 d2
Parameters for function evaluation 12:
                      2.0000000000e+00 d1
                      3.0000000000e+00 d2
```

Chapter 5

Design of Experiments Capabilities

5.1 Overview

DAKOTA contains several software packages for sampling and design of experiments: LHS (Latin hypercube sampling), DDACE (distributed design and analysis for computer experiments), FSUDace (Florida State University's Design and Analysis of Computer Experiments package), and PSUADE (Problem Solving Environment for Uncertainty Analysis and Design Exploration). LHS [125] is a general-purpose sampling package developed at Sandia that has been used by the DOE national labs for several decades. DDACE is a more recent package for computer experiments that is under development by staff at Sandia Labs [127]. DDACE provides the capability for generating orthogonal arrays, Box-Behnken designs, Central Composite designs, and random designs. The FSUDace package provides the following sampling techniques: quasi-Monte Carlo sampling based on Halton or Hammersley sequences, and Centroidal Voronoi Tessellation. Lawrence Livermore National Lab's PSUADE [126] includes several methods for model exploration, but only the Morris screeing method is exposed in DAKOTA.

This chapter describes DDACE, FSUDace, and PSUADE, with the primary goal of designing computer experiments. Latin Hypercube Sampling, also used in uncertainty quantification, is discussed in Section 6.2. The differences between sampling used in design of experiments and sampling used in uncertainty quantification is discussed in more detail in the following paragraphs. In brief, we consider design of experiment methods to generate sets of uniform random variables on the interval $[0, 1]$. These sets are mapped to the lower/upper bounds of the problem variables and then the response functions are evaluated at the sample input points with the goal of characterizing the behavior of the response functions over the input parameter ranges of interest. Uncertainty quantification via LHS sampling, in contrast, involves characterizing the uncertain input variables with probability distributions such as normal, Weibull, triangular, etc., sampling from the input distributions, and propagating the input uncertainties to obtain a cumulative distribution function on the output. There is significant overlap between design of experiments and sampling. Often, both techniques can be used to obtain similar results about the behavior of the response functions and about the relative importance of the input variables.

5.2 Design of Computer Experiments

Computer experiments are often different from physical experiments, such as those performed in agriculture, manufacturing, or biology. In physical experiments, one often applies the same *treatment* or *factor level* in an

experiment several times to get an understanding of the variability of the output when that treatment is applied. For example, in an agricultural experiment, several fields (e.g., 8) may be subject to a low level of fertilizer and the same number of fields may be subject to a high level of fertilizer to see if the amount of fertilizer has a significant effect on crop output. In addition, one is often interested in the variability of the output within a treatment group: is the variability of the crop yields in the low fertilizer group much higher than that in the high fertilizer group, or not?

In physical experiments, the process we are trying to examine is stochastic: that is, the same treatment may result in different outcomes. By contrast, in computer experiments, often we have a deterministic code. If we run the code with a particular set of input parameters, the code will always produce the same output. There certainly are stochastic codes, but the main focus of computer experimentation has been on deterministic codes. Thus, in computer experiments we often do not have the need to do replicates (running the code with the exact same input parameters several times to see differences in outputs). Instead, a major concern in computer experiments is to create an experimental design which can sample a high-dimensional space in a representative way with a minimum number of samples. The number of factors or parameters that we wish to explore in computer experiments is usually much higher than physical experiments. In physical experiments, one may be interesting in varying a few parameters, usually five or less, while in computer experiments we often have dozens of parameters of interest. Choosing the levels of these parameters so that the samples adequately explore the input space is a challenging problem. There are many experimental designs and sampling methods which address the issue of adequate and representative sample selection. Classical experimental designs which are often used in physical experiments include Central Composite designs and Box-Behnken designs.

There are many goals of running a computer experiment: one may want to explore the input domain or the design space and get a better understanding of the range in the outputs for a particular domain. Another objective is to determine which inputs have the most influence on the output, or how changes in the inputs change the output. This is usually called *sensitivity analysis*. Another goal is to compare the relative importance of model input uncertainties on the uncertainty in the model outputs, *uncertainty analysis*. Yet another goal is to use the sampled inputs points and their corresponding output to create a *response surface approximation* for the computer code. The response surface approximation (e.g., a polynomial regression model, a kriging model, a neural net) can then be used to emulate the computer code. Constructing a response surface approximation is particularly important for applications where running a computational model is extremely expensive: the computer model may take 10 or 20 hours to run on a high performance machine, whereas the response surface model may only take a few seconds. Thus, one often optimizes the response surface model or uses it within a framework such as surrogate-based optimization. Response surface models are also valuable in cases where the gradient (first derivative) and/or Hessian (second derivative) information required by optimization techniques are either not available, expensive to compute, or inaccurate because the derivatives are poorly approximated or the function evaluation is itself noisy due to roundoff errors. Furthermore, many optimization methods require a good initial point to ensure fast convergence or to converge to good solutions (e.g. for problems with multiple local minima). Under these circumstances, a good design of computer experiment framework coupled with response surface approximations can offer great advantages.

In addition to the sensitivity analysis, uncertainty analysis, and response surface modeling mentioned above, we also may want to do *uncertainty quantification* on a computer model. Uncertainty quantification (UQ) refers to taking a particular set of distributions on the inputs, and propagating them through the model to obtain a distribution on the outputs. For example, if input parameter A follows a normal with mean 5 and variance 1, the computer produces a random draw from that distribution. If input parameter B follows a weibull distribution with alpha = 0.5 and beta = 1, the computer produces a random draw from that distribution. When all of the uncertain variables have samples drawn from their input distributions, we run the model with the sampled values as inputs. We do this repeatedly to build up a distribution of outputs. We can then use the cumulative distribution function of the output to ask questions such as: what is the probability that the output is greater than 10? What is the 99th percentile of the output?

Note that sampling-based uncertainty quantification and design of computer experiments are very similar. *There is significant overlap* in the purpose and methods used for UQ and for DACE. We have attempted to delineate the differences within DAKOTA as follows: we use the methods DDACE, FSUDACE, and PSUADE primarily for design of experiments, where we are interested in understanding the main effects of parameters and where we want to sample over an input domain to obtain values for constructing a response surface. We use the nondeterministic sampling methods (`nond_sampling`) for uncertainty quantification, where we are propagating specific input distributions and interested in obtaining (for example) a cumulative distribution function on the output. If one has a problem with no distributional information, we recommend starting with a design of experiments approach. Note that DDACE, FSUDACE, and PSUADE currently do *not* support distributional information: they take an upper and lower bound for each uncertain input variable and sample within that. The uncertainty quantification methods in `nond_sampling` (primarily Latin Hypercube sampling) offer the capability to sample from many distributional types. The distinction between UQ and DACE is somewhat arbitrary: both approaches often can yield insight about important parameters and both can determine sample points for response surface approximations.

5.3 DDACE Background

The DACE package includes both classical design of experiments methods [127] and stochastic sampling methods. The classical design of experiments methods in DDACE are central composite design (CCD) and Box-Behnken (BB) sampling. A grid-based sampling method also is available. The stochastic methods are orthogonal array sampling [84], Monte Carlo (random) sampling, and Latin hypercube sampling. Note that the DDACE version of LHS available through the DAKOTA interface only supports uniform distributions. DDACE does not currently support enforcement of user-specified correlation structure among the variables.

The sampling methods in DDACE can be used alone or in conjunction with other methods. For example, DDACE sampling can be used with both the surrogate-based optimization strategy and the optimization under uncertainty strategy. See Figure 11.5 for an example of how the DDACE settings are used in DAKOTA.

More information on DDACE is available on the web at: `http://csmr.ca.sandia.gov/projects/ddace`

The following sections provide more detail about the sampling methods available for design of experiments in DDACE.

5.3.1 Central Composite Design

A Box-Wilson Central Composite Design, commonly called a central composite design (CCD), contains an embedded factorial or fractional factorial design with center points that is augmented with a group of 'star points' that allow estimation of curvature. If the distance from the center of the design space to a factorial point is ± 1 unit for each factor, the distance from the center of the design space to a star point is $\pm\alpha$ with $\mid \alpha \mid > 1$. The precise value of depends on certain properties desired for the design and on the number of factors involved. The CCD design is specified in DAKOTA with the method command `dace central_composite`.

As an example, with two input variables or factors, each having two levels, the factorial design is shown in Table 5.1 .

With a CCD, the design in Table 5.1 would be augmented with the following points shown in Table 5.2 if α = 1.3. These points define a circle around the original factorial design.

Note that the number of samples points specified in a CCD,`samples`, is a function of the number of variables in

Table 5.1: Simple Factorial Design

Input 1	Input 2
-1	-1
-1	+1
+1	-1
+1	+1

Table 5.2: Additional Points to make the factorial design a CCD

Input 1	Input 2
0	+1.3
0	-1.3
1.3	0
-1.3	0
0	0

the problem:

$$samples = 1 + 2 * NumVar + 2^{NumVar}$$

5.3.2 Box-Behnken Design

The Box-Behnken design is similar to a Central Composite design, with some differences. The Box-Behnken design is a quadratic design in that it does not contain an embedded factorial or fractional factorial design. In this design the treatment combinations are at the midpoints of edges of the process space and at the center, as compared with CCD designs where the extra points are placed at 'star points' on a circle outside of the process space. Box-Behken designs are rotatable (or near rotatable) and require 3 levels of each factor. The designs have limited capability for orthogonal blocking compared to the central composite designs. Box-Behnken requires fewer runs than CCD for 3 factors, but this advantage goes away as the number of factors increases. The Box-Behnken design is specified in DAKOTA with the method command `dace box_behnken`.

Note that the number of samples points specified in a Box-Behnken design, `samples`, is a function of the number of variables in the problem:

$$samples = 1 + 4 * NumVar + (NumVar - 1)/2$$

5.3.3 Orthogonal Array Designs

Orthogonal array (OA) sampling is a widely used technique for running experiments and systematically testing factor effects [72]. An orthogonal array sample can be described as an 4-tuple (m, n, s, r), where m is the number of sample points, n is the number of input variables, s is the number of symbols, and r is the strength of the orthogonal array. The number of sample points, m, must be a multiple of the number of symbols, s. The

number of symbols refers to the number of levels per input variable. The strength refers to the number of columns where we are guaranteed to see all the possibilities an equal number of times.

For example, Table 5.3 shows an orthogonal array of strength 2 for m = 8, with 7 variables:

Table 5.3: Orthogonal Array for Seven Variables

Input 1	Input 2	Input 3	Input 4	Input 5	Input 6	Input 7
0	0	0	0	0	0	0
0	0	0	1	1	1	1
0	1	1	0	0	1	1
0	1	1	1	1	0	0
1	0	1	0	1	0	1
1	0	1	1	0	1	0
1	1	0	0	1	1	0
1	1	0	1	0	0	1

If one picks any two columns, say the first and the third, note that each of the four possible rows we might see there, 0 0, 0 1, 1 0, 1 1, appears exactly the same number of times, twice in this case.

DDACE creates orthogonal arrays of strength 2. Further, the OAs generated by DDACE do not treat the factor levels as one fixed value (0 or 1 in the above example). Instead, once a level for a variable is determined in the array, DDACE samples a random variable from within that level. The orthogonal array design is specified in DAKOTA with the method command `dace oas`.

The orthogonal array method in DDACE is the only method that allows for the calculation of main effects, specified with the command `main_effects`. Main effects is a sensitivity analysis method which identifies the input variables that have the most influence on the output. In main effects, the idea is to look at the mean of the response function when variable A (for example) is at level 1 vs. when variable A is at level 2 or level 3. If these mean responses of the output are statistically significantly different at different levels of variable A, this is an indication that variable A has a significant effect on the response. The orthogonality of the columns is critical in performing main effects analysis, since the column orthogonality means that the effects of the other variables 'cancel out' when looking at the overall effect from one variable at its different levels. There are ways of developing orthogonal arrays to calculate higher order interactions, such as two-way interactions (what is the influence of Variable A * Variable B on the output?), but this is not available in DDACE currently. At present, one way interactions are supported in the calculation of orthogonal array main effects within DDACE. The main effects are presented as a series of ANOVA tables. For each objective function and constraint, the decomposition of variance of that objective or constraint is presented as a function of the input variables. The p-value in the ANOVA table is used to indicate if the input factor is significant. The p-value is the probability that you would have obtained samples more extreme than you did if the input factor has no effect on the response. For example, if you set a level of significance at 0.05 for your p-value, and the actual p-value is 0.03, then the input factor has a significant effect on the response.

5.3.4 Grid Design

In a grid design, a grid is placed over the input variable space. This is very similar to a multi-dimensional parameter study where the samples are taken over a set of partitions on each variable (see Section 4.5). The main difference is that in grid sampling, a small random perturbation is added to each sample value so that the grid points are not on a perfect grid. This is done to help capture certain features in the output such as periodic

functions. A purely structured grid, with the samples exactly on the grid points, has the disadvantage of not being able to capture important features such as periodic functions with relatively high frequency (due to aliasing). Adding a random perturbation to the grid samples helps remedy this problem.

Another disadvantage with grid sampling is that the number of sample points required depends exponentially on the input dimensions. In grid sampling, the number of samples is the number of symbols (grid partitions) raised to the number of variables. For example, if there are 2 variables, each with 5 partitions, the number of samples would be 5^2. In this case, doubling the number of variables squares the sample size. The grid design is specified in DAKOTA with the method command `dace grid`.

5.3.5 Monte Carlo Design

Monte Carlo designs simply involve pure Monte-Carlo random sampling from uniform distributions between the lower and upper bounds on each of the input variables. Monte Carlo designs, specified by `dace random`, are a way to generate a set of random samples over an input domain.

5.3.6 LHS Design

DDACE offers the capability to generate Latin Hypercube designs. For more information on Latin Hypercube sampling, see Section 6.2. Note that the version of LHS in DDACE generates uniform samples (uniform between the variable bounds). The version of LHS offered with nondeterministic sampling can generate LHS samples according to a number of distribution types, including normal, lognormal, weibull, beta, etc. To specify the DDACE version of LHS, use the method command `dace lhs`.

5.3.7 OA-LHS Design

DDACE offers a hybrid design which is combination of an orthogonal array and a Latin Hypercube sample. This design is specified with the method command `dace oa_lhs`. This design has the advantages of both orthogonality of the inputs as well as stratification of the samples (see [102]).

5.4 FSUDace Background

The FSUDace package includes quasi-Monte Carlo sampling methods (Halton and Hammersley sequences) and Centroidal Voronoi Tesselation sampling. All three methods generate sets of uniform random variables on the interval $[0, 1]$. The quasi-Monte Carlo and CVT methods are designed with the goal of low discrepancy. Discrepancy refers to the nonuniformity of the sample points within the unit hypercube. Low discrepancy sequences tend to cover the unit hypercube reasonably uniformly. Quasi-Monte Carlo methods produce low discrepancy sequences, especially if one is interested in the uniformity of projections of the point sets onto lower dimensional faces of the hypercube (usually 1-D: how well do the marginal distributions approximate a uniform?) CVT does very well volumetrically: it spaces the points fairly equally throughout the space, so that the points cover the region and are isotropically distributed with no directional bias in the point placement. There are various measures of volumetric uniformity which take into account the distances between pairs of points, regularity measures, etc. Note that CVT does not produce low-discrepancy sequences in lower dimensions, however: the lower-dimension (such as 1-D) projections of CVT can have high discrepancy.

The quasi-Monte Carlo sequences of Halton and Hammersley are deterministic sequences determined by a set of prime bases. A Halton design is specified in DAKOTA with the method command `fsu_quasi_mc halton`, and the Hammersley design is specified with the command `fsu_quasi_mc hammersley`. For more details about the input specification, see the Reference Manual. CVT points tend to arrange themselves in a pattern of cells that are roughly the same shape. To produce CVT points, an almost arbitrary set of initial points is chosen, and then an internal set of iterations is carried out. These iterations repeatedly replace the current set of sample points by an estimate of the centroids of the corresponding Voronoi subregions [25]. A CVT design is specified in DAKOTA with the method command `fsu_cvt`.

The methods in FSUDace are useful for design of experiments because they provide good coverage of the input space, thus allowing global sensitivity analysis.

5.5 PSUADE MOAT Background

This section describes the Morris One-At-A-Time (MOAT) screening method as implemented in LLNL's PSUADE (Problem Solving Environment for Uncertainty Analysis and Design Exploration) [126]. DAKOTA includes a prototype interface to this DACE method, which is useful for global sensitivity analysis, including interaction studies.

The Morris One-At-A-Time method, originally proposed by M. D. Morris [93], is a screening method, designed to explore a computational model to distinguish between input variables that have negligible, linear and additive, or nonlinear or interaction effects on the output. The computer experiments performed consist of individually randomized designs which vary one input factor at a time to create a sample of its elementary effects.

With MOAT, each dimension of a $k-$dimensional input space is uniformly partitioned into p levels, creating a grid of p^k points $\mathbf{x} \in \mathbb{R}^k$ at which evaluations of the model $y(\mathbf{x})$ might take place. An elementary effect corresponding to input i is computed by a forward difference

$$d_i(\mathbf{x}) = \frac{y(\mathbf{x} + \Delta \mathbf{e}_i) - y(\mathbf{x})}{\Delta}, \tag{5.1}$$

where e_i is the ith coordinate vector, and the step Δ is typically taken to be large (this is not intended to be a local derivative approximation). In the present implementation of MOAT, for an input variable scaled to $[0, 1]$, $\Delta = \frac{p}{2(p-1)}$, so the step used to find elementary effects is slightly larger than half the input range.

The distribution of elementary effects d_i over the input space characterizes the effect of input i on the output of interest. After generating r samples from this distribution, their mean,

$$\mu_i = \frac{1}{r} \sum_{j=1}^{r} d_i^{(j)}, \tag{5.2}$$

modified mean

$$\mu_i^* = \frac{1}{r} \sum_{j=1}^{r} |d_i^{(j)}|, \tag{5.3}$$

(using absolute value) and standard deviation

$$\sigma_i = \sqrt{\frac{1}{r} \sum_{j=1}^{r} \left(d_i^{(j)} - \mu_i \right)^2} \tag{5.4}$$

are computed for each input i. The mean and modified mean give an indication of the overall effect of an input on the output. Standard deviation indicates nonlinear effects or interactions, since it is an indicator of elementary effects varying throughout the input space.

The MOAT method is selected with method keyword `psuade_moat` as shown in the sample DAKOTA input deck in Figure 5.1. The number of samples (`samples`) must be a positive integer multiple of (number of continuous design variables k + 1) and will be automatically adjusted if misspecified. The number of partitions (`partitions`) applies to each variable being studied and must be odd (the number of MOAT levels p per variable is partitions + 1, similar to DAKOTA multidimensional parameter studies). This will also be adjusted at runtime as necessary. Finite user-specified lower and upper bounds are required and will be scaled as needed by the method. For more information on use of MOAT sampling, see the Morris example in Section 22.12, or Saltelli, et al. [115].

```
strategy,
        single_method

method,
        psuade_moat
          samples = 84
          partitions = 3
          seed = 5

variables,
        continuous_design = 20
          lower_bounds =    0.0 0.0 0.0 0.0 0.0
                                0.0 0.0 0.0 0.0 0.0
                                0.0 0.0 0.0 0.0 0.0
                                0.0 0.0 0.0 0.0 0.0
          upper_bounds =    1.0 1.0 1.0 1.0 1.0
                                1.0 1.0 1.0 1.0 1.0
                                1.0 1.0 1.0 1.0 1.0
                                1.0 1.0 1.0 1.0 1.0

interface,
        system asynchronous evaluation_concurrency = 5
          analysis_driver = 'morris'

responses,
        num_objective_functions = 1
        no_gradients
        no_hessians
```

Figure 5.1: DAKOTA input file `dakota_psuade.in` to invoke the PSUADE MOAT method on the Morris test problem described in Section 22.12.

5.6 Sensitivity Analysis

Like parameter studies (see Chapter 4), the DACE techniques are useful for characterizing the behavior of the response functions of interest through the parameter ranges of interest. In addition to direct interrogation and

visualization of the sampling results, a number of techniques have been developed for assessing the parameters which are most influential in the observed variability in the response functions. One example of this is the well-known technique of scatter plots, in which the set of samples is projected down and plotted against one parameter dimension, for each parameter in turn. Scatter plots with a uniformly distributed cloud of points indicate parameters with little influence on the results, whereas scatter plots with a defined shape to the cloud indicate parameters which are more significant. Related techniques include analysis of variance (ANOVA) [94] and main effects analysis, in which the parameters which have the greatest influence on the results are identified from sampling results. Scatter plots and ANOVA may be accessed through import of DAKOTA tabular results (see Section 16.3) into external statistical analysis programs such as S-plus, Minitab, etc.

Running any of the design of experiments or sampling methods allows the user to save the results in a tabular data file, which then can be read into a spreadsheet or statistical package for further analysis. In addition, we have provided some functions to help determine the most important variables.

We take the definition of uncertainty analysis from [115]: "The study of how uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model input."

As a default, DAKOTA provides correlation analyses when running LHS. Correlation tables are printed with the simple, partial, and rank correlations between inputs and outputs. These can be useful to get a quick sense of how correlated the inputs are to each other, and how correlated various outputs are to inputs. The correlation analyses are explained further in Chapter 6.2.

We also have the capability to calculate sensitivity indices through Variance-based Decomposition (VBD). Variance-based decomposition is a way of using sets of samples to understand how the variance of the output behaves, with respect to each input variable. A larger value of the sensitivity index, S_i (`Si` in the DAKOTA output), means that the uncertainty in the input variable i has a larger effect on the variance of the output. More details on the calculations and interpretation of the sensitivity indices can be found in [115]. VBD can be specified for any of the sampling methods using the command `variance_based_decomposition`. Note that VBD is extremely computationally intensive since replicated sets of sample values are evaluated. If the user specified a number of samples, N, and a number of nondeterministic variables, M, variance-based decomposition requires the evaluation of $N(M + 2)$ samples. To obtain sensitivity indices that are reasonably accurate, we recommend that N, the number of samples, be at least one hundred and preferably several hundred or thousands. Because of the computational cost, Variance-based decomposition is turned off as a default.

Finally, we have the capability to calculate a set of quality metrics for a particular input sample. These quality metrics measure various aspects relating to the volumetric spacing of the samples: are the points equally spaced, do they cover the region, are they isotropically distributed, do they have directional bias, etc.? The quality metrics are explained in more detail in the Reference Manual.

Chapter 6

Uncertainty Quantification Capabilities

6.1 Overview

Uncertainty quantification (UQ) is the process of determining the effect of input uncertainties on response metrics of interest. These input uncertainties may be characterized as either aleatory uncertainties, which are irreducible variabilities inherent in nature, or epistemic uncertainties, which are reducible uncertainties resulting from a lack of knowledge. Since sufficient data is generally available for aleatory uncertainties, probabilistic methods are commonly used for computing response distribution statistics based on input probability distribution specifications. Conversely, for epistemic uncertainties, data is generally sparse, making the use of probability theory questionable and leading to nonprobabilistic methods based on interval specifications.

DAKOTA contains capabilities for performing nondeterministic analysis. The methods for uncertainty quantification in DAKOTA have been developed by a group of researchers at Sandia Labs, in conjunction with collaborators in academia [53, 54, 32]. These methods perform a forward uncertainty propagation in which probability or interval information for input parameters is mapped to probability or interval information for output response functions. The m functions in the DAKOTA response data set are interpreted as m general response functions by the DAKOTA methods (with no specific interpretation of the functions as for optimization and least squares).

Within the variables specification, uncertain variable descriptions are employed to define the parameter probability distributions (see Section 12.3). The distribution types include: normal (Gaussian), lognormal, uniform, loguniform, triangular, exponential, beta, gamma, gumbel, frechet, weibull, histogram, and interval. All uncertain variables are treated as continuous variables in DAKOTA. When gradient and/or Hessian information is used in an uncertainty assessment, derivative components are normally computed with respect to the active continuous variables, or in this case, the *uncertain variables*.

6.2 Sampling Methods

Sampling techniques are selected using the `nond_sampling` method selection. This method generates sets of samples according to the probability distributions of the uncertain variables and maps them into corresponding sets of response functions, where the number of samples is specified by the `samples` integer specification. Means, standard deviations, coefficients of variation (COVs), and 95% confidence intervals are computed for the response functions. Probabilities and reliabilities may be computed for `response_levels` specifications, and response levels may be computed for either `probability_levels` or `reliability_levels` specifications (refer

to the Method Commands chapter in the DAKOTA Reference Manual [30] for additional information).

Currently, traditional Monte Carlo (MC) and Latin hypercube sampling (LHS) are supported by DAKOTA and are chosen by specifying `sample_type` as `random` or `lhs`. In Monte Carlo sampling, the samples are selected randomly according to the user-specified probability distributions. Latin hypercube sampling is a stratified sampling technique for which the range of each uncertain variable is divided into N_s segments of equal probability, where N_s is the number of samples requested. The relative lengths of the segments are determined by the nature of the specified probability distribution (e.g., uniform has segments of equal width, normal has small segments near the mean and larger segments in the tails). For each of the uncertain variables, a sample is selected randomly from each of these equal probability segments. These N_s values for each of the individual parameters are then combined in a shuffling operation to create a set of N_s parameter vectors with a specified correlation structure. A feature of the resulting sample set is that *every row and column in the hypercube of partitions has exactly one sample*. Since the total number of samples is exactly equal to the number of partitions used for each uncertain variable, an arbitrary number of desired samples is easily accommodated (as compared to less flexible approaches in which the total number of samples is a product or exponential function of the number of intervals for each variable, i.e., many classical design of experiments methods).

Advantages of sampling-based methods include their relatively simple implementation and their independence from the scientific disciplines involved in the analysis. The main drawback of these techniques is the large number of function evaluations needed to generate converged statistics, which can render such an analysis computationally very expensive, if not intractable, for real-world engineering applications. LHS techniques, in general, require fewer samples than traditional Monte Carlo for the same accuracy in statistics, but they still can be prohibitively expensive. For further information on the method and its relationship to other sampling techniques, one is referred to the works by McKay, et al. [91], Iman and Shortencarier [80], and Helton and Davis [73]. Note that under certain separability conditions associated with the function to be sampled, Latin hypercube sampling provides a more accurate estimate of the mean value than does random sampling. That is, given an equal number of samples, the LHS estimate of the mean will have less variance than the mean value obtained through random sampling.

Figure 6.1 demonstrates Latin hypercube sampling on a two-variable parameter space. Here, the range of both parameters, x_1 and x_2, is $[0, 1]$. Also, for this example both x_1 and x_2 have uniform statistical distributions. For Latin hypercube sampling, the range of each parameter is divided into p "bins" of equal probability. For parameters with uniform distributions, this corresponds to partitions of equal size. For n design parameters, this partitioning yields a total of p^n bins in the parameter space. Next, p samples are randomly selected in the parameter space, with the following restrictions: (a) each sample is randomly placed inside a bin, and (b) for all one-dimensional projections of the p samples and bins, there will be one and only one sample in each bin. In a two-dimensional example such as that shown in Figure 6.1, these LHS rules guarantee that only one bin can be selected in each row and column. For $p = 4$, there are four partitions in both x_1 and x_2. This gives a total of 16 bins, of which four will be chosen according to the criteria described above. Note that there is more than one possible arrangement of bins that meet the LHS criteria. The dots in Figure 6.1 represent the four sample sites in this example, where each sample is randomly located in its bin. There is no restriction on the number of bins in the range of each parameter, however, all parameters must have the same number of bins.

The actual algorithm for generating Latin hypercube samples is more complex than indicated by the description given above. For example, the Latin hypercube sampling method implemented in the LHS code [125] takes into account a user-specified correlation structure when selecting the sample sites. For more details on the implementation of the LHS algorithm, see Reference [125].

6.2.1 Uncertainty Quantification Example using Sampling Methods

The two-variable Textbook example problem (see Equation 2.3) will be used to demonstrate the application of sampling methods for uncertainty quantification where it is assumed that x_1 and x_2 are uniform uncertain variables

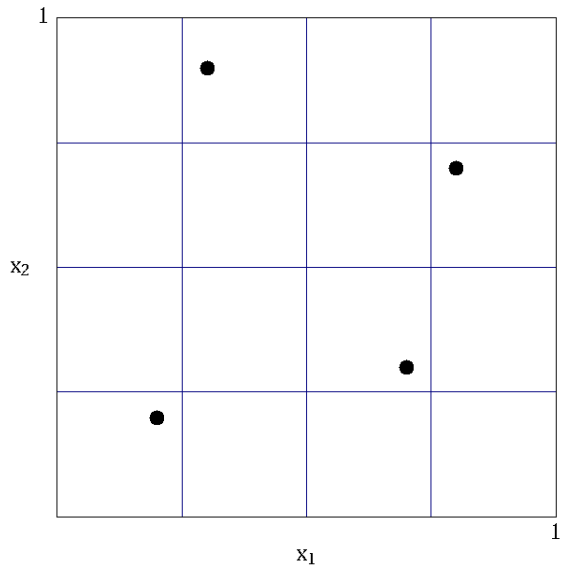


Figure 6.1: An example of Latin hypercube sampling with four bins in design parameters x_1 and x_2. The dots are the sample sites.

on the interval $[0, 1]$. The DAKOTA input file for this problem is shown in Figure 6.2. The number of samples to perform is controlled with the `samples` specification, the type of sampling algorithm to use is controlled with the `sample_type` specification, the levels used for computing statistics on the response functions is specified with the `response_levels` input, and the `seed` specification controls the sequence of the pseudo-random numbers generated by the sampling algorithms. The input samples generated are shown in Figure 6.3 for the case where `samples` = 5 and `samples` = 10 for both `random` ($\circ$) and `lhs` ($+$) sample types.

Latin hypercube sampling ensures full coverage of the range of the input variables, which is often a problem with Monte Carlo sampling when the number of samples is small. In the case of `samples = 5`, poor stratification is evident in x_1 as four out of the five Monte Carlo samples are clustered in the range $0.35 < x_1 < 0.55$, and the regions $x_1 < 0.3$ and $0.6 < x_1 < 0.9$ are completely missed. For the case where `samples = 10`, some clustering in the Monte Carlo samples is again evident with `4` samples in the range $0.5 < x_1 < 0.55$. In both cases, the stratification with LHS is superior. The response function statistics returned by DAKOTA are shown in Figure 6.4. The first two blocks of output specify the response sample means and sample standard deviations and confidence intervals for these statistics, as well as coefficients of variation. The last section of the output defines CDF pairs (`distribution cumulative` was specified) for the response functions by presenting the probability levels corresponding to the specified response levels (`response_levels` were set and the default `compute probabilities` was used). Alternatively, DAKOTA could have provided CCDF pairings, reliability levels corresponding to prescribed response levels, or response levels corresponding to prescribed probability or reliability levels.

In addition to obtaining statistical summary information of the type shown in Figure 6.4, the results of LHS sampling also include correlations. Four types of correlations are returned in the output: simple and partial "raw" correlations, and simple and partial "rank" correlations. The raw correlations refer to correlations performed on the actual input and output data. Rank correlations refer to correlations performed on the ranks of the data. Ranks are obtained by replacing the actual data by the ranked values, which are obtained by ordering the data in ascending order. For example, the smallest value in a set of input samples would be given a rank 1, the next smallest value a rank 2, etc. Rank correlations are useful when some of the inputs and outputs differ greatly in

```
strategy
        single_method
          tabular_graphics_data
          method_pointer = 'UQ'

method,
        id_method = 'UQ'
        nond_sampling,
          samples = 10 seed = 98765
          response_levels = 0.1 0.2 0.6
                            0.1 0.2 0.6
                            0.1 0.2 0.6
          sample_type lhs
          distribution cumulative




variables,
        uniform_uncertain = 2
          lower_bounds =  0.   0.
          upper_bounds =  1.   1.
          descriptors  = 'x1' 'x2'

interface,
        id_interface = 'I1'
        system asynch evaluation_concurrency = 5
          analysis_driver = 'text_book'

responses,
        num_response_functions = 3
        no_gradients
        no_hessians
```

Figure 6.2: DAKOTA input file for UQ example using LHS sampling.

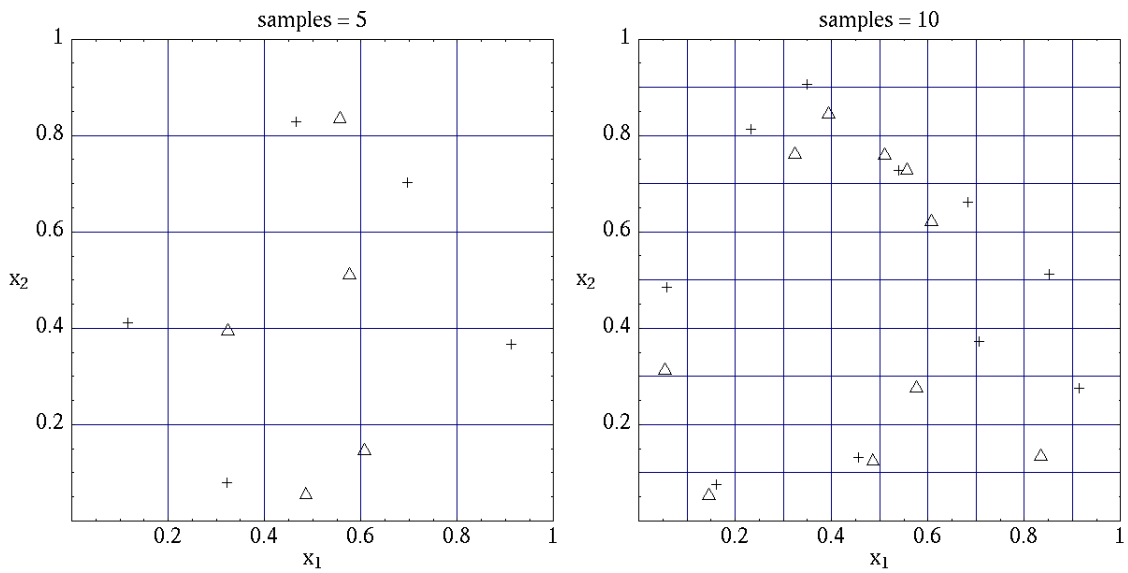


Figure 6.3: Distribution of input sample points for random (△) and lhs (+) sampling for `samples=5` and `10`.

magnitude: then it is easier to compare if the smallest ranked input sample is correlated with the smallest ranked output, for example.

Correlations are always calculated between two sets of sample data. One can calculate correlation coefficients between two input variables, between an input and an output variable (probably the most useful), or between two output variables. The simple correlation coefficients presented in the output tables are Pearson's correlation coefficient, which is defined for two variables x and y as: $\texttt{Corr}(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$. Partial correlation coefficients are similar to simple correlations, but a partial correlation coefficient between two variables measures their correlation while adjusting for the effects of the other variables. For example, say one has a problem with two inputs and one output; and the two inputs are highly correlated. Then the correlation of the second input and the output may be very low after accounting for the effect of the first input. The rank correlations in DAKOTA are obtained using Spearman's rank correlation. Spearman's rank is the same as the Pearson correlation coefficient except that it is calculated on the rank data.

Figure 6.5 shows an example of the correlation output provided by DAKOTA for the input file in Figure 6.2. Note that these correlations are presently only available when one specifies `lhs` as the sampling method under `nond_sampling`. Also note that the simple and partial correlations should be similar in most cases (in terms of values of correlation coefficients). This is because we use a default "restricted pairing" method in the LHS routine which forces near-zero correlation amongst uncorrelated inputs.

Finally, note that the LHS package can be used in design of experiments mode by including the `all_variables` flag in the method specification section of the DAKOTA input file. Then, instead of iterating on only the uncertain variables, the LHS package will sample on all of the continuous variables, where continuous design and continuous state variables are treated as having uniform probability distributions within their upper and lower bounds and any uncertain variables are sampled within their specified probability distributions.

```
Statistics based on 10 samples:

Moments for each response function:
response_fn_1:  Mean = 3.83840e-01  Std. Dev. = 4.02815e-01
                Coeff. of Variation = 1.04944e+00
response_fn_2:  Mean = 7.47987e-02  Std. Dev. = 3.46861e-01
                Coeff. of Variation = 4.63726e+00
response_fn_3:  Mean = 7.09462e-02  Std. Dev. = 3.41532e-01
                Coeff. of Variation = 4.81397e+00

95% confidence intervals for each response function:
response_fn_1:  Mean = (  9.56831e-02, 6.71997e-01 ),
             Std Dev = (  2.77071e-01, 7.35384e-01 )
response_fn_2:  Mean = ( -1.73331e-01, 3.22928e-01 ),
             Std Dev = (  2.38583e-01, 6.33233e-01 )
response_fn_3:  Mean = ( -1.73371e-01, 3.15264e-01 ),
             Std Dev = (  2.34918e-01, 6.23505e-01 )

Probabilities for each response function:
Cumulative Distribution Function (CDF) for response_fn_1:
     Response Level  Probability Level  Reliability Index
     --------------  -----------------  -----------------
   1.0000000000e-01   3.0000000000e-01
   2.0000000000e-01   5.0000000000e-01
   6.0000000000e-01   7.0000000000e-01
Cumulative Distribution Function (CDF) for response_fn_2:
     Response Level  Probability Level  Reliability Index
     --------------  -----------------  -----------------
   1.0000000000e-01   5.0000000000e-01
   2.0000000000e-01   7.0000000000e-01
   6.0000000000e-01   9.0000000000e-01
Cumulative Distribution Function (CDF) for response_fn_3:
     Response Level  Probability Level  Reliability Index
     --------------  -----------------  -----------------
   1.0000000000e-01   6.0000000000e-01
   2.0000000000e-01   6.0000000000e-01
   6.0000000000e-01   9.0000000000e-01
```

Figure 6.4: DAKOTA response function statistics from UQ sampling example.

```
Simple Correlation Matrix between input and output:
                        x1           x2 response_fn_1 response_fn_2 response_fn_3
          x1  1.00000e+00
          x2 -7.22482e-02  1.00000e+00
response_fn_1 -7.04965e-01 -6.27351e-01  1.00000e+00
response_fn_2  8.61628e-01 -5.31298e-01 -2.60486e-01  1.00000e+00
response_fn_3 -5.83075e-01  8.33989e-01 -1.23374e-01 -8.92771e-01  1.00000e+00

Partial Correlation Matrix between input and output:
             response_fn_1 response_fn_2 response_fn_3
          x1 -9.65994e-01  9.74285e-01 -9.49997e-01
          x2 -9.58854e-01 -9.26578e-01  9.77252e-01

Simple Rank Correlation Matrix between input and output:
                        x1           x2 response_fn_1 response_fn_2 response_fn_3
          x1  1.00000e+00
          x2 -6.66667e-02  1.00000e+00
response_fn_1 -6.60606e-01 -5.27273e-01  1.00000e+00
response_fn_2  8.18182e-01 -6.00000e-01 -2.36364e-01  1.00000e+00
response_fn_3 -6.24242e-01  7.93939e-01 -5.45455e-02 -9.27273e-01  1.00000e+00

Partial Rank Correlation Matrix between input and output:
             response_fn_1 response_fn_2 response_fn_3
          x1 -8.20657e-01  9.74896e-01 -9.41760e-01
          x2 -7.62704e-01 -9.50799e-01  9.65145e-01
```

Figure 6.5: Correlation results using LHS Sampling.

6.2.2 Incremental Sampling

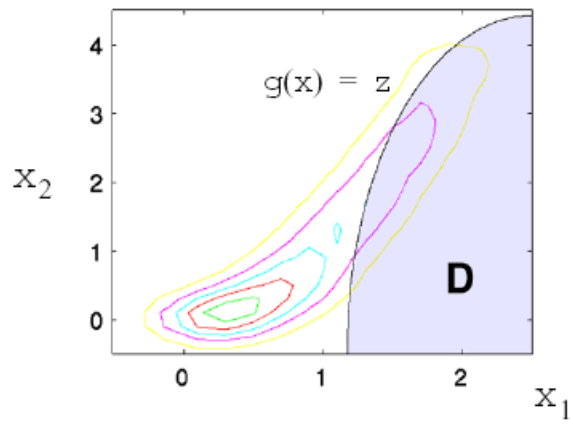
In many situations, one may run an initial sample set and then need to perform further sampling to get better estimates of the mean, variance, and percentiles, and to obtain more comprehensive sample coverage. We call this capability incremental sampling. Currently, the LHS incremental sampling capability we have in DAKOTA requires that the incremental samples are double the sample size of the previous sample. That is, if one starts with a very small sample size of 10, then one can use the incremental sampling capability to generate sample sizes of 20, 40, 80, etc. Also, a DAKOTA restart file (dakota.rst) must be available from the original sample. There are two cases, random incremental and Latin Hypercube incremental sampling. We assume that LHS incremental will be most frequently used. One major advantage of LHS incremental sampling is that it maintains the stratification and correlation structure of the original LHS sample. That is, if one generated 2 independent LHS samples and just merged them, the calculation of the accuracy of statistical measures such as the mean and the variance would be slightly incorrect. However, in the incremental case, the full sample (double the original size) is a Latin Hypercube sample itself and statistical measures and their accuracy can be properly calculated. The incremental sampling capability is most useful when one is starting off with very small samples. Once the sample size is more than a few hundred, the benefit of incremental sampling diminishes.

1. Incremental Random Sampling. With incremental random sampling, the original sample set with N samples must be generated using `sample_type` as `random`. Then, the user can create a new DAKOTA input file that is very similar to the original one except the `sample_type` should be defined to be `incremental_random`. Random incremental sampling does not require a doubling of samples each time. Thus, the user can specify the number of samples (`samples`) to be a desired number (it can range from an additional one sample to a large integer), and the `previous_samples` should be specified as N. For example, if the first sample has 50 samples, and 10 more samples are desired, in the second DAKOTA run, the number of samples should be set to 60 and the number of previous samples set to 50. In this situation, only 10 new samples will be generated and the final statistics will be reported on the full sample of 60. The command line syntax for running the second sample is `dakota -i input2.in -r dakota.rst` where input2.in is the input file with the incremental sampling specification and dakota.rst is the restart file. Note that if the restart file has a different name, that is fine; the correct restart file name should be used.

2. Incremental Latin Hypercube Sampling. With incremental LHS sampling, the original sample set with N samples must be generated using `sample_type` as `lhs`. Then, the user can create a new DAKOTA input file that is very similar to the original one except the `sample_type` should be defined to be `incremental_lhs`, the number of samples (`samples`) should be 2N (twice the number of original samples), and `previous_samples` should be specified as N. For example, if the first sample has 50 samples, in the second DAKOTA run, the number of samples should be set to 100 and the number of previous samples set to 50. In this situation, only 50 new samples will be generated and the final statistics will be reported on the full sample of 100. The command line syntax for running the second sample is `dakota -i input2.in -r dakota.rst`, where input2.in is the input file with the incremental sampling specification and dakota.rst is the restart file. Note that if the restart file has a different name, that is fine; the correct restart file name should be used.

6.3 Reliability Methods

Reliability methods provide an alternative approach to uncertainty quantification which can be less computationally demanding than sampling techniques. Reliability methods for uncertainty quantification are based on probabilistic approaches that compute approximate response function distribution statistics based on specified uncertain variable distributions. These response statistics include response mean, response standard deviation, and cumulative or complementary cumulative distribution functions (CDF/CCDF). These methods are often more ef-



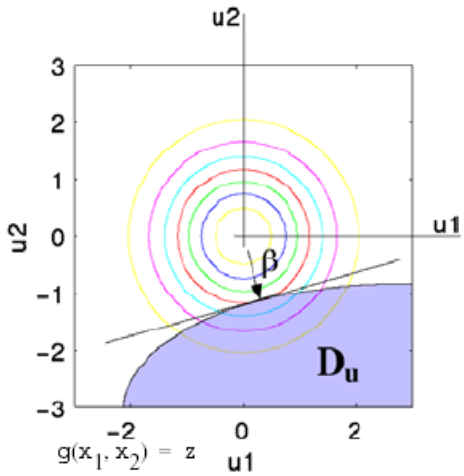
$$P[g(\mathbf{X}) < z] = \iint_{\mathbf{x} \in D} p(x_1, x_2) d\mathbf{x} = P[(\mathbf{x} \in \mathbf{D})]$$

Figure 6.6: Graphical depiction of calculation of cumulative distribution function in the original uncertain variable space.

ficient at computing statistics in the tails of the response distributions (events with low probability) than sampling based approaches since the number of samples required to resolve a low probability can be prohibitive.

The methods all answer the fundamental question: "Given a set of uncertain input variables, $\mathbf{X}$, and a scalar response function, g, what is the probability that the response function is below or above a certain level, $\bar{z}$?" The former can be written as $P[g(\mathbf{X}) \leq \bar{z}] = F_g(\bar{z})$ where $F_g(\bar{z})$ is the cumulative distribution function (CDF) of the uncertain response $g(\mathbf{X})$ over a set of response levels. The latter can be written as $P[g(\mathbf{X}) > \bar{z}]$ and defines the complementary cumulative distribution function (CCDF).

This probability calculation involves a multi-dimensional integral over an irregularly shaped domain of interest, $\mathbf{D}$, where $g(\mathbf{X}) < z$ as displayed in Figure 6.6 for the case of two variables. The reliability methods all involve the transformation of the user-specified uncertain variables, $\mathbf{X}$, with probability density function, $p(x_1, x_2)$, which can be non-normal and correlated, to a space of independent Gaussian random variables, $\mathbf{u}$, possessing a mean value of zero and unit variance (i.e., standard normal variables). The region of interest, $\mathbf{D}$, is also mapped to the transformed space to yield, $\mathbf{D_u}$, where $g(\mathbf{U}) < z$ as shown in Figure 6.7. The Nataf transformation [23], which is identical to the Rosenblatt transformation [113] in the case of independent random variables, is used in DAKOTA to accomplish this mapping. This transformation is performed to make the probability calculation more tractable. In the transformed space, probability contours are circular in nature as shown in Figure 6.7 unlike in the original uncertain variable space, Figure 6.6. Also, the multi-dimensional integrals can be approximated by simple functions of a single parameter, β, called the reliability index. β is the minimum Euclidean distance from the origin in the transformed space to the response surface. This point is also known as the most probable point (MPP) of failure. Note, however, the methodology is equally applicable for generic functions, not simply those corresponding to failure criteria; this nomenclature is due to the origin of these methods within the disciplines of structural safety and reliability. Note that there are local and global reliability methods. The majority of the methods available are local, meaning that a local optimization formulation is used to locate one MPP. In contrast, global methods can find multiple MPPs if they exist.



$$P(\mathbf{X} \in D) = P(U \in D_U) \approx f(\beta)$$

Figure 6.7: Graphical depiction of integration for the calculation of cumulative distribution function in the transformed uncertain variable space.

6.3.1 Mean Value

The Mean Value method (MV, also known as MVFOSM in [70]) is the simplest, least-expensive reliability method because it estimates the response means, response standard deviations, and all CDF/CCDF response-probability-reliability levels from a single evaluation of response functions and their gradients at the uncertain variable means. This approximation can have acceptable accuracy when the response functions are nearly linear and their distributions are approximately Gaussian, but can have poor accuracy in other situations. The expressions for approximate response mean μ_g, approximate response standard deviation σ_g, response target to approximate probability/reliability level mapping ($\bar{z} \rightarrow p, \beta$), and probability/reliability target to approximate response level mapping ($\bar{p}, \bar{\beta} \rightarrow z$) are

$$\mu_g = g(\mu_{\mathbf{x}}) \tag{6.1}$$

$$\sigma_g = \sum_i \sum_j Cov(i,j) \frac{dg}{dx_i}(\mu_{\mathbf{x}}) \frac{dg}{dx_j}(\mu_{\mathbf{x}}) \tag{6.2}$$

$$\beta_{cdf} = \frac{\mu_g - \bar{z}}{\sigma_g} \tag{6.3}$$

$$\beta_{ccdf} = \frac{\bar{z} - \mu_g}{\sigma_g} \tag{6.4}$$

$$z = \mu_g - \sigma_g \bar{\beta}_{cdf} \tag{6.5}$$

$$z = \mu_g + \sigma_g \bar{\beta}_{ccdf} \tag{6.6}$$

respectively, where $\mathbf{x}$ are the uncertain values in the space of the original uncertain variables ("x-space"), $g(\mathbf{x})$ is the limit state function (the response function for which probability-response level pairs are needed), and β_{cdf} and β_{ccdf} are the CDF and CCDF reliability indices, respectively.

With the introduction of second-order limit state information, MVSOSM calculates a second-order mean as

$$\mu_g = g(\mu_{\mathbf{x}}) + \frac{1}{2}\sum_i \sum_j Cov(i,j)\frac{d^2 g}{dx_i dx_j}(\mu_{\mathbf{x}}) \tag{6.7}$$

This is commonly combined with a first-order variance (Equation 6.2), since second-order variance involves higher order distribution moments (skewness, kurtosis) [70] which are often unavailable.

The first-order CDF probability $p(g \leq z)$, first-order CCDF probability $p(g > z)$, β_{cdf}, and β_{ccdf} are related to one another through

$$\begin{aligned}
p(g \leq z) &= \Phi(-\beta_{cdf}) && (6.8)\\
p(g > z) &= \Phi(-\beta_{ccdf}) && (6.9)\\
\beta_{cdf} &= -\Phi^{-1}(p(g \leq z)) && (6.10)\\
\beta_{ccdf} &= -\Phi^{-1}(p(g > z)) && (6.11)\\
\beta_{cdf} &= -\beta_{ccdf} && (6.12)\\
p(g \leq z) &= 1 - p(g > z) && (6.13)
\end{aligned}$$

where $\Phi()$ is the standard normal cumulative distribution function. A common convention in the literature is to define g in such a way that the CDF probability for a response level z of zero (i.e., $p(g \leq 0)$) is the response metric of interest. DAKOTA is not restricted to this convention and is designed to support CDF or CCDF mappings for general response, probability, and reliability level sequences.

With the Mean Value method, it is possible to obtain importance factors indicating the relative importance of input variables. The importance factors can be viewed as an extension of linear sensitivity analysis combining deterministic gradient information with input uncertainty information, *i.e.* input variable standard deviations. The accuracy of the importance factors is contingent of the validity of the linear approximation used to approximate the true response functions. The importance factors are determined as:

$$ImpFactor_i = (\frac{\sigma_{x_i}}{\sigma_g}\frac{dg}{dx_i}(\mu_{\mathbf{x}}))^2 \tag{6.14}$$

6.3.2 MPP Search Methods

All other local reliability methods solve an equality-constrained nonlinear optimization problem to compute a most probable point (MPP) and then integrate about this point to compute probabilities. The MPP search is performed in uncorrelated standard normal space ("u-space") since it simplifies the probability integration: the distance of the MPP from the origin has the meaning of the number of input standard deviations separating the mean response from a particular response threshold. The transformation from correlated non-normal distributions (x-space) to uncorrelated standard normal distributions (u-space) is denoted as $\mathbf{u} = T(\mathbf{x})$ with the reverse transformation denoted as $\mathbf{x} = T^{-1}(\mathbf{u})$. These transformations are nonlinear in general, and possible approaches include the Rosenblatt [113], Nataf [23], and Box-Cox [11] transformations. The nonlinear transformations may also be linearized, and common approaches for this include the Rackwitz-Fiessler [108] two-parameter equivalent normal and the Chen-Lind [16] and Wu-Wirsching [141] three-parameter equivalent normals. DAKOTA employs the Nataf nonlinear transformation which occurs in the following two steps. To transform between the original correlated x-space variables and correlated standard normals ("z-space"), the CDF matching condition is used:

$$\Phi(z_i) = F(x_i) \tag{6.15}$$

where $F()$ is the cumulative distribution function of the original probability distribution. Then, to transform between correlated z-space variables and uncorrelated u-space variables, the Cholesky factor $\mathbf{L}$ of a modified correlation matrix is used:

$$\mathbf{z} = \mathbf{L}\mathbf{u} \tag{6.16}$$

where the original correlation matrix for non-normals in x-space has been modified to represent the corresponding correlation in z-space [23].

The forward reliability analysis algorithm of computing CDF/CCDF probability/reliability levels for specified response levels is called the reliability index approach (RIA), and the inverse reliability analysis algorithm of computing response levels for specified CDF/CCDF probability/reliability levels is called the performance measure approach (PMA) [128]. The differences between the RIA and PMA formulations appear in the objective function and equality constraint formulations used in the MPP searches. For RIA, the MPP search for achieving the specified response level $\bar{z}$ is formulated as computing the minimum distance in u-space from the origin to the $\bar{z}$ contour of the limit state response function:

$$\begin{array}{ll} \text{minimize} & \mathbf{u}^T\mathbf{u} \\ \text{subject to} & G(\mathbf{u}) = \bar{z} \end{array} \tag{6.17}$$

and for PMA, the MPP search for achieving the specified reliability/probability level $\bar{\beta}, \bar{p}$ is formulated as computing the minimum/maximum response function value corresponding to a prescribed distance from the origin in u-space:

$$\begin{array}{ll} \text{minimize} & \pm G(\mathbf{u}) \\ \text{subject to} & \mathbf{u}^T\mathbf{u} = \bar{\beta}^2 \end{array} \tag{6.18}$$

where $\mathbf{u}$ is a vector centered at the origin in u-space and $g(\mathbf{x}) \equiv G(\mathbf{u})$ by definition. In the RIA case, the optimal MPP solution $\mathbf{u}^*$ defines the reliability index from $\beta = \pm\|\mathbf{u}^*\|_2$, which in turn defines the CDF/CCDF probabilities (using Equations 6.8-6.9 in the case of first-order integration). The sign of β is defined by

$$G(\mathbf{u}^*) > G(\mathbf{0}) : \beta_{cdf} < 0, \beta_{ccdf} > 0 \tag{6.19}$$

$$G(\mathbf{u}^*) < G(\mathbf{0}) : \beta_{cdf} > 0, \beta_{ccdf} < 0 \tag{6.20}$$

where $G(\mathbf{0})$ is the median limit state response computed at the origin in u-space (where β_{cdf} = β_{ccdf} = 0 and first-order $p(g \leq z)$ = $p(g > z)$ = 0.5). In the PMA case, the sign applied to $G(\mathbf{u})$ (equivalent to minimizing or maximizing $G(\mathbf{u})$) is similarly defined by $\bar{\beta}$

$$\bar{\beta}_{cdf} < 0, \bar{\beta}_{ccdf} > 0 : \text{maximize}\, G(\mathbf{u}) \tag{6.21}$$

$$\bar{\beta}_{cdf} > 0, \bar{\beta}_{ccdf} < 0 : \text{minimize}\, G(\mathbf{u}) \tag{6.22}$$

and the limit state at the MPP ($G(\mathbf{u}^*)$) defines the desired response level result.

Limit state approximations

There are a variety of algorithmic variations that are available for use within RIA/PMA reliability analyses. First, one may select among several different limit state approximations that can be used to reduce computational expense during the MPP searches. Local, multipoint, and global approximations of the limit state are possible. [32] investigated local first-order limit state approximations, and [33] investigated local second-order and multipoint approximations. These techniques include:

1. a single Taylor series per response/reliability/probability level in x-space centered at the uncertain variable means. The first-order approach is commonly known as the Advanced Mean Value (AMV) method:

$$g(\mathbf{x}) \cong g(\mu_{\mathbf{x}}) + \nabla_x g(\mu_{\mathbf{x}})^T (\mathbf{x} - \mu_{\mathbf{x}}) \tag{6.23}$$

and the second-order approach has been named AMV^2:

$$g(\mathbf{x}) \cong g(\mu_{\mathbf{x}}) + \nabla_{\mathbf{x}} g(\mu_{\mathbf{x}})^T (\mathbf{x} - \mu_{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \mu_{\mathbf{x}})^T \nabla^2_{\mathbf{x}} g(\mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}}) \tag{6.24}$$

2. same as AMV/AMV^2, except that the Taylor series is expanded in u-space. The first-order option has been termed the u-space AMV method:

$$G(\mathbf{u}) \cong G(\mu_{\mathbf{u}}) + \nabla_u G(\mu_{\mathbf{u}})^T (\mathbf{u} - \mu_{\mathbf{u}}) \tag{6.25}$$

where $\mu_{\mathbf{u}} = T(\mu_{\mathbf{x}})$ and is nonzero in general, and the second-order option has been named the u-space AMV^2 method:

$$G(\mathbf{u}) \cong G(\mu_{\mathbf{u}}) + \nabla_{\mathbf{u}} G(\mu_{\mathbf{u}})^T (\mathbf{u} - \mu_{\mathbf{u}}) + \frac{1}{2}(\mathbf{u} - \mu_{\mathbf{u}})^T \nabla^2_{\mathbf{u}} G(\mu_{\mathbf{u}})(\mathbf{u} - \mu_{\mathbf{u}}) \tag{6.26}$$

3. an initial Taylor series approximation in x-space at the uncertain variable means, with iterative expansion updates at each MPP estimate ($\mathbf{x}^*$) until the MPP converges. The first-order option is commonly known as AMV+:

$$g(\mathbf{x}) \cong g(\mathbf{x}^*) + \nabla_x g(\mathbf{x}^*)^T (\mathbf{x} - \mathbf{x}^*) \tag{6.27}$$

and the second-order option has been named AMV^2+:

$$g(\mathbf{x}) \cong g(\mathbf{x}^*) + \nabla_{\mathbf{x}} g(\mathbf{x}^*)^T (\mathbf{x} - \mathbf{x}^*) + \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T \nabla^2_{\mathbf{x}} g(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) \tag{6.28}$$

4. same as AMV+/AMV^2+, except that the expansions are performed in u-space. The first-order option has been termed the u-space AMV+ method.

$$G(\mathbf{u}) \cong G(\mathbf{u}^*) + \nabla_u G(\mathbf{u}^*)^T (\mathbf{u} - \mathbf{u}^*) \tag{6.29}$$

and the second-order option has been named the u-space AMV^2+ method:

$$G(\mathbf{u}) \cong G(\mathbf{u}^*) + \nabla_{\mathbf{u}} G(\mathbf{u}^*)^T (\mathbf{u} - \mathbf{u}^*) + \frac{1}{2}(\mathbf{u} - \mathbf{u}^*)^T \nabla^2_{\mathbf{u}} G(\mathbf{u}^*)(\mathbf{u} - \mathbf{u}^*) \tag{6.30}$$

5. a multipoint approximation in x-space. This approach involves a Taylor series approximation in intermediate variables where the powers used for the intermediate variables are selected to match information at the current and previous expansion points. Based on the two-point exponential approximation concept (TPEA, [43]), the two-point adaptive nonlinearity approximation (TANA-3, [146]) approximates the limit state as:

$$g(\mathbf{x}) \cong g(\mathbf{x}_2) + \sum_{i=1}^{n} \frac{\partial g}{\partial x_i}(\mathbf{x}_2) \frac{x_{i,2}^{1-p_i}}{p_i}(x_i^{p_i} - x_{i,2}^{p_i}) + \frac{1}{2}\epsilon(\mathbf{x}) \sum_{i=1}^{n} (x_i^{p_i} - x_{i,2}^{p_i})^2 \tag{6.31}$$

where n is the number of uncertain variables and:

$$p_i = 1 + \ln\left[\frac{\frac{\partial g}{\partial x_i}(\mathbf{x}_1)}{\frac{\partial g}{\partial x_i}(\mathbf{x}_2)}\right] \Big/ \ln\left[\frac{x_{i,1}}{x_{i,2}}\right] \tag{6.32}$$

$$\epsilon(\mathbf{x}) = \frac{H}{\sum_{i=1}^{n}(x_i^{p_i} - x_{i,1}^{p_i})^2 + \sum_{i=1}^{n}(x_i^{p_i} - x_{i,2}^{p_i})^2} \tag{6.33}$$

$$H = 2\left[g(\mathbf{x}_1) - g(\mathbf{x}_2) - \sum_{i=1}^{n} \frac{\partial g}{\partial x_i}(\mathbf{x}_2) \frac{x_{i,2}^{1-p_i}}{p_i}(x_{i,1}^{p_i} - x_{i,2}^{p_i})\right] \tag{6.34}$$

and $\mathbf{x}_2$ and $\mathbf{x}_1$ are the current and previous MPP estimates in x-space, respectively. Prior to the availability of two MPP estimates, x-space AMV+ is used.

6. a multipoint approximation in u-space. The u-space TANA-3 approximates the limit state as:

$$G(\mathbf{u}) \cong G(\mathbf{u}_2) + \sum_{i=1}^{n} \frac{\partial G}{\partial u_i}(\mathbf{u}_2) \frac{u_{i,2}^{1-p_i}}{p_i}(u_i^{p_i} - u_{i,2}^{p_i}) + \frac{1}{2}\epsilon(\mathbf{u}) \sum_{i=1}^{n}(u_i^{p_i} - u_{i,2}^{p_i})^2 \tag{6.35}$$

where:

$$p_i = 1 + \ln\left[\frac{\frac{\partial G}{\partial u_i}(\mathbf{u}_1)}{\frac{\partial G}{\partial u_i}(\mathbf{u}_2)}\right] \Big/ \ln\left[\frac{u_{i,1}}{u_{i,2}}\right] \tag{6.36}$$

$$\epsilon(\mathbf{u}) = \frac{H}{\sum_{i=1}^{n}(u_i^{p_i} - u_{i,1}^{p_i})^2 + \sum_{i=1}^{n}(u_i^{p_i} - u_{i,2}^{p_i})^2} \tag{6.37}$$

$$H = 2\left[G(\mathbf{u}_1) - G(\mathbf{u}_2) - \sum_{i=1}^{n} \frac{\partial G}{\partial u_i}(\mathbf{u}_2) \frac{u_{i,2}^{1-p_i}}{p_i}(u_{i,1}^{p_i} - u_{i,2}^{p_i})\right] \tag{6.38}$$

and $\mathbf{u}_2$ and $\mathbf{u}_1$ are the current and previous MPP estimates in u-space, respectively. Prior to the availability of two MPP estimates, u-space AMV+ is used.

7. the MPP search on the original response functions without the use of any approximations. Combining this option with first-order and second-order integration approaches (see next section) results in the traditional first-order and second-order reliability methods (FORM and SORM).

The Hessian matrices in AMV^2 and AMV^2+ may be available analytically, estimated numerically, or approximated through quasi-Newton updates. The selection between x-space or u-space for performing approximations depends on where the approximation will be more accurate, since this will result in more accurate MPP estimates (AMV, AMV^2) or faster convergence (AMV+, AMV^2+, TANA). Since this relative accuracy depends on the forms of the limit state $g(x)$ and the transformation $T(x)$ and is therefore application dependent in general, DAKOTA supports both options. A concern with approximation-based iterative search methods (i.e., AMV+, AMV^2+ and TANA) is the robustness of their convergence to the MPP. It is possible for the MPP iterates to oscillate or even diverge. However, to date, this occurrence has been relatively rare, and DAKOTA contains checks that monitor for this behavior. Another concern with TANA is numerical safeguarding (e.g., the possibility of raising negative x_i or u_i values to nonintegral p_i exponents in Equations 6.31, 6.33-6.35, and 6.37-6.38). Safeguarding involves offseting negative x_i or u_i and, for potential numerical difficulties with the logarithm ratios in Equations 6.32 and 6.36, reverting to either the linear ($p_i = 1$) or reciprocal ($p_i = -1$) approximation based on which approximation has lower error in $\frac{\partial g}{\partial x_i}(\mathbf{x}_1)$ or $\frac{\partial G}{\partial u_i}(\mathbf{u}_1)$.

Probability integrations

The second algorithmic variation involves the integration approach for computing probabilities at the MPP, which can be selected to be first-order (Equations 6.8-6.9) or second-order integration. Second-order integration involves applying a curvature correction [12, 76, 77]. Breitung applies a correction based on asymptotic analysis [12]:

$$p = \Phi(-\beta_p) \prod_{i=1}^{n-1} \frac{1}{\sqrt{1 + \beta_p \kappa_i}} \tag{6.39}$$

where κ_i are the principal curvatures of the limit state function (the eigenvalues of an orthonormal transformation of $\nabla^2_{\mathbf{u}} G$, taken positive for a convex limit state) and $\beta_p \geq 0$ (a CDF or CCDF probability correction is selected to

Table 6.1: Mapping from DAKOTA options to standard reliability methods.

MPP search	Order of approximation and integration	
	First order	Second order
none	MVFOSM	MVSOSM
x_taylor_mean	AMV	AMV^2
u_taylor_mean	u-space AMV	u-space AMV^2
x_taylor_mpp	AMV+	AMV^2+
u_taylor_mpp	u-space AMV+	u-space AMV^2+
x_two_point	TANA	
u_two_point	u-space TANA	
no_approx	FORM	SORM

obtain the correct sign for β_p). An alternate correction in [76] is consistent in the asymptotic regime ($\beta_p \rightarrow \infty$) but does not collapse to first-order integration for $\beta_p = 0$:

$$p = \Phi(-\beta_p) \prod_{i=1}^{n-1} \frac{1}{\sqrt{1 + \psi(-\beta_p)\kappa_i}} \tag{6.40}$$

where $\psi() = \frac{\phi()}{\Phi()}$ and $\phi()$ is the standard normal density function. [77] applies further corrections to Equation 6.40 based on point concentration methods. At this time, all three approaches are available within the code, but the Breitung correction is used by default (switching the correction is not currently supported in the input specification and requires minor source modification and recompile).

Method mapping

Given settings for limit state approximation, approximation order, integration approach, and other details presented to this point, it is evident that the number of algorithmic combinations is high. Table 6.1 provides a succinct mapping for some of these combinations to common method names from the reliability literature, where blue indicates the most well-known combinations and gray indicates other supported combinations.

Within the DAKOTA specification (refer to the Method Commands chapter within the Reference Manual), the MPP search and integration order selections are explicit in the method specification, but the order of the approximation is inferred from the associated response specification (as is done with local taylor series approximations described in Section 11.4.1). Thus, reliability methods do not have to be synchronized in order as shown in the table; however, it is often desirable to do so.

6.3.3 Global Reliability Methods

Global reliability methods are designed to handle nonsmooth and multimodal failure surfaces, by creating global approximations based on Gaussian process models. They accurately resolve a particular contour of a response function and then estimate probabilities using multimodal adaptive importance sampling.

The global reliability method in DAKOTA is called Efficient Global Reliability Analysis (EGRA) [9]. The name is due to its roots in efficient global optimization (EGO) [81, 79]. The main idea in EGO-type optimization methods is that a global approximation is made of the underlying function. This approximation, which is a Gaussian process model, is used to guide the search by finding points which maximize the expected improvement function (EIF). The EIF is used to select the location at which a new training point should be added to the Gaussian

process model by maximizing the amount of improvement in the objective function that can be expected by adding that point. A point could be expected to produce an improvement in the objective function if its predicted value is better than the current best solution, or if the uncertainty in its prediction is such that the probability of it producing a better solution is high. Because the uncertainty is higher in regions of the design space with fewer observations, this provides a balance between exploiting areas of the design space that predict good solutions, and exploring areas where more information is needed.

The general procedure of these EGO-type methods is:

1. Build an initial Gaussian process model of the objective function.
2. Find the point that maximizes the EIF. If the EIF value at this point is sufficiently small, stop.
3. Evaluate the objective function at the point where the EIF is maximized. Update the Gaussian process model using this new point. Go to Step 2.

Gaussian process (GP) models are used because they provide not just a predicted value at an unsampled point, but also an estimate of the prediction variance. This variance gives an indication of the uncertainty in the GP model, which results from the construction of the covariance function. This function is based on the idea that when input points are near one another, the correlation between their corresponding outputs will be high. As a result, the uncertainty associated with the model's predictions will be small for input points which are near the points used to train the model, and will increase as one moves further from the training points.

The expected improvement function is used in EGO algorithms to select the location at which a new training point should be added. The EIF is defined as the expectation that any point in the search space will provide a better solution than the current best solution based on the expected values and variances predicted by the GP model. It is important to understand how the use of this EIF leads to optimal solutions. The EIF indicates how much the objective function value at a new potential location is expected to be less than the predicted value at the current best solution. Because the GP model provides a Gaussian distribution at each predicted point, expectations can be calculated. Points with good expected values and even a small variance will have a significant expectation of producing a better solution (exploitation), but so will points that have relatively poor expected values and greater variance (exploration).

The application of EGO to reliability analysis, however, is made more complicated due to the inclusion of equality constraints. In forward reliability analysis, the response function appears as a constraint rather than the objective. That is, we want to satisfy the constraint that the response equals a threshold value and is on the limit state: $G(\mathbf{u}) = \bar{z}$. Therefore, the EIF function was modified to focus on feasibility, and instead of using an expected improvement function, we use an expected feasibility function (EFF) [9]. The EFF provides an indication of how well the response is expected to satisfy the equality constraint. Points where the expected value is close to the threshold ($\mu_G \approx \bar{z}$) and points with a large uncertainty in the prediction will have large expected feasibility values.

The general outline of the EGRA algorithm is as follows: LHS sampling is used to generate a small number of samples from the true response function. Then, an initial Gaussian process model is constructed. Based on the EFF, the point with maximum EFF is found using the global optimizer DIRECT. The true response function is then evaluated at this new point, and this point is added to the sample set and the process of building a new GP model and maximizing the EFF is repeated until the maximum EFF is small. At this stage, the GP model is accurate in the vicinity of the limit state. The GP model is then used to calculate the probability of failure using multimodal importance sampling, which is explained below.

One method to calculate the probability of failure is to directly perform the probability integration numerically by sampling the response function. Sampling methods can be prohibitively expensive because they generally require a large number of response function evaluations. Importance sampling methods reduce this expense by focusing the samples in the important regions of the uncertain space. They do this by centering the sampling

density function at the MPP rather than at the mean. This ensures the samples will lie the region of interest, thus increasing the efficiency of the sampling method. Adaptive importance sampling (AIS) further improves the efficiency by adaptively updating the sampling density function. Multimodal adaptive importance sampling [24] is a variation of AIS that allows for the use of multiple sampling densities making it better suited for cases where multiple sections of the limit state are highly probable.

Note that importance sampling methods require that the location of at least one MPP be known because it is used to center the initial sampling density. However, current gradient-based, local search methods used in MPP search may fail to converge or may converge to poor solutions for highly nonlinear problems, possibly making these methods inapplicable. The EGRA algorithm described above does not depend on the availability of accurate gradient information, making convergence more reliable for nonsmooth response functions. Moreover, EGRA has the ability to locate multiple failure points, which can provide multiple starting points and thus a good multimodal sampling density for the initial steps of multimodal AIS. The probability assessment using multimodal AIS thus incorporates probability of failure at multiple points.

6.3.4 Uncertainty Quantification Example using Reliability Analysis

In summary, the user can choose to perform either forward (RIA) or inverse (PMA) mappings when performing a reliability analysis. With either approach, there are a variety of methods from which to choose in terms of limit state approximations (MVFOSM, MVSOSM, x-/u-space AMV, x-/u-space AMV^2, x-/u-space AMV+, x-/u-space AMV^2+, x-/u-space TANA, and FORM/SORM), probability integrations (first-order or second-order), limit state Hessian selection (analytic, finite difference, BFGS, or SR1), and MPP optimization algorithm (SQP or NIP) selections.

All reliability methods output approximate values of the CDF/CCDF response-probability-reliability levels for prescribed response levels (RIA) or prescribed probability or reliability levels (PMA). In addition, the MV methods additionally output estimates of the mean and standard deviation of the response functions along with importance factors for each of the uncertain variables in the case of independent random variables.

This example quantifies the uncertainty in the "log ratio" response function:

$$g(x_1, x_2) = \frac{x_1}{x_2} \tag{6.41}$$

by computing approximate response statistics using reliability analysis to determine the response cumulative distribution function:

$$P[g(x_1, x_2) < \bar{z}] \tag{6.42}$$

where X_1 and X_2 are identically distributed lognormal random variables with means of `1`, standard deviations of `0.5`, and correlation coefficient of `0.3`.

A DAKOTA input file showing RIA using FORM (option 7 in limit state approximations combined with first-order integration) is listed in Figure 6.8. The user first specifies the `nond_local_reliability` method, followed by the MPP search approach and integration order. In this example, we specify `mpp_search no_approx` and utilize the default first-order integration to select FORM. Finally, the user specifies response levels or probability/reliability levels to determine if the problem will be solved using an RIA approach or a PMA approach. In the example figure of 6.8, we use RIA by specifying a range of `response_levels` for the problem. The resulting output for this input is shown in Figure 6.9, with probability and reliability levels listed for each response level. Figure 6.10 shows that FORM compares favorably to an analytic solution for this problem (note: the response levels differ from those shown in Figure 6.9).

If the user specifies `nond_local_reliability` as a method with no additional specification on how to do the MPP search, then no MPP search is done: the Mean Value method is used. The MV results are shown in

```
strategy,
        single_method graphics

method,
        nond_local_reliability
          mpp_search no_approx
          response_levels = .4 .5 .55 .6 .65 .7
           .75 .8 .85 .9 1. 1.05 1.15 1.2 1.25 1.3
           1.35 1.4 1.5 1.55 1.6 1.65 1.7 1.75

variables,
        lognormal_uncertain = 2
          means             =  1.  1
          std_deviations    =  0.5 0.5
          descriptors       =  'TF1ln'   'TF2ln'
        uncertain_correlation_matrix =  1   0.3
                                        0.3 1

interface,
        system asynch
          analysis_driver = 'log_ratio'

responses,
        num_response_functions = 1
        numerical_gradients
          method_source dakota
          interval_type central
          fd_gradient_step_size = 1.e-4
        no_hessians
```

Figure 6.8: DAKOTA input file for Reliability UQ example using FORM.

```
Cumulative Distribution Function (CDF) for response_fn_1:
     Response Level  Probability Level  Reliability Index
     --------------  -----------------  -----------------
   4.0000000000e-01   4.7624085962e-02   1.6683404020e+00
   5.0000000000e-01   1.0346525475e-01   1.2620507942e+00
   5.5000000000e-01   1.3818404972e-01   1.0885143628e+00
   6.0000000000e-01   1.7616275822e-01   9.3008801339e-01
   6.5000000000e-01   2.1641741368e-01   7.8434989943e-01
   7.0000000000e-01   2.5803428381e-01   6.4941748143e-01
   7.5000000000e-01   3.0020938124e-01   5.2379840558e-01
   8.0000000000e-01   3.4226491013e-01   4.0628960782e-01
   8.5000000000e-01   3.8365052982e-01   2.9590705956e-01
   9.0000000000e-01   4.2393548232e-01   1.9183562480e-01
   1.0000000000e+00   5.0000000000e-01   6.8682233460e-12
   1.0500000000e+00   5.3539344228e-01  -8.8834907167e-02
   1.1500000000e+00   6.0043460094e-01  -2.5447217462e-01
   1.2000000000e+00   6.3004131827e-01  -3.3196278078e-01
   1.2500000000e+00   6.5773508987e-01  -4.0628960782e-01
   1.3000000000e+00   6.8356844630e-01  -4.7770089473e-01
   1.3500000000e+00   7.0761025532e-01  -5.4641676380e-01
   1.4000000000e+00   7.2994058691e-01  -6.1263331274e-01
   1.5000000000e+00   7.6981945355e-01  -7.3825238860e-01
   1.5500000000e+00   7.8755158269e-01  -7.9795460350e-01
   1.6000000000e+00   8.0393505584e-01  -8.5576118635e-01
   1.6500000000e+00   8.1906005158e-01  -9.1178881995e-01
   1.7000000000e+00   8.3301386860e-01  -9.6614373461e-01
   1.7500000000e+00   8.4588021938e-01  -1.0189229206e+00
```

Figure 6.9: Output from Reliability UQ example using FORM.

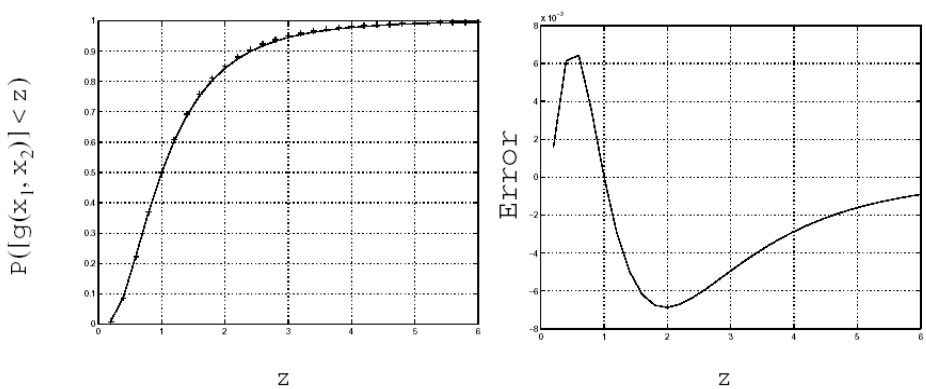


Figure 6.10: Comparison of the cumulative distribution function (CDF) computed by FORM (+ marks) and the exact CDF for $g(x_1, x_2) = \frac{x_1}{x_2}$

```
MV Statistics for response_fn_1:
  Approximate Mean Response                  =  1.0000000000e+00
  Approximate Standard Deviation of Response =  5.9160798127e-01
  Importance Factors not available.
Cumulative Distribution Function (CDF) for response_fn_1:
     Response Level  Probability Level  Reliability Index
     --------------  -----------------  -----------------
   4.0000000000e-01   1.5524721837e-01   1.0141851006e+00
   5.0000000000e-01   1.9901236093e-01   8.4515425050e-01
   5.5000000000e-01   2.2343641149e-01   7.6063882545e-01
   6.0000000000e-01   2.4948115037e-01   6.7612340040e-01
   6.5000000000e-01   2.7705656603e-01   5.9160797535e-01
   7.0000000000e-01   3.0604494093e-01   5.0709255030e-01
   7.5000000000e-01   3.3630190949e-01   4.2257712525e-01
   8.0000000000e-01   3.6765834596e-01   3.3806170020e-01
   8.5000000000e-01   3.9992305332e-01   2.5354627515e-01
   9.0000000000e-01   4.3288618783e-01   1.6903085010e-01
   1.0000000000e+00   5.0000000000e-01   0.0000000000e+00
   1.0500000000e+00   5.3367668035e-01  -8.4515425050e-02
   1.1500000000e+00   6.0007694668e-01  -2.5354627515e-01
   1.2000000000e+00   6.3234165404e-01  -3.3806170020e-01
   1.2500000000e+00   6.6369809051e-01  -4.2257712525e-01
   1.3000000000e+00   6.9395505907e-01  -5.0709255030e-01
   1.3500000000e+00   7.2294343397e-01  -5.9160797535e-01
   1.4000000000e+00   7.5051884963e-01  -6.7612340040e-01
   1.5000000000e+00   8.0098763907e-01  -8.4515425050e-01
   1.5500000000e+00   8.2372893005e-01  -9.2966967555e-01
   1.6000000000e+00   8.4475278163e-01  -1.0141851006e+00
   1.6500000000e+00   8.6405064339e-01  -1.0987005257e+00
   1.7000000000e+00   8.8163821351e-01  -1.1832159507e+00
   1.7500000000e+00   8.9755305196e-01  -1.2677313758e+00
```

Figure 6.11: Output from Reliability UQ example using MV.

Figure 6.11 and consist of approximate mean and standard deviation of the response, the importance factors for each uncertain variable, and approximate probability/reliability levels for the prescribed response levels that have been inferred from the approximate mean and standard deviation (using Equations 6.3 and 6.8). It is evident that the statistics are considerably different from the fully converged FORM results; however, these rough approximations are also much less expensive to calculate. The importance factors are a measure of the sensitivity of the response function(s) to the uncertain input variables, but in this case, are not separable due to the presence of input correlation coefficients.

Additional reliability analysis and design results are provided in Sections 22.5-22.10.

6.4 Stochastic Expansion Methods

The objective of these techniques is to characterize the response of systems whose governing equations involve stochastic coefficients. The development of these techniques mirrors that of deterministic finite element analysis through the utilization of the concepts of projection, orthogonality, and weak convergence. The polynomial chaos

expansion is based on a multidimensional orthogonal polynomial approximation in standardized random variables, and the stochastic collocation approach is based on a multidimensional Lagrange interpolation in standardized random variables. A distinguishing feature of these methodologies is that the final solution is expressed as a random process, and not merely as a set of statistics as is the case for many nondeterministic methodologies. This makes these techniques particularly attractive for use in multi-physics applications which link different analysis packages.

The first stochastic expansion method is the polynomial chaos expansion (PCE) described in Ghanem, et al. [53], [54]. DAKOTA implements the generalized polynomial chaos approach using the Wiener-Askey scheme [145], in which Hermite, Legendre, Laguerre, Jacobi, and generalized Laguerre orthogonal polynomials are used for modeling the effect of continuous random variables described by normal, uniform, exponential, beta, and gamma probability distributions, respectively. These orthogonal polynomial selections are optimal for these distribution types since the inner product weighting function and its corresponding support range correspond to the probability density functions for these continuous distributions. In theory, exponential convergence rates in response moments can be obtained with the optimal basis. When transformations to independent standard random variables (in some cases, approximated by uncorrelated standard random variables) are used, the variable expansions are uncoupled, allowing the polynomial orthogonality properties to be applied on a per-dimension basis. This allows one to mix and match the polynomial basis used for each variable without interference with the spectral projection scheme for the response.

In non-intrusive PCE, simulations are used as black boxes and the calculation of chaos expansion coefficients for response metrics of interest is based on a set of simulation response evaluations. To calculate these response PCE coefficients, two primary classes of approaches have been proposed: spectral projection and linear regression. The spectral projection approach projects the response against each basis function using inner products and employs the polynomial orthogonality properties to extract each coefficient. Each inner product involves a multidimensional integral over the support range of the weighting function, which can be evaluated numerically using quadrature, cubature, or sampling approaches. The linear regression approach (also known as point collocation or stochastic response surfaces) uses a single linear least squares solution to solve for the PCE coefficients which best match a set of response values obtained from a design of computer experiments.

Stochastic collocation (SC) is another stochastic expansion technique for UQ that is closely related to PCE. Whereas PCE estimates coefficients for known orthogonal polynomial basis functions, SC forms Lagrange interpolation functions for known coefficients. Since the ith interpolation function is 1 at collocation point i and 0 for all other collocation points, it is easy to see that the expansion coefficients are just the response values at each of the collocation points. The formation of multidimensional interpolants with this property requires the use of structured collocation point sets derived from tensor product or sparse grids. The key to the approach is performing collocation using the Gauss points and weights from the same optimal orthogonal polynomials used in generalized PCE, which results in the same exponential convergence rates. A key distinction is that, whereas PCE must define an expansion order and a corresponding coefficient estimation approach, SC requires only a collocation grid definition from which the expansion polynomial order is derived based on Lagrange interpolation.

6.4.1 Askey scheme

Table 6.2 shows the set of polynomials which provide an optimal basis for different continuous probability distribution types. It is derived from the family of hypergeometric orthogonal polynomials known as the Askey scheme [6], for which the Hermite polynomials originally employed by Wiener [139] are a subset. The optimality of these basis selections derives from the use of inner product weighting functions that correspond to the probability density functions (PDFs) of the continuous distributions when placed in a standard form. The density and weighting functions differ by a constant factor due to the requirement that the integral of the PDF over the support range is one.

Table 6.2: Linkage between standard forms of continuous probability distributions and Askey scheme of continuous hyper-geometric polynomials.

Distribution	Density function	Polynomial	Weight function	Support range
Normal	$\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}$	Hermite $He_n(x)$	$e^{\frac{-x^2}{2}}$	$[-\infty,\infty]$
Uniform	$\frac{1}{2}$	Legendre $P_n(x)$	1	$[-1,1]$
Beta	$\frac{(1-x)^\alpha(1+x)^\beta}{2^{\alpha+\beta+1}B(\alpha+1,\beta+1)}$	Jacobi $P_n^{(\alpha,\beta)}(x)$	$(1-x)^\alpha(1+x)^\beta$	$[-1,1]$
Exponential	e^{-x}	Laguerre $L_n(x)$	e^{-x}	$[0,\infty]$
Gamma	$\frac{x^\alpha e^{-x}}{\Gamma(\alpha+1)}$	Generalized Laguerre $L_n^{(\alpha)}(x)$	$x^\alpha e^{-x}$	$[0,\infty]$

Note that Legendre is a special case of Jacobi for $\alpha=\beta=0$, Laguerre is a special case of generalized Laguerre for $\alpha=0$, and the Beta function is defined as $B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. Some care is necessary when specifying the α and β parameters for the Jacobi and generalized Laguerre polynomials since the orthogonal polynomial conventions [1] differ from the common statistical PDF conventions. The former conventions are used in Table 6.2.

6.4.2 Interpolation polynomials

Lagrange polynomials interpolate a set of points in a single dimension using the functional form

$$L_j = \prod_{\substack{k=1\\k\neq j}}^{m} \frac{\xi-\xi_k}{\xi_j-\xi_k} \tag{6.43}$$

where it is evident that L_j is 1 at $\xi=\xi_j$, is 0 for each of the points $\xi=\xi_k$, and has order $m-1$.

For interpolation of a response function R in one dimension over m points, the expression

$$R(\xi) \cong \sum_{j=1}^{m} r(\xi_j)\, L_j(\xi) \tag{6.44}$$

reproduces the response values $r(\xi_j)$ at the interpolation points and smoothly interpolates between these values at other points. For interpolation in multiple dimensions, a tensor-product approach is used wherein

$$R(\boldsymbol{\xi}) \cong \sum_{j_1=1}^{m_{i_1}} \cdots \sum_{j_n=1}^{m_{i_n}} r\left(\xi_{j_1}^{i_1},\ldots,\xi_{j_n}^{i_n}\right)\left(L_{j_1}^{i_1}\otimes\cdots\otimes L_{j_n}^{i_n}\right) = \sum_{j=1}^{N_p} r_j(\boldsymbol{\xi})\boldsymbol{L}_j(\boldsymbol{\xi}) \tag{6.45}$$

where $\boldsymbol{i}=(m_1,m_2,\cdots,m_n)$ are the number of nodes used in the n-dimensional interpolation and $\xi_{j_m}^{i_k}$ is the j_mth point in the kth direction. As will be seen later, interpolation on sparse grids involves a summation of these tensor products.

6.4.3 Generalized Polynomial Chaos

The set of polynomials from 6.4.1 are used as an orthogonal basis to approximate the functional form between the stochastic response output and each of its random inputs. The chaos expansion for a response R takes the form

$$R = a_0B_0 + \sum_{i_1=1}^{\infty} a_{i_1}B_1(\xi_{i_1}) + \sum_{i_1=1}^{\infty}\sum_{i_2=1}^{i_1} a_{i_1i_2}B_2(\xi_{i_1},\xi_{i_2}) + \sum_{i_1=1}^{\infty}\sum_{i_2=1}^{i_1}\sum_{i_3=1}^{i_2} a_{i_1i_2i_3}B_3(\xi_{i_1},\xi_{i_2},\xi_{i_3}) + \ldots \tag{6.46}$$

where each additional set of nested summations indicates an additional order of polynomials in the expansion. This expression can be simplified by reformulating the expansion from an order-based indexing to a term-based indexing

$$R = \sum_{j=0}^{\infty} \alpha_j \Psi_j(\boldsymbol{\xi}) \tag{6.47}$$

where there is a one-to-one correspondence between $a_{i_1 i_2 \ldots i_n}$ and α_j and between $B_n(\xi_{i_1}, \xi_{i_2}, ..., \xi_{i_n})$ and $\Psi_j(\boldsymbol{\xi})$. Each of the $\Psi_j(\boldsymbol{\xi})$ are multivariate polynomials which involve products of the one-dimensional polynomials. For example, a multivariate Hermite polynomial $B(\boldsymbol{\xi})$ of order n is defined from

$$B_n(\xi_{i_1}, ..., \xi_{i_n}) = e^{\frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\xi}}(-1)^n \frac{\partial^n}{\partial \xi_{i_1} ... \partial \xi_{i_n}} e^{-\frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\xi}} \tag{6.48}$$

which can be shown to be a product of one-dimensional Hermite polynomials involving a multi-index m_i^j:

$$B_n(\xi_{i_1}, ..., \xi_{i_n}) = \Psi_j(\boldsymbol{\xi}) = \prod_{i=1}^{n} \psi_{m_i^j}(\xi_i) \tag{6.49}$$

The first few multidimensional Hermite polynomials for a two-dimensional case (covering zeroth, first, and second order terms) are then

$$\begin{aligned}
\Psi_0(\boldsymbol{\xi}) &= \psi_0(\xi_1)\,\psi_0(\xi_2) &&= 1 \\
\Psi_1(\boldsymbol{\xi}) &= \psi_1(\xi_1)\,\psi_0(\xi_2) &&= \xi_1 \\
\Psi_2(\boldsymbol{\xi}) &= \psi_0(\xi_1)\,\psi_1(\xi_2) &&= \xi_2 \\
\Psi_3(\boldsymbol{\xi}) &= \psi_2(\xi_1)\,\psi_0(\xi_2) &&= \xi_1^2 - 1 \\
\Psi_4(\boldsymbol{\xi}) &= \psi_1(\xi_1)\,\psi_1(\xi_2) &&= \xi_1\xi_2 \\
\Psi_5(\boldsymbol{\xi}) &= \psi_0(\xi_1)\,\psi_2(\xi_2) &&= \xi_2^2 - 1
\end{aligned}$$

Thus, a generalized polynomial basis is generated by selecting the univariate basis that is most optimal for each random input and then applying the products as defined by the multi-index to define a mixed set of multivariate polynomials. Similarly, the multivariate weighting functions involve a product of the one-dimensional weighting functions and the multivariate quadrature rules involve a tensor product of the one-dimensional quadrature rules.

The use of independent standard random variables is the critical component that allows decoupling of the multidimensional integrals in a mixed basis expansion. The transformation from the original, potentially correlated random variables to uncorrelated standard random variables involves an extended Nataf transformation. When an optimal basis is available, then simple scaling is used when transforming non-standard normal, uniform, exponential, beta, and gamma distributions to their standard forms. When an optimal basis is not available for a particular random variable, then Eqs. 6.15 is used in most cases to transform non-normals to standard normals. The only current exception to this is that loguniform and triangular random variables are currently transformed to standard uniform random variables. In all supported cases (refer to the correlation support table in the Reference Manual), Eq. 6.16 is then applied to remove the correlation. It is assumed in this approach that the resulting uncorrelated standard random variables are also independent. This assumption is valid for uncorrelated standard normal variables (and motivates the popular approach of using a strictly Hermite basis), but is an approximation for uncorrelated standard uniform, exponential, beta, and gamma variables. For independent variables, the multidimensional integrals involved in the inner products of multivariate polynomials decouple to a product of one-dimensional integrals involving only the particular polynomial basis and corresponding weight function selected for each random dimension. The multidimensional inner products are nonzero only if each of the one-dimensional inner products is nonzero, which preserves the desired multivariate orthogonality properties for the case of a mixed basis.

In practice, one truncates the infinite expansion at a finite number of random variables and a finite expansion order

$$R = \sum_{j=0}^{P} \alpha_j \Psi_j(\boldsymbol{\xi}) \tag{6.50}$$

where the total number of terms N in a complete polynomial chaos expansion of arbitrary order p for a response function involving n uncertain input variables is given by

$$N \;=\; 1 + P \;=\; 1 + \sum_{s=1}^{p} \frac{1}{s!} \prod_{r=0}^{s-1} (n+r) \;=\; \frac{(n+p)!}{n!p!} \tag{6.51}$$

6.4.4 Stochastic Collocation

The SC expansion is formed as a sum of a set of multidimensional Lagrange interpolation polynomials, one polynomial per collocation point. Since these polynomials have the feature of being equal to 1 at their particular collocation point and 0 at all other points, the coefficients of the expansion are just the response values at each of the collocation points. This can be written as:

$$R \cong \sum_{j=1}^{N_p} r_j \boldsymbol{L}_j(\boldsymbol{\xi}) \tag{6.52}$$

where the set of collocation points could involve a tensor product grid as shown in Eq. 6.45 or a sum of tensor products defined from a sparse grid (refer to Section 6.4.5).

The major practical difference between PCE and SC is that, in PCE, one must estimate the coefficients for known basis functions, whereas in SC, one must form the interpolants for known coefficients. PCE estimates its coefficients using any of the approaches to follow: random sampling, tensor-product quadrature, Smolyak sparse grids, or linear regression. In SC, the multidimensional interpolants need to be formed over structured data sets, such as point sets from quadrature or sparse grids; approaches based on random sampling may not be used.

6.4.5 Spectral projection

The spectral projection approach projects the response against each basis function using inner products and employs the polynomial orthogonality properties to extract each coefficient. Similar to a Galerkin projection, the residual error from the approximation is rendered orthogonal to the selected basis. From Eq. 6.50, it is evident that

$$\alpha_j \;=\; \frac{\langle R, \Psi_j \rangle}{\langle \Psi_j^2 \rangle} \;=\; \frac{1}{\langle \Psi_j^2 \rangle} \int_{\Omega} R\, \Psi_j\, \varrho(\boldsymbol{\xi})\, d\boldsymbol{\xi}, \tag{6.53}$$

where each inner product involves a multidimensional integral over the support range of the weighting function. In particular, $\Omega = \Omega_1 \otimes \cdots \otimes \Omega_n$, with possibly unbounded intervals $\Omega_j \subset \mathbb{R}$ and the tensor product form $\varrho(\boldsymbol{\xi}) = \prod_{i=1}^{n} \varrho_i(\xi_i)$ of the joint probability density (weight) function. The denominator in Eq. 6.53 is the norm squared of the multivariate orthogonal polynomial, which can be computed analytically using the product of univariate norms squared

$$\langle \Psi_j^2 \rangle \;=\; \prod_{i=1}^{n} \langle \psi_{m_i^j}^2 \rangle \tag{6.54}$$

where the univariate inner products have simple closed form expressions for each polynomial in the Askey scheme [1]. Thus, the primary computational effort resides in evaluating the numerator, which is evaluated numerically using sampling, quadrature or sparse grid approaches (and this numerical approximation leads to use of the term "pseudo-spectral").

Sampling

In the sampling approach, the integral evaluation is equivalent to computing the expectation (mean) of the response-basis function product (the numerator in Eq. 6.53) for each term in the expansion when sampling within the density of the weighting function.

In computational practice, coefficient estimations based on sampling benefit from first estimating the response mean (the first PCE coefficient) and then removing the mean from the expectation evaluations for all subsequent coefficients. While this has no effect for quadrature/sparse grid methods (see following two sections) and little effect for fully-resolved sampling, it does have a small but noticeable beneficial effect for under-resolved sampling.

Tensor product quadrature

In quadrature-based approaches, the simplest general technique for approximating multidimensional integrals, as in Eq. 6.53, is to employ a tensor product of one-dimensional quadrature rules. In the case where Ω is a hypercube, i.e. $\Omega = [-1,1]^n$, there are several choices of nested abscissas, included Clenshaw-Curtis, Gauss-Patterson, etc. [96, 97, 52]. In the more general case, we propose to use Gaussian abscissas, i.e. the zeros of the orthogonal polynomials with respect to some positive weight, e.g. Gauss-Hermite, Gauss-Legendre, Gauss-Laguerre, generalized Gauss-Laguerre, or Gauss-Jacobi.

We first introduce an index $i \in \mathbb{N}_+$, $i \geq 1$. Then, for each value of i, let $\{\xi_1^i, \ldots, \xi_{m_i}^i\} \subset \Omega_i$ be a sequence of abscissas for quadrature on Ω_i. For $f \in C^0(\Omega_i)$ and $n = 1$ we introduce a sequence of one-dimensional quadrature operators

$$\mathscr{U}^i(f)(\xi) = \sum_{j=1}^{m_i} f(\xi_j^i)\, w_j^i, \tag{6.55}$$

with $m_i \in \mathbb{N}$ given. When utilizing Gaussian quadrature, Eq. 6.55 integrates exactly all polynomials of degree less than $2m_i - 1$, for each $i = 1, \ldots, n$. Given an expansion order p, the highest order coefficient evaluations (Eq. 6.53) can be assumed to involve integrands of at least polynomial order $2p$ (Ψ of order p and R modeled to order p) in each dimension such that a minimal Gaussian quadrature order of $p+1$ will be required to obtain good accuracy in these coefficients.

Now, in the multivariate case $n > 1$, for each $f \in C^0(\Omega)$ and the multi-index $\mathbf{i} = (i_1, \ldots, i_n) \in \mathbb{N}_+^n$ we define the full tensor product quadrature formulas

$$\mathcal{Q}_{\mathbf{i}}^n f(\xi) = \left(\mathscr{U}^{i_1} \otimes \cdots \otimes \mathscr{U}^{i_n}\right)(f)(\boldsymbol{\xi}) = \sum_{j_1=1}^{m_{i_1}} \cdots \sum_{j_n=1}^{m_{i_n}} f\left(\xi_{j_1}^{i_1}, \ldots, \xi_{j_n}^{i_n}\right) \left(w_{j_1}^{i_1} \otimes \cdots \otimes w_{j_n}^{i_n}\right). \tag{6.56}$$

Clearly, the above product needs $\prod_{j=1}^n m_{i_j}$ function evaluations. Therefore, when the number of input random variables is small, full tensor product quadrature is a very effective numerical tool. On the other hand, approximations based on tensor product grids suffer from the *curse of dimensionality* since the number of collocation points in a tensor grid grows exponentially fast in the number of input random variables. For example, if Eq. 6.56 employs the same order for all random dimensions, $m_{i_j} = m$, then Eq. 6.56 requires m^n function evaluations.

Smolyak sparse grids

If the number of random variables is moderately large, one should rather consider sparse tensor product spaces as first proposed by Smolyak [121] and further investigated by Refs. [52, 7, 47, 144, 96, 97] that reduce dramatically the number of collocation points, while preserving a high level of accuracy.

Here we follow the notation and extend the description in Ref. [96] to describe the Smolyak *isotropic* formulas $\mathscr{A}(\mathrm{w}, n)$, where w is a level that is independent of dimension[1]. The Smolyak formulas are just linear combinations of the product formulas in Eq. 6.56 with the following key property: only products with a relatively small number of points are used. With $\mathscr{U}^0 = 0$ and for $i \geq 1$ define

$$\Delta^i = \mathscr{U}^i - \mathscr{U}^{i-1}. \tag{6.57}$$

and we set $|\mathbf{i}| = i_1 + \cdots + i_n$. Then the isotropic Smolyak quadrature formula is given by

$$\mathscr{A}(\mathrm{w}, n) = \sum_{|\mathbf{i}| \leq \mathrm{w}+n} \left(\Delta^{i_1} \otimes \cdots \otimes \Delta^{i_n} \right). \tag{6.58}$$

Equivalently, formula Eq. 6.58 can be written as [137]

$$\mathscr{A}(\mathrm{w}, n) = \sum_{\mathrm{w}+1 \leq |\mathbf{i}| \leq \mathrm{w}+n} (-1)^{\mathrm{w}+n-|\mathbf{i}|} \binom{n-1}{\mathrm{w}+n-|\mathbf{i}|} \cdot \left(\mathscr{U}^{i_1} \otimes \cdots \otimes \mathscr{U}^{i_n} \right). \tag{6.59}$$

Given an index set of levels, growth rules must be defined for the one-dimensional quadrature orders. In order to take advantage of nesting and provide similar growth behavior for fully nested and weakly nested integration rules, the following growth rules are currently employed:

$$\text{Clenshaw} - \text{Curtis}: \; m = \begin{cases} 1 & w = 0 \\ 2^w + 1 & w \geq 1 \end{cases} \tag{6.60}$$

$$\text{Gaussian}: \; m = 2^{w+1} - 1 \tag{6.61}$$

Examples of isotropic sparse grids, constructed from the fully nested Clenshaw-Curtis abscissas and the weakly-nested Gaussian abscissas are shown in Figure 6.12, where $\Omega = [-1, 1]^2$. There, we consider a two-dimensional parameter space and a maximum level $\mathrm{w} = 5$ (sparse grid $\mathscr{A}(5, 2)$). To see the reduction in function evaluations with respect to full tensor product grids, we also include a plot of the corresponding Clenshaw-Curtis isotropic full tensor grid having the same maximum number of points in each direction, namely $2^{\mathrm{w}} + 1 = 33$. Whereas an isotropic tensor-product quadrature scales as m^n, an isotropic sparse grid scales as $m^{\log n}$, significantly mitigating the curse of dimensionality.

6.4.6 Linear regression

The linear regression approach (also known as point collocation or stochastic response surfaces [135, 78]) uses a single linear least squares solution of the form:

$$\boldsymbol{\Psi}\boldsymbol{\alpha} = \boldsymbol{R} \tag{6.62}$$

to solve for the complete set of PCE coefficients $\boldsymbol{\alpha}$ that best match a set of response values $\boldsymbol{R}$. The set of response values is typically obtained by performing a design of computer experiments within the density function

[1] Other common formulations use a dimension-dependent level q where $q \geq n$. We use $w = q - n$, where $w \geq 0$ for all n.

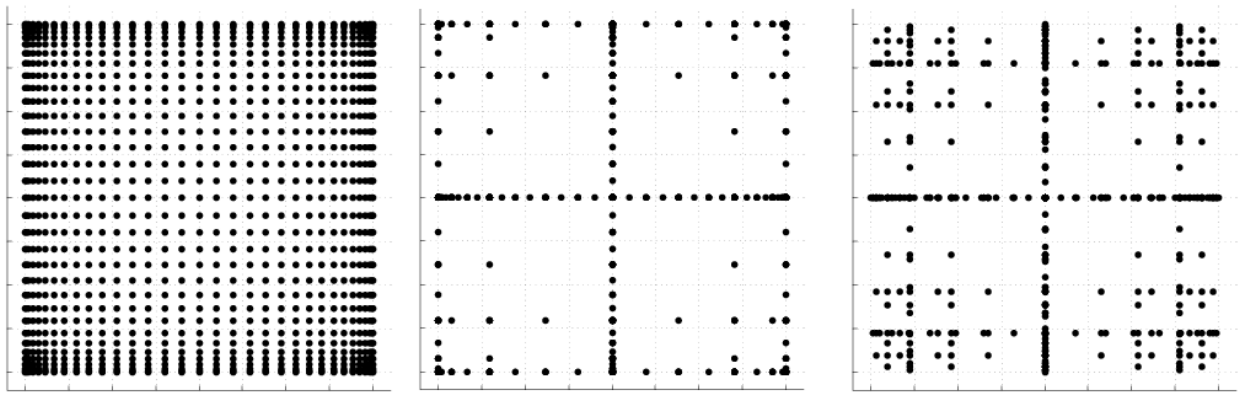


Figure 6.12: For a two-dimensional parameter space ($n = 2$) and maximum level $\mathrm{w} = 5$, we plot the full tensor product grid using the Clenshaw-Curtis abscissas (left) and isotropic Smolyak sparse grids $\mathscr{A}(5, 2)$, utilizing the Clenshaw-Curtis abscissas (middle) and the Gaussian abscissas (right).

of $\boldsymbol{\xi}$, where each row of the matrix $\boldsymbol{\Psi}$ contains the N_t multivariate polynomial terms Ψ_j evaluated at a particular $\boldsymbol{\xi}$ sample. An over-sampling is generally advisable (Ref. [78] recommends $2N_t$ samples), resulting in a least squares solution for the over-determined system. In the case of $2N_t$ oversampling, the simulation requirements for this approach scale as $\frac{2(n+p)!}{n!p!}$, which can be significantly more affordable than isotropic tensor-product quadrature (e.g., $(p+1)^n$) for larger problems.

A closely related technique is known as the "probabilistic collocation" approach. Rather than employing random over-sampling, this technique uses a selected subset of N_t Gaussian quadrature points (those with highest tensor-product weighting), which provides more optimal collocation locations and preserves interpolation properties. Finally, additional regression equations can be obtained through the use of derivative information (gradient and Hessians) from each collocation point, which aids greatly in scaling with respect to the number of random variables. Neither of these latter capabilities are currently available in DAKOTA.

6.4.7 Analytic moments

Mean and variance of the polynomial chaos expansion are available in simple closed form:

$$\mu_R = \langle R \rangle \cong \sum_{j=0}^{P} \alpha_j \langle \Psi_j(\boldsymbol{\xi}) \rangle = \alpha_0 \tag{6.63}$$

$$\sigma_R^2 = \langle (R - \mu_R)^2 \rangle \cong \langle (\sum_{j=1}^{P} \alpha_j \Psi_j(\boldsymbol{\xi}))^2 \rangle = \sum_{j=1}^{P} \sum_{k=1}^{P} \alpha_j \alpha_k \langle \Psi_j(\boldsymbol{\xi}) \Psi_k(\boldsymbol{\xi}) \rangle = \sum_{j=1}^{P} \alpha_j^2 \langle \Psi_j^2 \rangle \tag{6.64}$$

where the norm squared of each multivariate polynomial is computed from Eq. 6.54. The moments μ_R and σ_R are exact moments of the expansion, which converge to moments of the true response function. Higher moments are also available analytically and could be employed in moment fitting approaches (i.e., Pearson and Johnson models) in order to approximate a response PDF, although this is outside the scope of the current paper.

Similar expressions can be derived for stochastic collocation:

$$\mu_R = \langle R \rangle \cong \sum_{j=1}^{N_p} r_j \langle \boldsymbol{L}_j(\boldsymbol{\xi}) \rangle = \sum_{j=1}^{N_p} r_j w_j \tag{6.65}$$

$$\sigma_R^2 = \langle R^2 \rangle - \mu_R^2 \cong \sum_{j=1}^{N_p} \sum_{k=1}^{N_p} r_j r_k \langle \boldsymbol{L}_j(\boldsymbol{\xi}) \boldsymbol{L}_k(\boldsymbol{\xi}) \rangle - \mu_R^2 = \sum_{j=1}^{N_p} r_j^2 w_j - \mu_R^2 \tag{6.66}$$

where the expectation of a particular Lagrange polynomial constructed at Gauss points and then integrated at these same Gauss points leaves only the weight corresponding to the point for which the interpolation value is one.

6.4.8 Random variable sensitivity analysis

Polynomial chaos expansions are easily differentiated with respect to the random variables [109]. First, using Eq. 6.50,

$$\frac{dR}{d\xi_i} = \sum_{j=0}^{P} \alpha_j \frac{d\Psi_j(\boldsymbol{\xi})}{d\xi_i} \tag{6.67}$$

and then using Eq. 6.49,

$$\frac{d\Psi_j(\boldsymbol{\xi})}{d\xi_i} = \frac{d\psi_i}{d\xi_i} \prod_{\substack{k=1 \\ k \neq i}}^{n} \psi_{m_k^j}(\xi_k) \tag{6.68}$$

where the univariate polynomial derivatives $\frac{d\psi_i}{d\xi_i}$ have simple closed form expressions for each polynomial in the Askey scheme [1]. Finally, using the Jacobian of the (extended) Nataf variable transformation,

$$\frac{dR}{dx_i} = \frac{dR}{d\boldsymbol{\xi}} \frac{d\boldsymbol{\xi}}{dx_i} \tag{6.69}$$

which simplifies to $\frac{dR}{d\xi_i} \frac{d\xi_i}{dx_i}$ in the case of uncorrelated x_i.

Similar expressions may be derived for stochastic collocation, starting from Eq. 6.52:

$$\frac{dR}{d\xi_i} = \sum_{j=1}^{N_p} r_j \frac{d\boldsymbol{L}_j(\boldsymbol{\xi})}{d\xi_i} \tag{6.70}$$

where the multidimensional interpolant $\boldsymbol{L}_j$ is formed over either tensor-product quadrature points or a Smolyak sparse grid. For the former case, the derivative of the multidimensional interpolant $\boldsymbol{L}_j$ involves a product rule of the one-dimensional interpolants L_k:

$$\frac{d\boldsymbol{L}_j(\boldsymbol{\xi})}{d\xi_i} = \frac{dL_i}{d\xi_i} \prod_{\substack{k=1 \\ k \neq i}}^{n} L_k(\xi_k) \tag{6.71}$$

and for the latter case, the derivative involves a linear combination of these product rules, as dictated by the Smolyak recursion shown in Eq. 6.59. Finally, calculation of $\frac{dR}{dx_i}$ involves the same Jacobian application shown in Eq. 6.69.

```
strategy,
        single_method #graphics

method,
        nond_polynomial_chaos
          expansion_order    = 4
          quadrature_order   = 5 4
          samples = 10000
          seed    = 12347
          response_levels =
          .1 1. 50. 100.
          500. 1000.

variables,
        uniform_uncertain = 2
          lower_bounds      = -2.  -2.
          upper_bounds      =  2.   2.
          descriptors       = 'x1' 'x2'

interface,
        direct
        analysis_driver = 'rosenbrock'

responses,
        num_response_functions = 1
        no_gradients
        no_hessians
```

Figure 6.13: DAKOTA input file for performing UQ using polynomial chaos expansions.

6.4.9 Uncertainty Quantification Example using Polynomial Chaos

A typical DAKOTA input file for performing an uncertainty quantification using polynomial chaos expansions is shown in Figure 6.13. In this example, we compute CDF probabilities for six response levels of Rosenbrock's function. Since Rosenbrock is a fourth order polynomial and we employ a fourth-order expansion using an optimal basis (Legendre for uniform random variables), we can readily obtain a polynomial expansion which exactly matches the Rosenbrock function. In this example, we select Gaussian quadratures using an anisotropic approach (fifth-order quadrature in x_1 and fourth-order quadrature in x_2), resulting in a total of 20 function evaluations to compute the PCE coefficients.

Once the expansion coefficients have been calculated, some statistics are available analytically and others must be evaluated numerically. For the numerical portion, the input file specifies the use of 10000 samples, which will be evaluated on the expansion to compute the CDF probabilities. In Figure 6.14, excerpts from the results summary are presented, where we first see a summary of the PCE coefficients which exactly reproduce Rosenbrock for a Legendre polynomial basis. The analytic statistics for mean, standard deviation, and COV are then presented, as computed based on Eqs. 6.63-6.64, as well as analytic random variable sensitivities, as computed from Eqs. 6.67-6.69, evaluated at the mean values. Finally, we see the numerical results for the CDF probabilities based on 10000 samples performed on the expansion.

```
Polynomial Chaos coefficients for response_fn_1:
        coefficient   u1   u2
        ----------- ---- ----
   4.5566666667e+02   P0   P0
  -4.0000000000e+00   P1   P0
  -5.3333333333e+02   P0   P1
   9.1695238095e+02   P2   P0
   1.7585932710e-13   P1   P1
   2.6666666667e+02   P0   P2
  -1.9895196601e-13   P3   P0
  -1.0666666667e+03   P2   P1
   1.3988810110e-13   P1   P2
   1.3988810110e-14   P0   P3
   3.6571428571e+02   P4   P0
   1.0258460748e-13   P3   P1
   2.8865798640e-13   P2   P2
  -7.9269923958e-14   P1   P3
   1.2008391835e-13   P0   P4
-------------------------------------------------------------------
Statistics derived analytically from PC expansion:

Moments for each response function:
response_fn_1:  Mean = 4.55667e+02  Std. Dev. = 6.06560e+02
                Coeff. of Variation = 1.33115e+00

Sensitivities for each response function evaluated at uncertain variable means:
response_fn_1:
 [ -2.0000000000e+00 -6.7335026444e-14 ]

Statistics based on 10000 samples performed on PC expansion:

Probabilities for each response function:
Cumulative Distribution Function (CDF) for response_fn_1:
     Response Level  Probability Level  Reliability Index  General Rel Index
     --------------  -----------------  -----------------  -----------------
   1.0000000000e-01   1.9000000000e-03
   1.0000000000e+00   1.3600000000e-02
   5.0000000000e+01   2.4390000000e-01
   1.0000000000e+02   3.4230000000e-01
   5.0000000000e+02   7.1090000000e-01
   1.0000000000e+03   8.5240000000e-01
```

Figure 6.14: Excerpt of UQ output for polynomial chaos example.

6.5 Epistemic Nondeterministic Methods

Uncertainty quantification is often used as part of the risk assessment of performance, reliability, and safety of engineered systems. Increasingly, uncertainty is separated into two categories for analysis purposes: aleatory and epistemic uncertainty [99, 74]. Aleatory uncertainty is also referred to as variability, irreducible or inherent uncertainty, or uncertainty due to chance. Examples of aleatory uncertainty include the height of individuals in a population, or the temperature in a processing environment. Aleatory uncertainty is usually modeled with probability distributions, and sampling methods such as Latin Hypercube sampling in DAKOTA can be used to model aleatory uncertainty. In contrast, epistemic uncertainty refers to lack of knowledge or lack of information about a particular aspect of the simulation model, including the system and environment being modeled. An increase in knowledge or information relating to epistemic uncertainty will lead to a reduction in the predicted uncertainty of the system response or performance. For epistemic uncertain variables, typically one does not know enough to specify a probability distribution on a variable. Epistemic uncertainty is referred to as subjective, reducible, or lack of knowledge uncertainty. Examples of epistemic uncertainty include little or no experimental data for a fixed but unknown physical parameter, incomplete understanding of complex physical phenomena, uncertainty about the correct model form to use, etc.

There are many approaches which have been developed to model epistemic uncertainty, including fuzzy set theory, possibility theory, and evidence theory. We have chosen to pursue evidence theory at Sandia for modeling epistemic uncertainty, in part because evidence theory is a generalization of probability theory. Evidence theory is also referred to as Dempster-Shafer theory or the theory of random sets [99]. In evidence theory, there are two complementary measures of uncertainty: belief and plausibility. Together, belief and plausibility can be thought of as defining lower and upper bounds, respectively, on probabilities. Belief and plausibility define the lower and upper limits or intervals on probability values. Typical plots of cumulative and complementary cumulative belief and plausibility functions are shown in Figure 6.15 [74]. In evidence theory, it is not possible to specify one probability value. Instead, there is a range of values that is consistent with the evidence. The range of values is defined by belief and plausibility. Note that no statement or claim is made about one value within an interval being more or less likely than any other value.

This section focuses on Dempster-Shafer evidence theory. We also use a technique called second-order probability to perform uncertainty quantification when there is both epistemic and aleatory uncertainty present. Second-order probability is a nested technique with two levels of uncertainty quantification. The outer level UQ is typically linked to epistemic uncertainties and the inner level UQ is commonly associated with aleatory uncertainties. A common approach used is to sample possible realizations of epistemic variables in the outer loop, then send these to the inner loop for additional sampling over the aleatory variables. In this way one generates "families" or ensembles of cumulative distribution functions, where each individual CDF is based on aleatory uncertainty, and the ensemble is based on epistemic uncertainty. See Section 11.6.1 for more details.

In Dempster-Shafer evidence theory, the uncertain input variables are modeled as sets of intervals. The user assigns a basic probability assignment (BPA) to each interval, indicating how likely it is that the uncertain input falls within the interval. The BPAs for a particular uncertain input variable must sum to one. The intervals may be overlapping, contiguous, or have gaps. In DAKOTA, an interval uncertain variable is specified as `interval_uncertain`. When one defines an interval type variable in DAKOTA, it is also necessary to specify the number of intervals defined for each variable with `iuv_num_intervals` as well the basic probability assignments per interval, `iuv_interval_probs`, and the associated bounds per each interval, `iuv_interval_bounds`. Figure 6.16 shows the input specification for interval uncertain variables. The example shown in Figure 6.16 has two epistemic uncertain interval variables. The first uncertain variable has three intervals and the second has two. The basic probability assignments for the first variable are 0.5, 0.1, and 0.4, while the BPAs for the second variable are 0.7 and 0.3. Note that it is possible (and often the case) to define an interval uncertain variable with only ONE interval. This means that you only know that the possible value of that variable falls within the interval, and the BPA for that interval would be 1.0. In the case we have shown, the

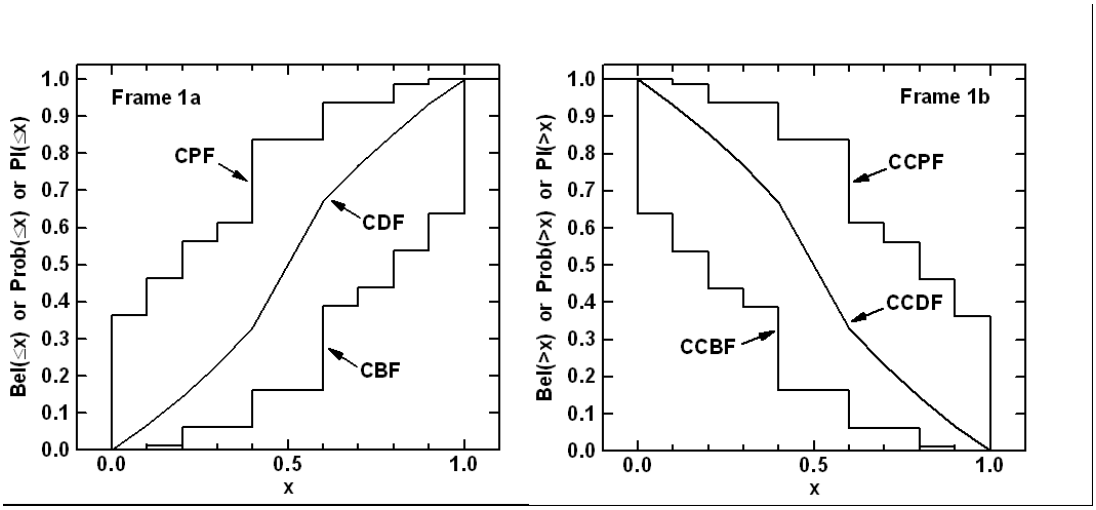


Figure 6.15: Example cumulative belief and plausibility distribution functions on left; complementary cumulative belief and plausibility distribution functions on right

interval bounds on the first interval for the first variable are 0.6 and 0.9, and the bounds for the second interval for the first variable are 0.1 to 0.5, etc.

Once the intervals, the BPAs, and the interval bounds are defined, the user can run an epistemic analysis by specifying the method as `nond_evidence` in the DAKOTA input file. The intervals and their associated BPAs are then propagated through the simulation to obtain cumulative distribution functions on belief and plausibility. As mentioned above, belief is the lower bound on a probability estimate that is consistent with the evidence, and plausibility is the upper bound on a probability estimate that is consistent with the evidence. Figure 6.17 shows results for the first response function obtained when running the example in Figure 6.16. In this example, there are 6 output intervals (as a result of the 2 interval input variables with 3 and 2 intervals, respectively). The output intervals are ordered to obtain cumulative bound functions for both belief and plausibility. The cumulative distribution function is presented for both belief (CBF) and plausibility (CPF). The CBF value is the cumulative belief corresponding to a certain output value. For example, the belief that the output value is less than or equal to 0.2 for response 1 is 0.27, and the plausibility that the output is less than or equal to 0.2 is 1 for response 1. The belief that the output value is less than 0.6217 is 0.75, while the plausbility that the output is less than 0.0806 is 0.75. The CBF and CPF may be plotted on a graph and interpreted as bounding the cumulative distribution function (CDF), which is the probability that the output is less than or equal to a certain value. The interval bounds on probability values show the value of epistemic uncertainty analysis: the intervals are usually much larger than expected, giving one a truer picture of the total output uncertainty caused by lack of knowledge or information about the epistemic input quantities.

As in other nondeterministic methods, with `nond_evidence`, one can specify probability levels and response levels. If response levels are specified, the belief and plausibility function values corresponding to those response levels are calculated (see Belief Prob Level and Plaus Prob Level in the tables shown in Figure 6.17). Similarly, if probability levels are specified, these are first interpreted to be belief values, and the corresponding response levels are calculated (see Belief Resp Level); then they are interpreted to be plausibility values and the corresponding response levels are calculated (see Plaus Resp Level in the table in Figure 6.17). We have recently added the capability to support generalized reliability mappings in `nond_evidence`. If the user specifies a generalized reliability level, it will be first converted to a probability, then interpreted as a belief and plausibility and the corresponding response levels will be calculated. Likewise, if response levels are specified, the corresponding

```
strategy,
        single_method
          tabular_graphics_data

method,
        nond_evidence
          samples = 1000
          seed = 59334
          response_levels = 0.001 0.03 0.2 0.8 0.001 0.2 0.6 0.8
          probability_levels = 0.25 0.5 0.75 0.25 0.5 0.75
          distribution cumulative

variables,
        interval_uncertain = 2
          num_intervals  =  3 2
          interval_probs = 0.5 0.1 0.4 0.7 0.3
          interval_bounds = 0.6 0.9 0.1 0.5 0.5 1.0 0.3 0.5 0.6 0.8

interface,
        direct
          analysis_driver = 'text_book'

responses,
        num_response_functions = 2
        no_gradients
        no_hessians
```

Figure 6.16: DAKOTA input file for UQ example using Evidence Theory.

```
Belief and Plausibility for each response function:
Cumulative Belief/Plausibility Functions (CBF/CPF) for response_fn_1:
     Response Level  Belief Prob Level   Plaus Prob Level
     --------------  -----------------   ----------------
   1.0000000000e-03   0.0000000000e+00   0.0000000000e+00
   3.0000000000e-02   0.0000000000e+00   2.7000000000e-01
   2.0000000000e-01   2.7000000000e-01   1.0000000000e+00
   8.0000000000e-01   9.3000000000e-01   1.0000000000e+00
  Probability Level  Belief Resp Level   Plaus Resp Level
  -----------------  -----------------   ----------------
   2.5000000000e-01   2.6187288772e-01   6.2609206069e-02
   5.0000000000e-01   2.9829775860e-01   6.3736734971e-02
   7.5000000000e-01   6.2173551556e-01   8.0596931719e-02
-----------------------------------------------------------------
```

Figure 6.17: Results of an Epistemic Uncertainty Quantification using Evidence Theory.

belief and plausibility values will be mapped to bounds on the generalized reliability levels.

6.6 Future Nondeterministic Methods

Uncertainty analysis methods under investigation for future inclusion into the DAKOTA framework include extensions to the reliability techniques and sampling capabilities supported. Advanced "smart sampling" techniques such as bootstrap sampling (BS) and Markov chain Monte Carlo simulation (McMC) are being investigated. Efforts have been initiated to allow for the possibility of non-traditional representations of uncertainty. We have implemented Dempster-Shafer theory of evidence, but may also pursue possibility theory or fuzzy sets, and combinations of epistemic and aleatory uncertainty methods. Finally, the tractability and efficacy of the more intrusive variant of stochastic finite element/polynomial chaos expansion methods, previously mentioned, is being assessed for possible implementation in DAKOTA.

Chapter 7

Optimization Capabilities

7.1 Overview

DAKOTA's optimization capabilities include a variety of gradient-based and nongradient-based optimization methods. Numerous packages are available, some of which are commercial packages, some of which are developed internally to Sandia, and some of which are free software packages from the open source community. The downloaded version of DAKOTA excludes the commercially developed packages but includes APPSPACK, COLINY, CONMIN, JEGA, OPT++, and PICO. Interfaces to DOT, NPSOL, and NLPQL are provided with DAKOTA, but to use these commercial optimizers, the user must obtain a software license and the source code for these packages separately. The commercial software can then be compiled into DAKOTA by following DAKOTA's installation procedures (see notes in `Dakota/INSTALL`).

DAKOTA's input commands permit the user to specify two-sided nonlinear inequality constraints of the form $g_{L_i} \leq g_i(\mathbf{x}) \leq g_{U_i}$, as well as nonlinear equality constraints of the form $h_j(\mathbf{x}) = h_{t_j}$ (see also Section 1.4.1). Some optimizers (e.g., NPSOL, OPT++, JEGA) can handle these constraint forms directly, whereas other optimizers (e.g., APPSPACK, DOT, CONMIN) require DAKOTA to perform an internal conversion of all constraints to one-sided inequality constraints of the form $g_i(\mathbf{x}) \leq 0$. In the latter case, the two-sided inequality constraints are treated as $g_i(\mathbf{x}) - g_{U_i} \leq 0$ and $g_{L_i} - g_i(\mathbf{x}) \leq 0$ and the equality constraints are treated as $h_j(\mathbf{x}) - h_{t_j} \leq 0$ and $h_{t_j} - h_j(\mathbf{x}) \leq 0$. The situation is similar for linear constraints: APPSPACK, NPSOL, OPT++, and JEGA support them directly, whereas DOT and CONMIN do not. For linear inequalities of the form $a_{L_i} \leq \mathbf{a}_i^T\mathbf{x} \leq a_{U_i}$ and linear equalities of the form $\mathbf{a}_i^T\mathbf{x} = a_{t_j}$, the nonlinear constraint arrays in DOT and CONMIN are further augmented to include $\mathbf{a}_i^T\mathbf{x} - a_{U_i} \leq 0$ and $a_{L_i} - \mathbf{a}_i^T\mathbf{x} \leq 0$ in the inequality case and $\mathbf{a}_i^T\mathbf{x} - a_{t_j} \leq 0$ and $a_{t_j} - \mathbf{a}_i^T\mathbf{x} \leq 0$ in the equality case. Awareness of these constraint augmentation procedures can be important for understanding the diagnostic data returned from the DOT and CONMIN algorithms. Other optimizers fall somewhere in between. NLPQL supports nonlinear equality constraints $h_j(\mathbf{x}) = 0$ and nonlinear one-sided inequalities $g_i(\mathbf{x}) \geq 0$, but does not natively support linear constraints. Constraint mappings are used with NLPQL for both linear and nonlinear cases. Most COLINY methods now support two-sided nonlinear inequality constraints and nonlinear constraints with targets, but do not natively support linear constraints. Constraint augmentation is not currently used with COLINY, since linear constraints will soon be supported natively.

When gradient and Hessian information is used in the optimization, derivative components are most commonly computed with respect to the active continuous variables, which in this case are the *continuous design variables*. This differs from parameter study methods (for which all continuous variables are active) and from nondeterministic analysis methods (for which the uncertain variables are active). Refer to Section 14.3 for additional

information on derivative components and active continuous variables.

7.2 Optimization Software Packages

7.2.1 APPSPACK Library

Asynchronous Parallel Pattern Search (APPS) [64] is available as method `asynch_pattern_search`. It is an asynchronous implementation of generating set search. APPSPACK can handle unconstrained problems as well as those with bound constraints, linear constraints [66], and general nonlinear constraints [65]. APPSPACK is best suited for simulation-based problems with less than 100 variables. It has significant advantages over gradient-based methods when the function is noisy and/or discontinuous. APPSPACK leverages the cddlib software [49] to handle degeneracy in linear constraints.

An example specification for APPSPACK with nonlinear constraints is:

```
method,
      asynch_pattern_search
         initial_delta = .5
         contraction_factor = 0.25
         threshold_delta = 1.e-4
         merit_function merit_max
         smoothing_factor = 10.0
```

See the DAKOTA Reference Manual [30] for additional detail on the APPSPACK commands and sample input specifications in `Dakota/test/dakota_apps.in`.

7.2.2 COLINY Library

The COLINY library [71] supersedes the SGOPT library and contains a variety of nongradient-based optimization algorithms. The suite of COLINY optimizers available in DAKOTA currently include the following:

- **Global Optimization Methods**
  - Several evolutionary algorithms, including genetic algorithms (`coliny_ea`)
  - DIRECT [105] (`coliny_direct`)
- **Local Optimization Methods**
  - Solis-Wets (`coliny_solis_wets`)
  - Pattern Search (`coliny_pattern_search`)
- **Interfaces to Third-Party Local Optimization Methods**
  - COBYLA2 (`coliny_cobyla`)

For expensive optimization problems, COLINY's global optimizers are best suited for identifying promising regions in the global design space. In multimodal design spaces, the combination of global identification (from COLINY) with efficient local convergence (from CONMIN, DOT, NLPQL, NPSOL, or OPT++) can be highly

effective. None of the COLINY methods are gradient-based, which makes them appropriate for problems for which gradient information is unavailable or is of questionable accuracy due to numerical noise. The COLINY methods support bound constraints and nonlinear constraints, but not linear constraints. The nonlinear constraints in COLINY are currently satisfied using penalty function formulations [106]. Support for methods which manage constraints internally is currently being developed and will be incorporated into future versions of DAKOTA. *Note that one observed drawback to* `coliny_solis_wets` *is that it does a poor job solving problems with nonlinear constraints*. Refer to Table 17.1 for additional method classification information.

An example specification for a simplex-based pattern search algorithm from COLINY is:

```
method,
      coliny_pattern_search
        max_function_evaluations = 2000
        solution_accuracy = 1.0e-4
        initial_delta = 0.05
        threshold_delta = 1.0e-8
        pattern_basis simplex
        exploratory_moves adaptive_pattern
        contraction_factor = 0.75
```

The DAKOTA Reference Manual [30] contains additional information on the COLINY options and settings.

7.2.3 Constrained Minimization (CONMIN) Library

The CONMIN library [130] contains two methods for gradient-based nonlinear optimization. For constrained optimization, the Method of Feasible Directions (DAKOTA's `conmin_mfd` method selection) is available, while for unconstrained optimization, the Fletcher-Reeves conjugate gradient method (DAKOTA's `conmin_frcg` method selection) is available. Both of these methods are most efficient at finding a local minimum in the vicinity of the starting point. The methods in CONMIN can be applied to global optimization problems, but there is no guarantee that they will find the globally optimal design point.

One observed drawback to CONMIN's Method of Feasible Directions is that it does a poor job handling equality constraints. This is the case even if the equality constraint is formulated as two inequality constraints. This problem is what motivates the modifications to MFD that are present in DOT's MMFD algorithm. For problems with equality constraints, it is better to use the OPT++ nonlinear interior point methods, NPSOL, NLPQL, or one of DOT's constrained optimization methods (see below).

An example specification for CONMIN's Method of Feasible Directions algorithm is:

```
method,
      conmin_mfd
        convergence_tolerance = 1.0e-4
        max_iterations = 100
        output quiet
```

Refer to the DAKOTA Reference Manual [30] for more information on the settings that can be used with CONMIN methods.

7.2.4 Design Optimization Tools (DOT) Library

The DOT library [132] contains nonlinear programming optimizers, specifically the Broyden-Fletcher-Goldfarb-Shanno (DAKOTA's `dot_bfgs` method selection) and Fletcher-Reeves conjugate gradient (DAKOTA's `dot_frcg` method selection) methods for unconstrained optimization, and the modified method of feasible directions (DAKOTA's `dot_mmfd` method selection), sequential linear programming (DAKOTA's `dot_slp` method selection), and sequential quadratic programming (DAKOTA's `dot_sqp` method selection) methods for constrained optimization.

All DOT methods are local gradient-based optimizers which are best suited for efficient navigation to a local minimum in the vicinity of the initial point. Global optima in nonconvex design spaces may be missed. Other gradient based optimizers for constrained optimization include the NPSOL, NLPQL, CONMIN, and OPT++ libraries.

Through the `optimization_type` specification, DOT can be used to solve either minimization or maximization problems. For all other optimizer libraries, it is up to the user to reformulate a maximization problem as a minimization problem by negating the objective function (i.e., maximize $f(x)$ is equivalent to minimize $-f(x)$). An example specification for DOT's BFGS quasi-Newton algorithm is:

```
method,
      dot_bfgs
        optimization_type maximize
        convergence_tolerance = 1.0e-4
        max_iterations = 100
        output quiet
```

See the DAKOTA Reference Manual [30] for additional detail on the DOT commands. More information on DOT can be obtained by contacting Vanderplaats Research and Development at `http://www.vrand.com`.

7.2.5 dl_solver — Solvers via Shared Libraries

On computer systems that permit use of shared libraries (most modern systems), DAKOTA can avail itself of optimization solvers contained in shared libraries. This is a first step toward allowing optional parts of DAKOTA, such as proprietary solvers, to be accessed from shared libraries. For example, the DAKOTA source distributions illustrate making a sample shared-library interface to SNOPT [57], whose use would be specified by

```
method,
      dl_solver = 'dl_snopt.dll'
```

The quoted string contains the name of the shared library, optionally followed by keyword assignments known to the library, such as

```
method,
      dl_solver = 'dl_snopt.dll outlev = 1'
```

which would turn on some diagnostic printing in the SNOPT example.

7.2.6 JEGA

The JEGA (John Eddy's Genetic Algorithms) library contains two global optimization methods. The first is a Multi-objective Genetic Algorithm (MOGA) which performs Pareto optimization. The second is a Single-objective Genetic Algorithm (SOGA) which performs optimization on a single objective function. These functions are accessed as (`moga` and `soga`) within DAKOTA.

The `moga` algorithm directly creates a population of Pareto optimal solutions. Over time, the selection operators of a genetic algorithm act to efficiently select non-dominated solutions along the Pareto front. Because a GA involves a population of solutions, many points along the Pareto front can be computed in a single study. Thus, although GAs are computationally expensive when compared to gradient-based methods, the advantage in the multiobjective setting is that one can obtain an entire Pareto set at the end of one genetic algorithm run, as compared with having to run the "weighted sum" single objective problem multiple times with different weights.

The DAKOTA Reference Manual [30] contains additional information on the JEGA options and settings. Section 7.3 discusses additional multiobjective optimization capabilities, and there are MOGA examples in Chapters 2 and 22.

7.2.7 NCSU DIRECT

We have an implementation of the global optimization method called DIRECT (DIviding RECTangles algorithm) that is detailed in [50]. DIRECT is a derivative free global optimization method that balances local search in promising regions of the design space with global search in unexplored regions. DIRECT adaptively subdivides the space of feasible design points to guarantee that iterates are generated in the neighborhood of a global minimum in finitely many iterations. In practice, DIRECT has proven an effective heuristic for many applications.

NCSU DIRECT is specified with `ncsu_direct`. One of the controls is `volume_boxsize_limit`, which terminates the optimization when the volume of the particular rectangle which contains the minimum function value found thus far is less than a certain percentage (given by the volume boxsize limit) of the whole volume of the hyperrectangle defined by the variable bounds. An example specification is given below:

```
method,
      ncsu_direct
      volume_boxsize_limit = 1.e-8
```

The DAKOTA Reference Manual [30] contains additional information on the NCSU DIRECT options and settings.

7.2.8 NLPQL Library

The NLPQL library contains a sequential quadratic programming (SQP) implementation (DAKOTA's `nlpql_sqp` method selection). The particular implementation used is NLPQLP [117], a variant with distributed and non-monotone line search. SQP is a nonlinear programming approach for constrained minimization which solves a series of quadratic programming (QP) subproblems, where each QP minimizes a quadratic approximation to the Lagrangian subject to linearized constraints. It uses an augmented Lagrangian merit function and a BFGS approximation to the Hessian of the Lagrangian. It is an infeasible method in that constraints will be satisfied at the final solution, but not necessarily during the solution process. The non-monotone line search used in NLPQLP is designed to be more robust in the presence of inaccurate or noisy gradients common in many engineering applications.

NLPQL's gradient-based approach is best suited for efficient navigation to a local minimum in the vicinity of the initial point. Global optima in nonconvex design spaces may be missed. Other gradient based optimizers for constrained optimization include the DOT, CONMIN, NPSOL, and OPT++ libraries.

See the DAKOTA Reference Manual [30] for additional detail on the NLPQL commands. More information on NLPQL can be obtained from Prof. Klaus Schittkowski at `http://www.uni-bayreuth.de/departments/math/~kschittkowski/nlpqlp20.htm`.

7.2.9 NPSOL Library

The NPSOL library [55] contains a sequential quadratic programming (SQP) implementation (DAKOTA's `npsol_sqp` method selection). Like NLPQL, it solves a series of QP subproblems, uses an augmented Lagrangian merit function and a BFGS approximation to the Hessian of the Lagrangian, and will not necessarily satisfy the constraints until the final solution. It uses a sufficient-decrease line search approach, which is a gradient-based line search for analytic, mixed, or DAKOTA-supplied numerical gradients and is a value-based line search in the vendor numerical case.

NPSOL's gradient-based approach is best suited for efficient navigation to a local minimum in the vicinity of the initial point. Global optima in nonconvex design spaces may be missed. Other gradient based optimizers for constrained optimization include the DOT, CONMIN, NLPQL, and OPT++ libraries.. For least squares methods based on NPSOL, refer to Section 8.2.2.

An example of an NPSOL specification is:

```
method,
      npsol_sqp
        convergence_tolerance = 1.0e-6
        max_iterations = 100
        output quiet
```

See the DAKOTA Reference Manual [30] for additional detail on the NPSOL commands. More information on NPSOL can be obtained by contacting Stanford Business Software at `http://www.sbsi-sol-optimize.com`.

The NPSOL library generates diagnostics in addition to those appearing in the DAKOTA output stream. These diagnostics are written to the default FORTRAN device 9 file (e.g., `ftn09` or `fort.9`, depending on the architecture) in the working directory.

7.2.10 OPT++ Library

The OPT++ library [92] contains primarily nonlinear programming optimizers for unconstrained, bound constrained, and nonlinearly constrained minimization: Polak-Ribiere conjugate gradient (DAKOTA's `optpp_cg` method selection), quasi-Newton (DAKOTA's `optpp_q_newton` method selection), finite difference Newton (DAKOTA's `optpp_fd_newton` method selection), and full Newton (DAKOTA's `optpp_newton` method selection). The library also contains the parallel direct search nongradient-based method [22] (specified as DAKOTA's `optpp_pds` method selection).

OPT++'s gradient-based optimizers are best suited for efficient navigation to a local minimum in the vicinity of the initial point. Global optima in nonconvex design spaces may be missed. OPT++'s PDS method does not use gradients and has some limited global identification abilities; it is best suited for problems for which gradient information is unavailable or is of questionable accuracy due to numerical noise. Some OPT++ methods are strictly unconstrained (`optpp_cg`) and some support bound constraints (`optpp_pds`), whereas the Newton-based methods (`optpp_q_newton`, `optpp_fd_newton`, and `optpp_newton`) all support general linear and nonlinear constraints (refer to Table 19.1). Other gradient-based optimizers include the DOT, CONMIN, NLPQL, and NPSOL libraries. For least squares methods based on OPT++, refer to Section 8.2.1.

An example specification for the OPT++ quasi-Newton algorithm is:

```
method,
      optpp_q_newton
```

```
max_iterations = 50
convergence_tolerance = 1e-4
output debug
```

See the DAKOTA Reference Manual [30] for additional detail on the OPT++ commands.

The OPT++ library generates diagnostics in addition to those appearing in the DAKOTA output stream. These diagnostics are written to the file `OPT_DEFAULT.out` in the working directory.

7.2.11 Parallel Integer Combinatorial Optimization (PICO)

DAKOTA employs the branch and bound capabilities of the PICO library for solving discrete and mixed continuous/discrete constrained nonlinear optimization problems. This capability is implemented in DAKOTA as a strategy and is discussed further in Section 10.5.

7.2.12 SGOPT

The SGOPT library has been deprecated, and all methods have been migrated to the COLINY library.

7.3 Additional Optimization Capabilities

DAKOTA provides several capabilities which extend the services provided by the optimization software packages described in Section 7.2. First, any of the optimization algorithms can be used for multiobjective optimization problems through the use of multiobjective transformation techniques (e.g., weighted sums). Finally, with any optimizer (or least squares solver described in Section 8.2), user-specified (and in some cases automatic or logarithmic) scaling may be applied to continuous design variables, objective functions (or least squares terms), and constraints.

7.3.1 Multiobjective Optimization

Multiobjective optimization means that there are two or more objective functions that you wish to optimize simultaneously. Often these are conflicting objectives, such as cost and performance. The answer to a multi-objective problem is usually not a single point. Rather, it is a set of points called the Pareto front. Each point on the Pareto front satisfies the Pareto optimality criterion, which is stated as follows: a feasible vector X^* is Pareto optimal if there exists no other feasible vector X which would improve some objective without causing a simultaneous worsening in at least one other objective. Thus, if a feasible point X' exists that CAN be improved on one or more objectives simultaneously, it is not Pareto optimal: it is said to be "dominated" and the points along the Pareto front are said to be "non-dominated."

There are three capabilities for multiobjective optimization in DAKOTA. First, there is the MOGA capability described previously in Section 7.2.6. This is a specialized algorithm capability. The second capability involves the use of response data transformations to recast a multiobjective problem as a single-objective problem. Currently, DAKOTA supports the simple weighted sum approach for this transformation, in which a composite objective function is constructed from a set of individual objective functions using a user-specified set of weighting factors. This approach is optimization algorithm independent, in that it works with any of the optimization methods listed previously in this chapter. The third capability is the Pareto-set optimization strategy described in Section 10.4.

This capability also utilizes the multiobjective response data transformations to allow optimization algorithm independence; however, it builds upon the basic approach by computing sets of optima in order to generate a Pareto trade-off surface.

In the multiobjective transformation approach in which multiple objectives are combined into one, an appropriate single-objective optimization technique is used to solve the problem. The advantage of this approach is that one can use any number of optimization methods that are especially suited for the particular problem class. One disadvantage of the weighted sum transformation approach is that a linear weighted sum objective cannot locate all optimal solutions in the Pareto set if the Pareto front is nonconvex. Also, if one wants to understand the effects of changing weights, this method can become computationally expensive. Since each optimization of a single weighted objective will find only one point near or on the Pareto front, many optimizations need to be performed to get a good parametric understanding of the influence of the weights.

The selection of a multiobjective optimization problem is made through the specification of multiple objective functions in the responses keyword block (i.e., the `num_objective_functions` specification is greater than `1`). The weighting factors on these objective functions can be optionally specified using the `multi_objective_weights` keyword (the default is equal weightings). The composite objective function for this optimization problem, F, is formed using these weights as follows: $F = \sum_{k=1}^{R} w_k f_k$, where the f_k terms are the individual objective function values, the w_k terms are the weights, and R is the number of objective functions. The weighting factors stipulate the relative importance of the design concerns represented by the individual objective functions; the higher the weighting factor, the more dominant a particular objective function will be in the optimization process. Constraints are not affected by the weighting factor mapping; therefore, both constrained and unconstrained multiobjective optimization problems can be formulated and solved with DAKOTA, assuming selection of an appropriate constrained or unconstrained single-objective optimization algorithm. Future multiobjective response data transformations for goal programming, normal boundary intersection, etc. are planned.

Figure 7.1 shows a DAKOTA input file for a multiobjective optimization problem based on the "textbook" test problem. This input file is named `dakota_multiobj1.in` in the `Dakota/test` directory. In the standard textbook formulation, there is one objective function and two constraints. In the multiobjective textbook formulation, all three of these functions are treated as objective functions (`num_objective_functions = 3`), with weights given by the `multi_objective_weights` keyword. Note that it is not required that the weights sum to a value of one. The multiobjective optimization capability also allows any number of constraints, although none are included in this example.

Figure 7.2 shows an excerpt of the results for this multiobjective optimization problem, with output in verbose mode. The data for function evaluation 9 show that the simulator is returning the values and gradients of the three objective functions and that this data is being combined by DAKOTA into the value and gradient of the composite objective function, as identified by the header "`Multiobjective transformation:`". This combination of value and gradient data from the individual objective functions employs the user-specified weightings of `.7`, `.2`, and `.1`. Convergence to the optimum of the multiobjective problem is indicated in this case by the gradient of the composite objective function going to zero (no constraints are active).

By performing multiple optimizations for different sets of weights, a family of optimal solutions can be generated which define the trade-offs that result when managing competing design concerns. This set of solutions is referred to as the Pareto set. Section 10.4 describes a solution strategy used for directly generating the Pareto set in order to investigate the trade-offs in multiobjective optimization problems.

7.3.2 Optimization with User-specified or Automatic Scaling

Some optimization problems involving design variables, objective functions, or constraints on vastly different scales may be solved more efficiently if these quantities are adjusted to a common scale (typically on the order

```
strategy,
        single_method
        tabular_graphics_data

method,
        npsol_sqp
          output verbose
          convergence_tolerance = 1.e-8

variables,
        continuous_design = 2
          initial_point    0.9    1.1
          upper_bounds     5.8    2.9
          lower_bounds     0.5   -2.9
          descriptors      'x1'   'x2'

interface,
        system asynchronous
          analysis_driver=  'text_book'

responses,
        num_objective_functions = 3
        multi_objective_weights = .7 .2 .1
        analytic_gradients
        no_hessians
```

Figure 7.1: Example DAKOTA input file for multiobjective optimization.

```
------------------------------
Begin Function Evaluation    9
------------------------------
Parameters for function evaluation 9:
                      5.9388064483e-01 x1
                      7.4158741198e-01 x2

(text_book /tmp/fileFNNH3v /tmp/fileRktLe9)
Removing /tmp/fileFNNH3v and /tmp/fileRktLe9

Active response data for function evaluation 9:
Active set vector = { 3 3 3 } Deriv vars vector = { 1 2 }
                      3.1662048106e-02 obj_fn_1
                     -1.8099485683e-02 obj_fn_2
                      2.5301156719e-01 obj_fn_3
 [ -2.6792982175e-01 -6.9024137415e-02 ] obj_fn_1 gradient
 [  1.1877612897e+00 -5.0000000000e-01 ] obj_fn_2 gradient
 [ -5.0000000000e-01  1.4831748240e+00 ] obj_fn_3 gradient



-----------------------------------
Post-processing Function Evaluation
-----------------------------------
Multiobjective transformation:
                      4.3844693257e-02 obj_fn
 [  1.3827084219e-06  5.8620632776e-07  ] obj_fn gradient

    7    1 1.0E+00    9  4.38446933E-02 1.5E-06    2 T TT

 Exit NPSOL - Optimal solution found.

 Final nonlinear objective value =   0.4384469E-01
```

Figure 7.2: DAKOTA results for the multiobjective optimization example.

of unity). With any optimizer (or least squares solver described in Section 8.2), user-specified characteristic value scaling may be applied to any of continuous design variables, functions/residuals, nonlinear inequality and equality constraints, and linear inequality and equality constraints. Automatic scaling is available for variables or responses with one- or two-sided bounds or equalities and may be combined with user-specified scaling values. Logarithmic ($\log_{10}$) scaling is available and may also be combined with characteristic values. Log scaling is not available for linear constraints. Moreover, when continuous design variables are log scaled, linear constraints are not permitted in the problem formulation. Discrete variable scaling is not supported.

Scaling is enabled on a per-method basis for optimizers and least squares minimizers by including the `scaling` keyword in the relevant `method` specification in the DAKOTA input deck. When scaling is enabled, variables, functions, gradients, Hessians, etc., are transformed such that the optimizer iterates in scaled variable space, whereas evaluations of the computational model as specified in the interface are performed on the original problem scale. Therefore using scaling does not require rewriting the interface to the simulation code. When the `scaling` keyword is omitted, all `*_scale_types` and `*_scales` specifications described below are ignored in the corresponding method, variables, and responses sections. When the method `output_level` is set above normal, scaling initialization and diagnostic information will be printed.

Scaling for a particular variable or response type is enabled through the `*_scale_types` specification (see the Reference Manual method section and references contained therein for a complete keyword list). Valid options for this string specification include `'none'` (default), `'value'`, `'auto'`, or `'log'`, for no, characteristic value, automatic, or logarithmic scaling, respectively (although not all types are valid for scaling all entities). If a single string is specified with any of these keywords it will apply to each component of the relevant vector, e.g., `cdv_scale_types = 'value'` will enable characteristic value scaling for each continuous design variable.

The user may additionally specify no, one, or a vector of characteristic scale values through the `*_scales` specification. These characteristic values are ignored for scaling type `'none'`, required for `'value'`, and optional for `'auto'` and `'log'`. If a single value is specified with any of these keywords it will apply to each component of the relevant vector, e.g., `cdv_scales = 3.0` will apply a characteristic scaling value of 3.0 to each continuous design variable.

When scaling is enabled, the following procedures determine the transformations used to scale each component of a variables or response vector. In all cases, if scaling would result in division by a value smaller in magnitude than 1.0e-3, a warning is issued and no scaling performed for that component.

- None (`'none'`): no scaling performed (`*_scales` ignored) on this component.
- Characteristic value (`'value'`): the corresponding quantity is scaled (divided) by the required characteristic value provided in the `*_scales` specification, and bounds are adjusted as necessary. If the value is negative, the sense of inequalities are changed accordingly.
- Automatic (`'auto'`): First, any characteristic values from the optional `*_scales` specification are applied. Then, automatic scaling will be attempted according to the following scheme:
  - two-sided bounds scaled into the interval [0,1];
  - one-sided bounds or targets are scaled by a characteristic value to move the bound or target to 1, and the sense of inequalities are changed if necessary;
  - no bounds or targets: no automatic scaling possible for this component

  Automatic scaling is not available for objective functions nor least squares terms since they lack bound constraints. Further, when automatically scaled, linear constraints are scaled by characteristic values only, not affinely scaled into [0,1].

- Logarithmic (`'log'`): First, any characteristic values from the optional `*_scales` specification are applied. Then, $\log_{10}$ scaling is applied. Logarithmic scaling is not available for linear constraints. Further, when continuous design variables are log scaled, linear constraints are not allowed.

Scaling for linear constraints specified through `linear_inequality_scales` or `linear_equality_scales` is applied *after* any (user-specified or automatic) continuous variable scaling. For example, for scaling mapping unscaled continuous design variables x to scaled variables $\tilde{x}$:

$$\tilde{x}^j = \frac{x^j - x_O^j}{x_M^j},$$

where x_M^j is the final component multiplier and x_O^j the offset, we have the following matrix system for linear inequality constraints

$$\begin{gathered} a_L \leq A_i x \leq a_U \\ a_L \leq A_i \left(\mathrm{diag}(x_M)\tilde{x} + x_O\right) \leq a_U \\ a_L - A_i x_O \leq A_i \mathrm{diag}(x_M)\tilde{x} \leq a_U - A_i x_O \\ \tilde{a}_L \leq \tilde{A}_i \tilde{x} \leq \tilde{a}_U, \end{gathered}$$

and user-specified or automatically computed scaling multipliers are applied to this final transformed system, which accounts for any continuous design variable scaling. When automatic scaling is in use for linear constraints they are linearly scaled by characteristic values only, not affinely scaled into the interval $[0, 1]$.

Figure 7.3 demonstrates the use of several scaling keywords for the textbook optimization problem. The continuous design variable `x1` is scaled by a characteristic value of 4.0, whereas `x2` is scaled automatically into $[0, 1]$ based on its bounds. The objective function will be scaled by a factor of 50.0, then logarithmically, the first nonlinear constraint by a factor of 15.0, and the second nonlinear constraint is not scaled.

```
strategy,
        single_method
          graphics,tabular_graphics_data

method,
        conmin_frcg
          scaling
          output verbose

model,
        single

variables,
        continuous_design = 2
          initial_point   -1.2  1.0
          lower_bounds    -2.0      0.001
          upper_bounds     2.0      2.0
          descriptors      'x1'     "x2"
          cdv_scale_types = 'value' 'log'
          cdv_scales = 4.0 0.1

interface,
        direct
          analysis_driver = 'rosenbrock'

responses,
        num_objective_functions = 1
          objective_function_scale_types = 'value'
          objective_function_scales = 50.0
        analytic_gradients
        no_hessians
```

Figure 7.3: Sample usage of scaling keywords in DAKOTA input specification.

Chapter 8

Nonlinear Least Squares Capabilities

8.1 Overview

Nonlinear least squares methods are optimization algorithms which exploit the special structure of a sum of the squares objective function [56]. These problems commonly arise in parameter estimation, system identification, and test/analysis reconciliation. In order to exploit the problem structure, more granularity is needed in the response data than that required for a typical optimization problem. That is, rather than using the sum-of-squares objective function and its gradient, least squares iterators require each term used in the sum-of-squares formulation along with its gradient. This means that the m functions in the DAKOTA response data set consist of the individual least squares terms along with any nonlinear inequality and equality constraints. These individual terms are often called *residuals* in cases where they denote errors of observed quantities from desired quantities.

The enhanced granularity needed for nonlinear least-squares algorithms allows for simplified computation of an approximate Hessian matrix. In Gauss-Newton-based methods for example, the true Hessian matrix is approximated by neglecting terms in which the residual function values appear, under the assumption that the residuals tend towards zero at the solution. As a result, residual function value and gradient information (first-order information) is sufficient to define the value, gradient, and approximate Hessian of the sum-of-squares objective function (second-order information). See Section 1.4.2 for additional details on this approximation.

In practice, least squares solvers will tend to be significantly more efficient than general-purpose optimization algorithms when the Hessian approximation is a good one, e.g., when the residuals tend towards zero at the solution. Specifically, they can exhibit the quadratic convergence rates of full Newton methods, even though only first-order information is used. Gauss-Newton-based least squares solvers may experience difficulty when the residuals at the solution are significant.

In order to specify a least-squares problem, the responses section of the DAKOTA input should be configured using `num_least_squares_terms` (as opposed to `num_objective_functions` in the case of optimization). Any linear or nonlinear constraints are handled in an identical way to that of optimization (see Section 7.1; note that neither Gauss-Newton nor NLSSOL require any constraint augmentation and NL2SOL supports neither linear nor nonlinear constraints). Gradients of the least squares terms and nonlinear constraints are required and should be specified using either `numerical_gradients`, `analytic_gradients`, or `mixed_gradients`. Since second derivatives of the least squares terms are not needed by nature of the Hessian approximations, the `no_hessians` specification should be used. DAKOTA's scaling options, described in Section 7.3.2 can be used on least squares problems, using the `least_squares_term_scales` keyword to scale least squares residuals, if desired.

8.2 Solution Techniques

Nonlinear least squares problems can be solved using the Gauss-Newton algorithm, which leverages the full Newton method from OPT++, the NLSSOL algorithm, which is closely related to NPSOL, or the NL2SOL algorithm, which uses a secant-based algorithm. Details for each are provided below.

8.2.1 Gauss-Newton

DAKOTA's Gauss-Newton algorithm consists of combining an implementation of the Gauss-Newton Hessian approximation (see Section 1.4.2) with full Newton optimization algorithms from the OPT++ package [92] (see Section 7.2.10). This approach can be selected using the `optpp_g_newton` method specification. An example specification follows:

```
method,
      optpp_g_newton
        max_iterations = 50
        convergence_tolerance = 1e-4
        output debug
```

Refer to the DAKOTA Reference Manual [30] for more detail on the input commands for the Gauss-Newton algorithm.

The Gauss-Newton algorithm is gradient-based and is best suited for efficient navigation to a local least squares solution in the vicinity of the initial point. Global optima in multimodal design spaces may be missed. Gauss-Newton supports bound, linear, and nonlinear constraints. For the nonlinearly-constrained case, constraint Hessians (required for full-Newton nonlinear interior point optimization algorithms) are approximated using quasi-Newton secant updates. Thus, both the objective and constraint Hessians are approximated using first-order information.

8.2.2 NLSSOL

The NLSSOL algorithm is a commercial software product of Stanford University that is bundled with current versions of the NPSOL library (see Section 7.2.9). It uses an SQP-based approach to solve generally-constrained nonlinear least squares problems. It periodically employs the Gauss-Newton Hessian approximation to accelerate the search. Like the Gauss-Newton algorithm of Section 8.2.1, its derivative order is balanced in that it requires only first-order information for the least squares terms and nonlinear constraints. This approach can be selected using the `nlssol_sqp` method specification. An example specification follows:

```
method,
      nlssol_sqp
        convergence_tolerance = 1e-8
```

Refer to the DAKOTA Reference Manual [30] for more detail on the input commands for NLSSOL.

8.2.3 NL2SOL

The NL2SOL algorithm [20] is a secant-based least-squares algorithm that is q-superlinearly convergent. It does not rely solely on the Gauss-Newton Hessian approximation and is appropriate for "large residual" problems, i.e., least squares problems for which the residuals do not tend towards zero at the solution.

8.2.4 Additional Features and Future plans

In DAKOTA 4.2, the capability to calculate confidence intervals on the estimated parameters was added. The confidence intervals are determined for individual parameters: these are not joint confidence intervals. The intervals are 95% intervals around the estimated parameters, and are calculated as the optimal value of the estimated parameters $+/-$ a t-test statistic times the standard error (SE) of the estimated parameter vector. The SE is based on a linearization approximation involving the matrix of the derivatives of the model with respect to the derivatives of the estimated parameters. In the case where these gradients are extremely inaccurate or the model is very nonlinear, the confidence intervals reported are likely to be inaccurate as well. Future work on generating confidence intervals on the estimated parameters for nonlinear least squares methods will involve adding Bonferroni confidence intervals and one or two methods for calculating joint confidence intervals (such as a linear approximation and the F-test method). See [119] and [133] for more details about confidence intervals. Note that confidence intervals are not calculated when scaling is used, when the number of least squares terms is less than the number of parameters to be estimated, or when using numerical gradients.

Dakota 4.2 also allows a form of weighted least squares. The user can specify a set of weights that are used to weight each residual term using the keyword `least_squares_weights`. Note that these weights must be pre-determined by the user and entered in the DAKOTA input file: they are not calculated on-the-fly. The user can also specify scaling for the least squares terms. Scaling is applied before weighting; usually one or the other would be applied but not both. The Responses section in the DAKOTA Reference Manual [30] has more detail about weighting and scaling of the residual terms.

The least squares branch in DAKOTA is an area of continuing enhancements, particularly through the addition of new least squares algorithms. One potential future addition is the orthogonal distance regression (ODR) algorithms which estimate values for both independent and dependent parameters.

8.3 Examples

Both the Rosenbrock and textbook example problems can be formulated as nonlinear least squares problems. Refer to Chapter 22 for more information on these formulations. Figure 8.1 shows an excerpt from the textbook example which demonstrates use of the Gauss-Newton approximation in computing the objective function value, gradient, and Hessian from values and gradients of the least squares terms.

Figure 8.2 shows an excerpt from the output obtained when running NL2SOL on a five-dimensional problem. This input file is named `dakota_nl2test.in` found in the `Dakota/test` directory. Note that the optimal parameter estimates are printed, followed by the residual norm and values of the individual residual terms, followed by the confidence intervals on the parameters.

The analysis driver script (the script being driven by DAKOTA) has to perform several tasks in the case of parameter estimation using nonlinear least squares methods. The analysis driver script must (1) read in the values of the parameters supplied by DAKOTA (2) run the computer simulation with these parameter values (3) retrieve the results from the computer simulation (4) compute the L2 norm of the difference between the experimental or measured values of the results and the computed simulation values of the results and (5) write out the L2 norm of the residuals to an external file that gets passed back to DAKOTA. Note there will be one line per residual term, specified with `num_least_squares_terms` in the DAKOTA input file. It is the last two steps which are different from most other DAKOTA applications.

To aid the calculation of the residuals, in DAKOTA 4.2 we have added the capability to have the user specify a data file with the experimental results. In this case, DAKOTA will calculate the residuals (that is, the simulation model results minus the experimental results) and their L2 norm and the user doesn't have to add this step

```
    Active response data for function evaluation 1:
    Active set vector = { 3 3 }
                          6.0000000000e-01 least_sq_term1
                          2.0000000000e-01 least_sq_term2
     [ -1.6000000000e+01  1.0000000000e+01  ] least_sq_term1 gradient
     [ -1.0000000000e+00  0.0000000000e+00  ] least_sq_term2 gradient


        nlf2_evaluator_gn results: objective fn. =
        4.0000000000e-01
        nlf2_evaluator_gn results: objective fn. gradient =
     [ -1.9600000000e+01  1.2000000000e+01 ]
        nlf2_evaluator_gn results: objective fn. Hessian =
    [[  5.1400000000e+02 -3.2000000000e+02
       -3.2000000000e+02  2.0000000000e+02 ]]
```

Figure 8.1: Example of the Gauss-Newton approximation.

```
<<<<< Iterator nl2sol completed.
<<<<< Function evaluation summary: 27 total (26 new, 1 duplicate)
<<<<< Best parameters          =
                      3.7541004764e-01 x1
                      1.9358463401e+00 x2
                     -1.4646865611e+00 x3
                      1.2867533504e-02 x4
                      2.2122702030e-02 x5
<<<<< Best residual norm =  7.3924926090e-03; 0.5 * norm^2 =  2.7324473487e-05
<<<<< Best residual terms      =
                     -2.5698266189e-03
                      4.4759880011e-03
                      9.9223430643e-04
                     -1.0634409194e-03

...

Confidence Interval for x1 is [  3.7116510206e-01,  3.7965499323e-01 ]
Confidence Interval for x2 is [  1.4845485507e+00,  2.3871441295e+00 ]
Confidence Interval for x3 is [ -1.9189348458e+00, -1.0104382765e+00 ]
Confidence Interval for x4 is [  1.1948590669e-02,  1.3786476338e-02 ]
Confidence Interval for x5 is [  2.0289951664e-02,  2.3955452397e-02 ]
```

Figure 8.2: Example of confidence intervals on optimal parameters

to their script: they can just return the simulation results. An example of this can be found in the file named `dakota_nls_datafile.in` in the `Dakota/examples/methods` directory. In this example, there are 3 residual terms. The data file of experimental results associated with this example is `least_squares_test.dat`. These three values are subtracted from the least squares terms to produce residuals for the nonlinear least squares problem.

Chapter 9

Surrogate-Based Minimization

9.1 Overview

Surrogate models approximate an original, high fidelity "truth" model, typically at reduced computational cost. In DAKOTA, several surrogate model selections are possible, which are categorized as data fits, multifidelity models, and reduced-order models, as described in Section 11.4. In the context of minimization (optimization or calibration), surrogate models can speed convergence by reducing function evaluation cost or smoothing noisy response functions.

9.2 Surrogate-Based Local Minimization

In the surrogate-based local minimization method (keyword: `surrogate_based_local`) the minimization algorithm operates on a surrogate model instead of directly operating on the computationally expensive simulation model. The surrogate model can be based on data fits, multifidelity models, or reduced-order models, as described in Section 11.4. Since the surrogate will generally have a limited range of accuracy, the surrogate-based local algorithm periodically checks the accuracy of the surrogate model against the original simulation model and adaptively manages the extent of the approximate optimization cycles using a trust region approach.

A generally-constrained nonlinear programming problem takes the form

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{g}_l \leq \mathbf{g}(\mathbf{x}) \leq \mathbf{g}_u \\ & \mathbf{h}(\mathbf{x}) = \mathbf{h}_t \\ & \mathbf{x}_l \leq \mathbf{x} \leq \mathbf{x}_u \end{aligned} \tag{9.1}$$

where $\mathbf{x} \in \Re^n$ is the vector of design variables, and f, $\mathbf{g}$, and $\mathbf{h}$ are the objective function, nonlinear inequality constraints, and nonlinear equality constraints, respectively[1]. Individual nonlinear inequality and equality constraints are enumerated using i and j, respectively (e.g., g_i and h_j). The corresponding surrogate-based optimization (SBO) algorithm may be formulated in several ways and applied to either optimization or least-squares calibration problems. In all cases, SBO solves a sequence of k approximate optimization subproblems subject to a trust region constraint Δ^k; however, many different forms of the surrogate objectives and constraints in the

[1] Any linear constraints are not approximated and may be added without modification to all formulations

Table 9.1: SBO approximate subproblem formulations.

	Original Objective	Lagrangian	Augmented Lagrangian
No constraints			TRAL
Linearized constraints		SQP-like	
Original constraints	Direct surrogate		IPTRSAO

approximate subproblem can be explored. In particular, the subproblem objective may be a surrogate of the original objective or a surrogate of a merit function (most commonly, the Lagrangian or augmented Lagrangian), and the subproblem constraints may be surrogates of the original constraints, linearized approximations of the surrogate constraints, or may be omitted entirely. Each of these combinations is shown in Table 9.1, where black indicates an inappropriate combination, gray indicates an acceptable combination, and blue indicates a common combination.

Initial approaches to nonlinearly-constrained SBO optimized an approximate merit function which incorporated the nonlinear constraints [112, 2]:

$$\begin{array}{ll} \text{minimize} & \hat{\Phi}^k(\mathbf{x}) \\ \text{subject to} & \| \mathbf{x} - \mathbf{x}_c^k \|_\infty \leq \Delta^k \end{array} \tag{9.2}$$

where the surrogate merit function is denoted as $\hat{\Phi}(\mathbf{x})$, $\mathbf{x}_c$ is the center point of the trust region, and the trust region is truncated at the global variable bounds as needed. The merit function to approximate was typically chosen to be a standard implementation [131, 98, 56] of the augmented Lagrangian merit function (see Eqs. 9.11–9.12), where the surrogate augmented Lagrangian is constructed from individual surrogate models of the objective and constraints (approximate and assemble, rather than assemble and approximate). In Table 9.1, this corresponds to row 1, column 3, and is known as the trust-region augmented Lagrangian (TRAL) approach. While this approach was provably convergent, convergence rates to constrained minima have been observed to be slowed by the required updating of Lagrange multipliers and penalty parameters [104]. Prior to converging these parameters, SBO iterates did not strictly respect constraint boundaries and were often infeasible. A subsequent approach (IPTRSAO [104]) that sought to directly address this shortcoming added explicit surrogate constraints (row 3, column 3 in Table 9.1):

$$\begin{array}{ll} \text{minimize} & \hat{\Phi}^k(\mathbf{x}) \\ \text{subject to} & \mathbf{g}_l \leq \hat{\mathbf{g}}^k(\mathbf{x}) \leq \mathbf{g}_u \\ & \hat{\mathbf{h}}^k(\mathbf{x}) = \mathbf{h}_t \\ & \| \mathbf{x} - \mathbf{x}_c^k \|_\infty \leq \Delta^k \ . \end{array} \tag{9.3}$$

While this approach does address infeasible iterates, it still shares the feature that the surrogate merit function may reflect inaccurate relative weightings of the objective and constraints prior to convergence of the Lagrange multipliers and penalty parameters. That is, one may benefit from more feasible intermediate iterates, but the process may still be slow to converge to optimality. The concept of this approach is similar to that of SQP-like SBO approaches [2] which use linearized constraints:

$$\begin{array}{ll} \text{minimize} & \hat{\Phi}^k(\mathbf{x}) \\ \text{subject to} & \mathbf{g}_l \leq \hat{\mathbf{g}}^k(\mathbf{x}_c^k) + \nabla\hat{\mathbf{g}}^k(\mathbf{x}_c^k)^T(\mathbf{x} - \mathbf{x}_c^k) \leq \mathbf{g}_u \\ & \hat{\mathbf{h}}^k(\mathbf{x}_c^k) + \nabla\hat{\mathbf{h}}^k(\mathbf{x}_c^k)^T(\mathbf{x} - \mathbf{x}_c^k) = \mathbf{h}_t \\ & \| \mathbf{x} - \mathbf{x}_c^k \|_\infty \leq \Delta^k \ . \end{array} \tag{9.4}$$

in that the primary concern is minimizing a composite merit function of the objective and constraints, but under the restriction that the original problem constraints may not be wildly violated prior to convergence of Lagrange

multiplier estimates. Here, the merit function selection of the Lagrangian function (row 2, column 2 in Table 9.1; see also Eq. 9.10) is most closely related to SQP, which includes the use of first-order Lagrange multiplier updates (Eq. 9.16) that should converge more rapidly near a constrained minimizer than the zeroth-order updates (Eqs. 9.13-9.14) used for the augmented Lagrangian.

All of these previous constrained SBO approaches involve a recasting of the approximate subproblem objective and constraints as a function of the original objective and constraint surrogates. A more direct approach is to use a formulation of:

$$\begin{aligned} \text{minimize} \quad & \hat{f}^k(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{g}_l \leq \hat{\mathbf{g}}^k(\mathbf{x}) \leq \mathbf{g}_u \\ & \hat{\mathbf{h}}^k(\mathbf{x}) = \mathbf{h}_t \\ & \| \mathbf{x} - \mathbf{x}_c^k \|_\infty \leq \Delta^k \end{aligned} \tag{9.5}$$

This approach has been termed the direct surrogate approach since it optimizes surrogates of the original objective and constraints (row 3, column 1 in Table 9.1) without any recasting. It is attractive both from its simplicity and potential for improved performance, and is the default approach taken in DAKOTA. Other DAKOTA defaults include the use of a filter method for iterate acceptance (see Section 9.2.1), an augmented Lagrangian merit function (see Section 9.2.2), Lagrangian hard convergence assessment (see Section 9.2.3), and no constraint relaxation (see Section 9.2.4).

While the formulation of Eq. 9.2 (and others from row 1 in Table 9.1) can suffer from infeasible intermediate iterates and slow convergence to constrained minima, each of the approximate subproblem formulations with explicit constraints (Eqs. 9.3-9.5, and others from rows 2-3 in Table 9.1) can suffer from the lack of a feasible solution within the current trust region. Techniques for dealing with this latter challenge involve some form of constraint relaxation. Homotopy approaches [104, 103] or composite step approaches such as Byrd-Omojokun [100], Celis-Dennis-Tapia [14], or MAESTRO [2] may be used for this purpose (see Section 9.2.4).

After each of the k iterations in the SBO method, the predicted step is validated by computing $f(\mathbf{x}_*^k)$, $\mathbf{g}(\mathbf{x}_*^k)$, and $\mathbf{h}(\mathbf{x}_*^k)$. One approach forms the trust region ratio ρ^k which measures the ratio of the actual improvement to the improvement predicted by optimization on the surrogate model. When optimizing on an approximate merit function (Eqs. 9.2–9.4), the following ratio is natural to compute

$$\rho^k = \frac{\Phi(\mathbf{x}_c^k) - \Phi(\mathbf{x}_*^k)}{\hat{\Phi}(\mathbf{x}_c^k) - \hat{\Phi}(\mathbf{x}_*^k)} \ . \tag{9.6}$$

The formulation in Eq. 9.5 may also form a merit function for computing the trust region ratio; however, the omission of this merit function from explicit use in the approximate optimization cycles can lead to synchronization problems with the optimizer.

Once computed, the value for ρ^k can be used to define the step acceptance and the next trust region size Δ^{k+1} using logic similar to that shown in Table 9.2. Typical factors for shrinking and expanding are 0.5 and 2.0, respectively, but these as well as the threshold ratio values are tunable parameters in the algorithm (see Surrogate-Based Method controls in the DAKOTA Reference Manual [30]). In addition, the use of discrete thresholds is not required, and continuous relationships using adaptive logic can also be explored [142, 143]. Iterate acceptance or rejection completes an SBO cycle, and the cycles are continued until either soft or hard convergence criteria (see Section 9.2.3) are satisfied.

9.2.1 Iterate acceptance logic

[2]Exception: retain if $\mathbf{x}_*^k$ in trust region interior for design of experiments-based surrogates (global data fits, S-ROM, global E-ROM)

Table 9.2: Sample trust region ratio logic.

Ratio Value	Surrogate Accuracy	Iterate Acceptance	Trust Region Sizing
$\rho^k \leq 0$	poor	reject step	shrink
$0 < \rho^k \leq 0.25$	marginal	accept step	shrink
$0.25 < \rho^k < 0.75$ or $\rho^k > 1.25$	moderate	accept step	retain
$0.75 \leq \rho^k \leq 1.25$	good	accept step	expand[2]

When a surrogate optimization is completed and the approximate solution has been validated, then the decision must be made to either accept or reject the step. The traditional approach is to base this decision on the value of the trust region ratio, as outlined previously in Table 9.2. An alternate approach is to utilize a filter method [45], which does not require penalty parameters or Lagrange multiplier estimates. The basic idea in a filter method is to apply the concept of Pareto optimality to the objective function and constraint violations and only accept an iterate if it is not dominated by any previous iterate. Mathematically, a new iterate is not dominated if at least one of the following:

$$\text{either} \quad f < f^{(i)} \quad \text{or} \quad c < c^{(i)} \tag{9.7}$$

is true for all i in the filter, where c is a selected norm of the constraint violation. This basic description can be augmented with mild requirements to prevent point accumulation and assure convergence, known as a slanting filter [45]. Figure 9.1 illustrates the filter concept, where objective values are plotted against constraint violation for accepted iterates (blue circles) to define the dominated region (denoted by the gray lines). A filter method relaxes the common enforcement of monotonicity in constraint violation reduction and, by allowing more flexibility in acceptable step generation, often allows the algorithm to be more efficient.

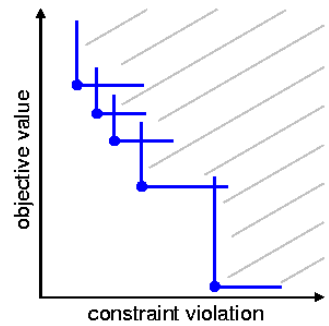


Figure 9.1: Depiction of filter method.

The use of a filter method is compatible with any of the SBO formulations in Eqs. 9.2–9.5.

9.2.2 Merit functions

The merit function $\Phi(\mathbf{x})$ used in Eqs. 9.2-9.4,9.6 may be selected to be a penalty function, an adaptive penalty function, a Lagrangian function, or an augmented Lagrangian function. In each of these cases, the more flexible inequality and equality constraint formulations with two-sided bounds and targets (Eqs. 9.1,9.3-9.5), have been converted to a standard form of $\mathbf{g}(\mathbf{x}) \leq 0$ and $\mathbf{h}(\mathbf{x}) = 0$ (in Eqs. 9.8,9.10-9.16). The active set of inequality constraints is denoted as $\mathbf{g}^+$.

The penalty function employed in this paper uses a quadratic penalty with the penalty schedule linked to SBO iteration number

$$\Phi(\mathbf{x}, r_p) = f(\mathbf{x}) + r_p \mathbf{g}^+(\mathbf{x})^T \mathbf{g}^+(\mathbf{x}) + r_p \mathbf{h}(\mathbf{x})^T \mathbf{h}(\mathbf{x}) \tag{9.8}$$

$$r_p = e^{(k+\text{offset})/10} \tag{9.9}$$

The adaptive penalty function is identical in form to Eq. 9.8, but adapts r_p using monotonic increases in the iteration offset value in order to accept any iterate that reduces the constraint violation.

The Lagrangian merit function is

$$\Phi(\mathbf{x}, \boldsymbol{\lambda}_g, \boldsymbol{\lambda}_h) = f(\mathbf{x}) + \boldsymbol{\lambda}_g^T \mathbf{g}^+(\mathbf{x}) + \boldsymbol{\lambda}_h^T \mathbf{h}(\mathbf{x}) \tag{9.10}$$

for which the Lagrange multiplier estimation is discussed in Section 9.2.3. Away from the optimum, it is possible for the least squares estimates of the Lagrange multipliers for active constraints to be zero, which equates to omitting the contribution of an active constraint from the merit function. This is undesirable for tracking SBO progress, so usage of the Lagrangian merit function is normally restricted to approximate subproblems and hard convergence assessments.

The augmented Lagrangian employed in this paper follows the sign conventions described in [131]

$$\Phi(\mathbf{x}, \boldsymbol{\lambda}_\psi, \boldsymbol{\lambda}_h, r_p) = f(\mathbf{x}) + \boldsymbol{\lambda}_\psi^T \boldsymbol{\psi}(\mathbf{x}) + r_p \boldsymbol{\psi}(\mathbf{x})^T \boldsymbol{\psi}(\mathbf{x}) + \boldsymbol{\lambda}_h^T \mathbf{h}(\mathbf{x}) + r_p \mathbf{h}(\mathbf{x})^T \mathbf{h}(\mathbf{x}) \tag{9.11}$$

$$\psi_i = \max\left\{ g_i, -\frac{\lambda_{\psi_i}}{2r_p} \right\} \tag{9.12}$$

where $\boldsymbol{\psi}(\mathbf{x})$ is derived from the elimination of slack variables for the inequality constraints. In this case, simple zeroth-order Lagrange multiplier updates may be used:

$$\boldsymbol{\lambda}_\psi^{k+1} = \boldsymbol{\lambda}_\psi^k + 2r_p \boldsymbol{\psi}(\mathbf{x}) \tag{9.13}$$

$$\boldsymbol{\lambda}_h^{k+1} = \boldsymbol{\lambda}_h^k + 2r_p \mathbf{h}(\mathbf{x}) \tag{9.14}$$

The updating of multipliers and penalties is carefully orchestrated [18] to drive reduction in constraint violation of the iterates. The penalty updates can be more conservative than in Eq. 9.9, often using an infrequent application of a constant multiplier rather than a fixed exponential progression.

9.2.3 Convergence assessment

To terminate the SBO process, hard and soft convergence metrics are monitored. It is preferable for SBO studies to satisfy hard convergence metrics, but this is not always practical (e.g., when gradients are unavailable or unreliable). Therefore, simple soft convergence criteria are also employed which monitor for diminishing returns (relative improvement in the merit function less than a tolerance for some number of consecutive iterations).

To assess hard convergence, one calculates the norm of the projected gradient of a merit function whenever the feasibility tolerance is satisfied. The best merit function for this purpose is the Lagrangian merit function from Eq. 9.10. This requires a least squares estimation for the Lagrange multipliers that best minimize the projected gradient:

$$\nabla_x \Phi(\mathbf{x}, \boldsymbol{\lambda}_g, \boldsymbol{\lambda}_h) = \nabla_x f(\mathbf{x}) + \boldsymbol{\lambda}_g^T \nabla_x \mathbf{g}^+(\mathbf{x}) + \boldsymbol{\lambda}_h^T \nabla_x \mathbf{h}(\mathbf{x}) \tag{9.15}$$

where gradient portions directed into active global variable bounds have been removed. This can be posed as a linear least squares problem for the multipliers:

$$\mathbf{A}\boldsymbol{\lambda} = -\nabla_x f \tag{9.16}$$

where $\mathbf{A}$ is the matrix of active constraint gradients, $\boldsymbol{\lambda}_g$ is constrained to be non-negative, and $\boldsymbol{\lambda}_h$ is unrestricted in sign. To estimate the multipliers using non-negative and bound-constrained linear least squares, the NNLS and BVLS routines [87] from NETLIB are used, respectively.

9.2.4 Constraint relaxation

The goal of constraint relaxation is to achieve efficiency through the balance of feasibility and optimality when the trust region restrictions prevent the location of feasible solutions to constrained approximate subproblems (Eqs. 9.3-9.5, and other formulations from rows 2-3 in Table 9.1). The SBO algorithm starting from infeasible

points will commonly generate iterates which seek to satisfy feasibility conditions without regard to objective reduction [103].

One approach for achieving this balance is to use *relaxed constraints* when iterates are infeasible with respect to the surrogate constraints. We follow Perez, Renaud, and Watson [104], and use a *global homotopy* mapping the relaxed constraints and the surrogate constraints. For formulations in Eqs. 9.3 and 9.5 (and others from row 3 in Table 9.1), the relaxed constraints are defined from

$$\tilde{\mathbf{g}}^k(\mathbf{x},\tau) = \hat{\mathbf{g}}^k(\mathbf{x}) + (1-\tau)\mathbf{b}_g \tag{9.17}$$

$$\tilde{\mathbf{h}}^k(\mathbf{x},\tau) = \hat{\mathbf{h}}^k(\mathbf{x}) + (1-\tau)\mathbf{b}_h \tag{9.18}$$

For Eq. 9.4 (and others from row 2 in Table 9.1), the original surrogate constraints $\hat{\mathbf{g}}^k(\mathbf{x})$ and $\hat{\mathbf{h}}^k(\mathbf{x})$ in Eqs. 9.17-9.18 are replaced with their linearized forms ($\hat{\mathbf{g}}^k(\mathbf{x}_c^k) + \nabla\hat{\mathbf{g}}^k(\mathbf{x}_c^k)^T(\mathbf{x}-\mathbf{x}_c^k)$ and $\hat{\mathbf{h}}^k(\mathbf{x}_c^k) + \nabla\hat{\mathbf{h}}^k(\mathbf{x}_c^k)^T(\mathbf{x}-\mathbf{x}_c^k)$, respectively). The approximate subproblem is then reposed using the relaxed constraints as

$$\begin{array}{ll} \text{minimize} & \hat{f}^k(\mathbf{x}) \ \text{ or } \ \hat{\Phi}^k(\mathbf{x}) \\ \text{subject to} & \mathbf{g}_l \leq \tilde{\mathbf{g}}^k(\mathbf{x},\tau^k) \leq \mathbf{g}_u \\ & \tilde{\mathbf{h}}^k(\mathbf{x},\tau^k) = \mathbf{h}_t \\ & \| \mathbf{x}-\mathbf{x}_c^k \|_\infty \leq \Delta^k \end{array} \tag{9.19}$$

in place of the corresponding subproblems in Eqs. 9.3-9.5. Alternatively, since the relaxation terms are constants for the kth iteration, it may be more convenient for the implementation to constrain $\hat{\mathbf{g}}^k(\mathbf{x})$ and $\hat{\mathbf{h}}^k(\mathbf{x})$ (or their linearized forms) subject to relaxed bounds and targets ($\tilde{\mathbf{g}}_l^k$, $\tilde{\mathbf{g}}_u^k$, $\tilde{\mathbf{h}}_t^k$). The parameter τ is the homotopy parameter controlling the extent of the relaxation: when $\tau = 0$, the constraints are fully relaxed, and when $\tau = 1$, the surrogate constraints are recovered. The vectors $\mathbf{b}_g, \mathbf{b}_h$ are chosen so that the starting point, $\mathbf{x}^0$, is feasible with respect to the fully relaxed constraints:

$$\mathbf{g}_l \leq \tilde{\mathbf{g}}^0(\mathbf{x}^0,0) \leq \mathbf{g}_u \tag{9.20}$$

$$\tilde{\mathbf{h}}^0(\mathbf{x}^0,0) = \mathbf{h}_t \tag{9.21}$$

At the start of the SBO algorithm, $\tau^0 = 0$ if $\mathbf{x}^0$ is infeasible with respect to the unrelaxed surrogate constraints; otherwise $\tau^0 = 1$ (i.e., no constraint relaxation is used). At the start of the kth SBO iteration where $\tau^{k-1} < 1$, τ^k is determined by solving the subproblem

$$\begin{array}{ll} \text{maximize} & \tau^k \\ \text{subject to} & \mathbf{g}_l \leq \tilde{\mathbf{g}}^k(\mathbf{x},\tau^k) \leq \mathbf{g}_u \\ & \tilde{\mathbf{h}}^k(\mathbf{x},\tau^k) = \mathbf{h}_t \\ & \| \mathbf{x}-\mathbf{x}_c^k \|_\infty \leq \Delta^k \\ & \tau^k \geq 0 \end{array} \tag{9.22}$$

starting at $(\mathbf{x}_*^{k-1}, \tau^{k-1})$, and then adjusted as follows:

$$\tau^k = \min\left\{1, \tau^{k-1} + \alpha\left(\tau_{\max}^k - \tau^{k-1}\right)\right\} \tag{9.23}$$

The adjustment parameter $0 < \alpha < 1$ is chosen so that that the feasible region with respect to the relaxed constraints has positive volume within the trust region. Determining the optimal value for α remains an open question and will be explored in future work.

After τ^k is determined using this procedure, the problem in Eq. 9.19 is solved for $\mathbf{x}_*^k$. If the step is accepted, then

the value of τ^k is updated using the current iterate $\mathbf{x}_*^k$ and the validated constraints $\mathbf{g}(\mathbf{x}_*^k)$ and $\mathbf{h}(\mathbf{x}_*^k)$:

$$\tau^k = \min\{1, \min_i \tau_i, \min_j \tau_j\} \tag{9.24}$$

$$\text{where } \tau_i = 1 + \frac{\min\{g_i(\mathbf{x}_*^k) - g_{l_i}, g_{u_i} - g_i(\mathbf{x}_*^k)\}}{b_{g_i}} \tag{9.25}$$

$$\tau_j = 1 - \frac{|h_j(\mathbf{x}_*^k) - h_{t_j}|}{b_{h_j}} \tag{9.26}$$

Figure 9.2 illustrates the SBO algorithm on a two-dimensional problem with one inequality constraint starting from an infeasible point, $\mathbf{x}^0$. The minimizer of the problem is denoted as $\mathbf{x}^*$. Iterates generated using the surrogate constraints are shown in red, where feasibility is achieved first, and then progress is made toward the optimal point. The iterates generated using the relaxed constraints are shown in blue, where a balance of satisfying feasibility and optimality has been achieved, leading to fewer overall SBO iterations.

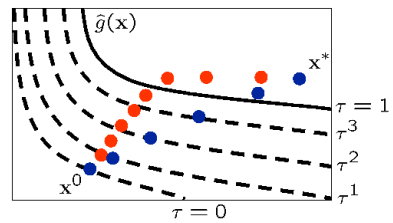


Figure 9.2: Illustration of SBO iterates using surrogate (red) and relaxed (blue) constraints.

The behavior illustrated in Fig. 9.2 is an example where using the relaxed constraints over the surrogate constraints may improve the overall performance of the SBO algorithm by reducing the number of iterations performed. This improvement comes at the cost of solving the minimization subproblem in Eq. 9.22, which can be significant in some cases (i.e., when the cost of evaluating $\hat{\mathbf{g}}^k(\mathbf{x})$ and $\hat{\mathbf{h}}^k(\mathbf{x})$ is not negligible, such as with multifidelity or ROM surrogates). As shown in the numerical experiments involving the Barnes problem presented at the end of this paper, the directions toward constraint violation reduction and objective function reduction may be in opposing directions. In such cases, the use of the relaxed constraints may result in an *increase* in the overall number of SBO iterations since feasibility must ultimately take precedence.

9.2.5 SBO with Data Fits

When performing SBO with local, multipoint, and global data fit surrogates, it is necessary to regenerate or update the data fit for each new trust region. In the global data fit case, this can mean performing a new design of experiments on the original high-fidelity model for each trust region, which can effectively limit the approach to use on problems with, at most, tens of variables. Figure 9.3 displays this case. However, an important benefit of the global sampling is that the global data fits can tame poorly-behaved, nonsmooth, discontinuous response variations within the original model into smooth, differentiable, easily navigated surrogates. This allows SBO with global data fits to extract the relevant global design trends from noisy simulation data.

When enforcing local consistency between a global data fit surrogate and a high-fidelity model at a point, care must be taken to balance this local consistency requirement with the global accuracy of the surrogate. In particular, performing a correction on an existing global data fit in order to enforce local consistency can skew the data fit and destroy its global accuracy. One approach for achieving this balance is to include the consistency requirement within the data fit process by constraining the global data fit calculation (e.g., using constrained linear least squares). This allows the data fit to satisfy the consistency requirement while still addressing global accuracy with its remaining degrees of freedom. Embedding the consistency within the data fit also reduces the sampling requirements. For example, a quadratic polynomial normally requires at least $(n+1)(n+2)/2$ samples for n variables to

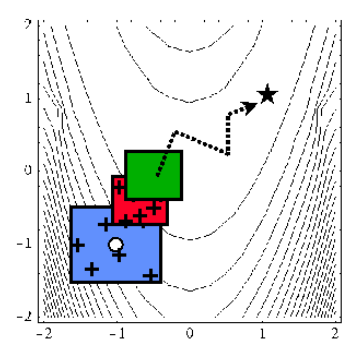


Figure 9.3: SBO iteration progression for global data fits.

perform the fit. However, with embedded first-order consistency constraints, the minimum number of samples is reduced by $n + 1$ to $(n^2 + n)/2$.

In the local and multipoint data fit cases, the iteration progression will appear as in Fig. 9.5. Both cases involve a single new evaluation of the original high-fidelity model per trust region, with the distinction that multipoint approximations reuse information from previous SBO iterates. Like model hierarchy surrogates, these techniques scale to larger numbers of design variables. Unlike model hierarchy surrogates, they generally do not require surrogate corrections, since the matching conditions are embedded in the surrogate form (as discussed for the global Taylor series approach above). The primary disadvantage to these surrogates is that the region of accuracy tends to be smaller than for global data fits and multifidelity surrogates, requiring more SBO cycles with smaller trust regions. More information on the design of experiments methods is available in Chapter 5, and the data fit surrogates are described in Section 11.4.1.

Figure 9.4 shows a DAKOTA input file that implements surrogate-based optimization on Rosenbrock's function. This input file is named `dakota_sbo_rosen.in` in the `Dakota/test` directory. The first method keyword block contains the SBO keyword `surrogate_based_local`, plus the commands for specifying the trust region size and scaling factors. The optimization portion of SBO, using the CONMIN Fletcher-Reeves conjugate gradient method, is specified in the following keyword blocks for `method`, `model`, `variables`, and `responses`. The model used by the optimization method specifies that a global surrogate will be used to map variables into responses (no `interface` specification is used by the surrogate model). The global surrogate is constructed using a DACE method which is identified with the `'SAMPLING'` identifier. This data sampling portion of SBO is specified in the final set of keyword blocks for `method`, `model`, `interface`, and `responses` (the earlier `variables` specification is reused). This example problem uses the Latin hypercube sampling method in the LHS software to select 10 design points in each trust region. A single surrogate model is constructed for the objective function using a quadratic polynomial. The initial trust region is centered at the design point $(x_1, x_2) = (-1.2, 1.0)$, and extends ± 0.4 (10% of the global bounds) from this point in the x_1 and x_2 coordinate directions.

If this input file is executed in DAKOTA, it will converge to the optimal design point at $(x_1, x_2) = (1, 1)$ in approximately 1000 function evaluations. While this solution is correct, it is obtained at a much higher cost than a traditional gradient-based optimizer (e.g., see the results obtained from `dakota_rosenbrock.in`). This demonstrates that the SBO method with global data fits is not really intended for use with smooth continuous optimization problems; direct gradient-based optimization can be more efficient for such applications. Rather, SBO with global data fits is best-suited for the types of problems that occur in engineering design where the response quantities may be discontinuous, nonsmooth, or may have multiple local optima [59]. In these types of engineering design problems, traditional gradient-based optimizers often are ineffective, whereas global data fits can extract the global trends of interest despite the presence of local nonsmoothness (for an example problem with multiple local optima, look in `Dakota/test` for the file `dakota_sbo_sine_fcn.in` [60]).

The surrogate-based local minimizer is only mathematically guaranteed to find a local minimum. However, in practice, SBO can often find the global minimum. Due to the random sampling method used within the SBO algorithm, the SBO method will solve a given problem a little differently each time it is run (unless the user specifies a particular random number seed in the dakota input file as is shown in Figure 9.4). Our experience on the quasi-sine function mentioned above is that if you run this problem 10 times with the same starting conditions but different seeds, then you will find the global minimum in about 70-80% of the trials. This is good performance for what is mathematically only a local optimization method.

9.2.6 SBO with Multifidelity Models

When performing SBO with model hierarchies, the low-fidelity model is normally fixed, requiring only a single high-fidelity evaluation to compute a new correction for each new trust region. Figure 9.5 displays this case. This

```
strategy,
        single_method
        graphics
        tabular_graphics_data
        method_pointer = 'SBLO'

method,
        id_method = 'SBLO'
        surrogate_based_local
        model_pointer = 'SURROGATE'
        approx_method_pointer = 'NLP'
        max_iterations = 500
        trust_region
          initial_size = 0.10
          minimum_size = 1.0e-6
          contract_threshold = 0.25
          expand_threshold   = 0.75
          contraction_factor = 0.50
          expansion_factor   = 1.50

method,
        id_method = 'NLP'
        conmin_frcg,
          max_iterations = 50,
          convergence_tolerance = 1e-8

model,
        id_model = 'SURROGATE'
        surrogate global
          responses_pointer = 'SURROGATE_RESP'
          dace_method_pointer = 'SAMPLING'
          correction additive zeroth_order
          polynomial quadratic

variables,
        continuous_design = 2
          initial_point   -1.2  1.0
          lower_bounds    -2.0     -2.0
          upper_bounds     2.0  2.0
          descriptors      'x1'     'x2'

responses,
        id_responses = 'SURROGATE_RESP'
        num_objective_functions = 1
        numerical_gradients
          method_source dakota
          interval_type central
          fd_gradient_step_size = 1.e-6
        no_hessians

method,
        id_method = 'SAMPLING'
        model_pointer = 'TRUTH'
        nond_sampling
          samples = 10
          seed = 531
          sample_type lhs
          all_variables

model,
        id_model = 'TRUTH'
        single
          interface_pointer = 'TRUE_FN'
          responses_pointer = 'TRUE_RESP'

interface,
        direct
        id_interface = 'TRUE_FN'
          analysis_driver = 'rosenbrock'
          deactivate evaluation_cache restart_file

responses,
        id_responses = 'TRUE_RESP'
        num_objective_functions = 1
        no_gradients
        no_hessians
```

Figure 9.4: DAKOTA input file for the surrogate-based local optimization example.

renders the multifidelity SBO technique more scalable to larger numbers of design variables since the number of high-fidelity evaluations per iteration (assuming no finite differencing for derivatives) is independent of the scale of the design problem. However, the ability to smooth poorly-behaved response variations in the high-fidelity model is lost, and the technique becomes dependent on having a well-behaved low-fidelity model[3]. In addition, the parameterizations for the low and high-fidelity models may differ, requiring the use of a mapping between these parameterizations. Space mapping, corrected space mapping, POD mapping, and hybrid POD space mapping are being explored for this purpose [110, 111].

When applying corrections to the low-fidelity model, there is no concern for balancing global accuracy with the local consistency requirements. However, with only a single high-fidelity model evaluation at the center of each trust region, it is critical to use the best correction possible on the low-fidelity model in order to achieve rapid convergence rates to the optimum of the high-fidelity model [35].

A multifidelity test problem named `dakota_sbo_hierarchical.in` is available in `Dakota/test` to demonstrate this SBO approach. This test problem uses the Rosenbrock function as the high fidelity model and a function named "lf_rosenbrock" as the low fidelity model. Here, lf_rosenbrock is a variant of the Rosenbrock function (see `Dakota/test/lf_rosenbrock.C` for formulation) with the minimum point at $(x_1, x_2) = (0.80, 0.44)$, whereas the minimum of the original Rosenbrock function is $(x_1, x_2) = (1, 1)$. Multifidelity SBO locates the high-fidelity minimum in 11 high fidelity evaluations for additive second-order corrections and in 208 high fidelity evaluations for additive first-order corrections, but fails for zeroth-order additive corrections by converging to the low-fidelity minimum.

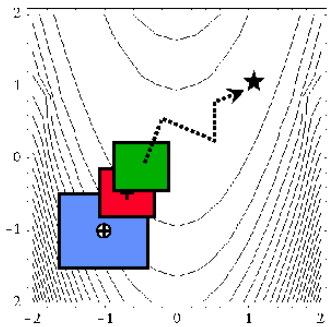


Figure 9.5: SBO iteration progression for model hierarchies.

9.2.7 SBO with Reduced Order Models

When performing SBO with reduced-order models (ROMs), the ROM is mathematically generated from the high-fidelity model. A critical issue in this ROM generation is the ability to capture the effect of parametric changes within the ROM. Two approaches to parametric ROM are extended ROM (E-ROM) and spanning ROM (S-ROM) techniques [138]. Closely related techniques include tensor singular value decomposition (SVD) methods [86]. In the single-point and multipoint E-ROM cases, the SBO iteration can appear as in Fig. 9.5, whereas in the S-ROM, global E-ROM, and tensor SVD cases, the SBO iteration will appear as in Fig. 9.3. In addition to the high-fidelity model analysis requirements, procedures for updating the system matrices and basis vectors are also required.

Relative to data fits and multifidelity models, ROMs have some attractive advantages. Compared to data fits such as regression-based polynomial models, they are more physics-based and would be expected to be more predictive (e.g., in extrapolating away from the immediate data). Compared to multifidelity models, ROMS may be more practical in that they do not require multiple computational models or meshes which are not always available. The primary disadvantage is potential invasiveness to the simulation code for projecting the system using the reduced basis.

[3]It is also possible to use a hybrid data fit/multifidelity approach in which a smooth data fit of a noisy low fidelity model is used in combination with a high fidelity model

9.3 Surrogate-Based Global Minimization

Surrogate-based global minimization differs from the surrogate-based local minimization approach discussed in the previous section in several ways: it is not a trust-region approach; initially there is one surrogate constructed over a set of sample points and the optimizer operates on that surrogate (as opposed to adaptively selecting points and re-building a surrogate in each trust region); and there is no guarantee of convergence.

The `surrogate_based_global` method was developed to address two needs. The first is the case where a user wishes to use existing function evaluations or a fixed sample size (perhaps based on computational cost and allocation of resources) to build a surrogate once and optimize on it. In this case (a single global optimization on a surrogate model), the set of surrogate building points is determined in advance as opposed to the trust-region local surrogate optimization in which the number of "true" function evaluations depends on the location and size of the trust region, the goodness of the surrogate within the trust-region, and problem characteristics.

In the second `surrogate_based_global` use case, we want to update the surrogate, but globally. That is, we add points to the sample set used to create the surrogate, rebuild the surrogate, and then perform another global optimization on the new surrogate. Thus, surrogate-based global optimization can be used in an iterative scheme. In one iteration, minimizers of the surrogate model are found, and a selected subset of these are passed to the next iteration. In the next iteration, these surrogate points are evaluated with the "truth" model, and then added to the set of points upon which the next surrogate is constructed. This presents a more accurate surrogate to the minimizer at each subsequent iteration, presumably driving to optimality quickly. Note that a global surrogate is constructed using the same bounds in each iteration. This approach has no guarantee of convergence.

The surrogate-based global method was originally designed for MOGA (a multi-objective genetic algorithm). Since genetic algorithms often need thousands or tens of thousands of points to produce optimal or near-optimal solutions, surrogates can help by reducing the necessary truth model evaluations. Instead of creating one set of surrogates for the individual objectives and running the optimization algorithm on the surrogate once, the idea is to select points along the (surrogate) Pareto frontier, which can be used to supplement the existing points. In this way, one does not need to use many points initially to get a very accurate surrogate. The surrogate becomes more accurate as the iterations progress.

Most single objective optimization methods will return only a single optimal point. In that case, only one point from the surrogate model will be evaluated with the "true" function and added to the pointset upon which the surrogate is based. In this case, it will take many iterations of the surrogate-based global optimization for the approach to converge, and its utility may not be as great as for the multi-objective case when multiple optimal solutions are passed from one iteration to the next to supplement the surrogate. Note that the user has the option of appending the optimal points from the surrogate model to the current set of truth points or using the optimal points from the surrogate model to replace the optimal set of points from the previous iteration. Although appending to the set is the default behavior, at this time we strongly recommend using the option `replace_points` because it appears to be more accurate and robust.

When using the surrogate-based global method, we first recommend running one optimization on a single surrogate model. That is, set `max_iterations` to 1. This will allow one to get a sense of where the optima are located and also what surrogate types are the most accurate to use for the problem. Note that by fixing the seed of the sample on which the surrogate is built, one can take a DAKOTA input file, change the surrogate type, and re-run the problem without any additional function evaluations by specifying the use of the dakota restart file which will pick up the existing function evaluations, create the new surrogate type, and run the optimization on that new surrogate. Also note that one can specify that surrogates be built for all primary functions and constraints or for only a subset of these functions and constraints. This allows one to use a "truth" model directly for some of the response functions, perhaps due to them being much less expensive than other functions. Finally, a diagnostic threshold can be used to stop the method if the surrogate is so poor that it is unlikely to provide useful points. If the goodness-of-fit has an R-squared value less than 0.5, meaning that less than half the variance of the output can

be explained or accounted for by the surrogate model, the surrogate-based global optimization stops and outputs an error message. This is an arbitrary threshold, but generally one would want to have an R-squared value as close to 1.0 as possible, and an R-squared value below 0.5 indicates a very poor fit.

For the surrogate-based global method, we initially recommend a small number of maximum iterations, such as 3–5, to get a sense of how the optimization is evolving as the surrogate gets updated globally. If it appears to be changing significantly, then a larger number (used in combination with restart) may be needed.

Figure 9.6 shows a DAKOTA input file that implements surrogate-based global optimization on a multi-objective test function. This input file is named `dakota_su_mogatest1.in` in the `Dakota/test` directory. The first method keyword block contains the keyword `surrogate_based_global`, plus the commands for specifying five as the maximum iterations and the option to replace points in the global surrogate construction. The method block identified as MOGA specifies a multi-objective genetic algorithm optimizer and its controls. The model keyword block specifies a surrogate model. In this case, a `gaussian_process` model is used as a surrogate. The `dace_method_pointer` specifies that the surrogate will be build on 100 Latin Hypercube samples with a seed = 531. The remainder of the input specification deals with the interface to the actual analysis driver and the 2 responses being returned as objective functions from that driver.

9.4 Efficient Global Minimization

Efficient Global Optimization (EGO) is a global optimization technique that employs response surface surrogates [81, 79]. In each EGO iteration, a Gaussian process (GP) approximation for the objective function is constructed based on sample points of the true simulation. The GP allows one to specify the prediction at a new input location as well as the uncertainty associated with that prediction. The key idea in EGO is to maximize an Expected Improvement Function (EIF), defined as the expectation that any point in the search space will provide a better solution than the current best solution, based on the expected values and variances predicted by the GP model. It is important to understand how the use of this EIF leads to optimal solutions. The EIF indicates how much the objective function value at a new potential location is expected to be less than the predicted value at the current best solution. Because the GP model provides a Gaussian distribution at each predicted point, expectations can be calculated. Points with good expected values and even a small variance will have a significant expectation of producing a better solution (exploitation), but so will points that have relatively poor expected values and greater variance (exploration). The EIF incorporates both the idea of choosing points which minimize the objective and choosing points about which there is large prediction uncertainty (e.g., there are few or no samples in that area of the space, and thus the probability may be high that a sample value is potentially lower than other values). Because the uncertainty is higher in regions of the design space with few observations, this provides a balance between exploiting areas of the design space that predict good solutions, and exploring areas where more information is needed.

There are two major differences between our implementation and that of [81]: we do not use a branch and bound method to find points which maximize the EIF. Rather, we use the DIRECT algorithm. Second, we allow for multiobjective optimization and nonlinear least squares including general nonlinear constraints. Constraints are handled through an augmented Lagrangian merit function approach (see Section 9.2.2 and Eqs. 9.11-9.14).

The method is specified as `efficient_global`. Currently we do not expose any specification controls for the underlying Gaussian process model used or for the optimization of the expected improvement function, which is currently performed by the NCSU DIRECT algorithm. An example specification for the EGO algorithm is shown in Figure 9.7:

```
strategy,
    single_method
    tabular_graphics_data
    method_pointer = 'SBGO'

method,
    id_method = 'SBGO'
    surrogate_based_global
    model_pointer = 'SURROGATE'
    approx_method_pointer = 'MOGA'
    max_iterations = 5
    replace_points
    output verbose

method,
    id_method = 'MOGA'
    moga
    output silent
        seed = 10983
    population_size = 300
    max_function_evaluations = 5000
    initialization_type unique_random
    crossover_type shuffle_random
        num_offspring = 2 num_parents = 2
        crossover_rate = 0.8
    mutation_type replace_uniform
        mutation_rate = 0.1
    fitness_type domination_count
    replacement_type below_limit = 6
        shrinkage_percentage = 0.9
    niching_type distance 0.05 0.05
    postprocessor_type
        orthogonal_distance 0.05 0.05
    convergence_type metric_tracker
        percent_change = 0.05 num_generations = 10

model,
    id_model = 'SURROGATE'
    surrogate global
        responses_pointer = 'SURROGATE_RESP'
        dace_method_pointer = 'SAMPLING'
        correction additive zeroth_order
      #  polynomial quadratic
        gaussian_process

variables,
    continuous_design = 3
        initial_point      0     0     0
        upper_bounds       4     4     4
        lower_bounds      -4    -4    -4
        descriptors      'x1'  'x2'  'x3'

responses,
    id_responses = 'SURROGATE_RESP'
    num_objective_functions = 2
    no_gradients
    no_hessians

method,
    id_method = 'SAMPLING'
    model_pointer = 'TRUTH'
    nond_sampling
      samples = 100
      seed = 531
      sample_type lhs
      all_variables

model,
    id_model = 'TRUTH'
    single
      interface_pointer = 'TRUE_FN'
      responses_pointer = 'TRUE_RESP'

interface,
    id_interface = 'TRUE_FN'
    system
      analysis_driver = 'mogatest1'

responses,
    id_responses = 'TRUE_RESP'
    num_objective_functions = 2
    no_gradients
    no_hessians
```

Figure 9.6: DAKOTA input file for the surrogate-based global optimization example.

```
method,
        efficient_global
          seed = 123457

variables,
        continuous_design = 2
          lower_bounds   -2.0 -2.0
          upper_bounds    2.0  2.0
          descriptors     'x1' 'x2'

interface,
        system
          analysis_driver = 'rosenbrock'

responses,
        num_objective_functions = 1
        no_gradients
        no_hessians
```

Figure 9.7: DAKOTA input file for the efficient global optimization example.

Chapter 10

Advanced Strategies

10.1 Overview

DAKOTA's strategy capabilities were developed in order to provide a control layer for managing multiple iterators and models. It was driven by the observed need for "meta-optimization" and other high level systems analysis procedures in real-world engineering design problems. This capability allows the use of existing iterative algorithm and computational model software components as building blocks to accomplish more sophisticated studies, such as hybrid minimization, multistart local minimization, Pareto optimization, or mixed integer nonlinear programming (MINLP). Other strategy-like capabilities are enabled by the model recursion capabilities described in Chapter 11. When these model recursion specifications are sufficient to completely describe a multi-iterator, multi-model solution approach, then a separate strategy specification is not used (see Section 11.6 for examples). In addition, some previous strategy capabilities (i.e., the surrogate-based minimization approaches descibed in Chapter 9) have migrated into the method specification to allow their componentization and reuse elsewhere. This trend will continue in future releases with migration of the MINLP strategy to the method specification, such that only the most generic coordination approaches will remain.

10.2 Hybrid Minimization

In the hybrid minimization strategy (keyword: `hybrid`), a sequence of minimization methods are applied to find an optimal design point. The goal of this strategy is to exploit the strengths of different minimization algorithms through different stages of the minimization process. Global/local optimization hybrids (e.g., genetic algorithms combined with nonlinear programming) are a common example in which the desire for a global optimum is balanced with the need for efficient navigation to a local optimum. An important related feature is that the sequence of minimization algorithms can employ models of varying fidelity. In the global/local case, for example, it would often be advantageous to use a low-fidelity model in the global search phase, followed by use of a more refined model in the local search phase.

The specification for hybrid minimization involves a list of method identifier strings, and each of the corresponding method specifications has the responsibility for identifying the model specification (which may in turn identify variables, interface, and responses specifications) that each method will use (see the DAKOTA Reference Manual [30] and the example discussed below). Currently, only the sequential hybrid approach is available. The `embedded` and `collaborative` approaches are not fully functional at this time.

In the `sequential` hybrid minimization approach, a sequence of minimization methods is invoked in the order specified in the DAKOTA input file. In the default mode, the best solution from each method is used as the starting point for the following method. If the user specifies `num_solutions_transferred` and the method returns multiple optimal solutions (for example, with an evolutionary algorithm or DIRECT or a sampline method), then the specified number of solutions from the previous method will be used to initialize the subsequent method. If the subsequent method cannot accept multiple input points (currently only a few methods such as the genetic algorithms in JEGA allow multiple input points), then multiple instances of the subsequent method are generated, each one initialized by one of the optimal solutions from the previous method. For example, if LHS sampling were run as the first method and the number of solutions transferred was 10 and the DOT conjugate gradient was the second method, there would be 10 instances of `dot_frcg` started, each with a separate LHS sample solution as its initial point. Method switching is governed by the separate convergence controls of each method; that is, *each method is allowed to run to its own internal definition of completion without interference*. Individual method completion may be determined by convergence criteria (e.g., `convergence_tolerance`) or iteration limits (e.g., `max_iterations`). The `adaptive` option is similar, with the difference that the progress of each method is monitored and method switching is enforced according to externally-defined relative progress metrics.

Figure 10.1 shows a DAKOTA input file that specifies a sequential hybrid optimization strategy to solve the "textbook" optimization test problem. This input file is named `dakota_hybrid.in` in the `Dakota/test` directory. The three optimization methods are identified using the `method_list` specification in the strategy section of the input file. The identifier strings listed in the specification are '`GA`' for genetic algorithm, '`PS`' for pattern search, and '`NLP`' for nonlinear programming. Following the strategy keyword block are the three corresponding method keyword blocks. Note that each method has a tag following the `id_method` keyword that corresponds to one of the method names listed in the strategy keyword block. By following the identifier tags from `method` to `model` and from `model` to `variables`, `interface`, and `responses`, it is easy to see the specification linkages for this problem. The GA optimizer runs first and uses model '`M1`' which includes variables '`V1`', interface '`I1`', and responses '`R1`'. Once the GA is complete, the PS optimizer starts from the best GA result and again uses model '`M1`'. Since both GA and PS are nongradient-based optimization methods, there is no need for gradient or Hessian information in the '`R1`' response keyword block. The NLP optimizer runs last, using the best result from the PS method as its starting point. It uses model '`M2`' which includes the same '`V1`' and '`I1`' keyword blocks, but uses the responses keyword block '`R2`' since the full Newton optimizer used in this example (`optpp_newton`) needs analytic gradient and Hessian data to perform its search.

10.3 Multistart Local Minimization

A simple, heuristic, global minimization technique is to use many local minimization runs, each of which is started from a different initial point in the parameter space. This is known as multistart local minimization. This is an attractive strategy in situations where multiple local optima are known or expected to exist in the parameter space. However, there is no theoretical guarantee that the global optimum will be found. This approach combines the efficiency of local minimization methods with a user-specified global stratification (using a specified `starting_points` list, a number of specified `random_starts`, or both; see the DAKOTA Reference Manual [30] for additional specification details). Since solutions for different starting points are independent, parallel computing may be used to concurrently run the local minimizations.

An example input file for multistart local optimization on the "quasi_sine" test function (see `quasi_sine_fcn.C` in `Dakota/test`) is shown in Figure 10.2. The strategy keyword block in the input file contains the keyword `multi_start`, along with the set of starting points (3 random and 5 listed) that will be used for the optimization runs. The other keyword blocks in the input file are similar to what would be used in a single optimization run.

The `quasi_sine` test function has multiple local minima, but there is an overall trend in the function that tends

```
strategy,
        graphics
        hybrid sequential
          method_list = 'GA' 'PS' 'NLP'

method,
        id_method = 'GA'
        model_pointer = 'M1'
        coliny_ea
          seed = 1234
          population_size = 10
          verbose output

method,
        id_method = 'PS'
        model_pointer = 'M1'
        coliny_pattern_search stochastic
          seed = 1234
          initial_delta = 0.1
          threshold_delta = 1.e-4
          solution_accuracy = 1.e-10
          exploratory_moves basic_pattern
          verbose output

method,
        id_method = 'PS2'
        model_pointer = 'M1'
        max_function_evaluations = 10
        coliny_pattern_search stochastic
          seed = 1234
          initial_delta = 0.1
          threshold_delta = 1.e-4
          solution_accuracy = 1.e-10
          exploratory_moves basic_pattern
          verbose output

method,
        id_method = 'NLP'
        model_pointer = 'M2'
        optpp_newton
          gradient_tolerance = 1.e-12
          convergence_tolerance = 1.e-15
          verbose output

model,
        id_model = 'M1'
        single
          variables_pointer = 'V1'
          interface_pointer = 'I1'
          responses_pointer = 'R1'

model,
        id_model = 'M2'
        single
          variables_pointer = 'V1'
          interface_pointer = 'I1'
          responses_pointer = 'R2'

variables,
        id_variables = 'V1'
        continuous_design = 2
          initial_point    0.6    0.7
          upper_bounds     5.8    2.9
          lower_bounds     0.5   -2.9
          descriptors      'x1'   'x2'

interface,
        id_interface = 'I1'
        direct
          analysis_driver=  'text_book'

responses,
        id_responses = 'R1'
        num_objective_functions = 1
        no_gradients
        no_hessians

responses,
        id_responses = 'R2'
        num_objective_functions = 1
        analytic_gradients
        analytic_hessians
```

Figure 10.1: DAKOTA input file for a hybrid optimization strategy.

```
strategy,
        multi_start graphics
          method_pointer = 'NLP'
          random_starts = 3 seed = 123
          starting_points = -.8  -.8
                            -.8   .8
                             .8  -.8
                             .8   .8
                              0.   0.

method,
        id_method = 'NLP'
        dot_bfgs

variables,
        continuous_design = 2
          lower_bounds    -1.0     -1.0
          upper_bounds     1.0  1.0
          descriptors      'x1'     'x2'

interface,
        system #asynchronous
          analysis_driver = 'quasi_sine_fcn'

responses,
        num_objective_functions = 1
        analytic_gradients
        no_hessians
```

Figure 10.2: DAKOTA input file for a multistart local optimization strategy.

```
<<<<< Results summary:
   set_id             x1             x2            x1*            x2*         obj_fn
        1           -0.8           -0.8  -0.8543728666  -0.8543728666   0.5584096919
        2           -0.8            0.8  -0.9998398719    0.177092822    0.291406596
        3            0.8           -0.8    0.177092822  -0.9998398719    0.291406596
        4            0.8            0.8   0.1770928217   0.1770928217   0.0602471946
        5              0              0  0.03572926375  0.03572926375  0.08730499239
        6  -0.7767971993  0.01810943539  -0.7024118387  0.03572951143   0.3165522387
        7  -0.3291571008  -0.7697378755   0.3167607374  -0.4009188363   0.2471403213
        8   0.8704730469   0.7720679005    0.177092899   0.3167611757  0.08256082751
```

Figure 10.3: DAKOTA results summary for a multistart local optimization strategy.

toward the global minimum at $(x1, x2) = (0.177, 0.177)$. See [60] for more information on this test function. Figure 10.3 shows the results summary for the eight local optimizations performed. From the five specified starting points and the 3 random starting points (as identified by the `x1`, `x2` headers), the eight local optima (as identified by the `x1*`, `x2*` headers) are all different and only one of the local optimizations finds the global minimum.

10.4 Pareto Optimization

The Pareto optimization strategy (keyword: `pareto_set`) is one of three multiobjective optimization capabilities discussed in Section 7.3.1. In the Pareto optimization strategy, multiple sets of multiobjective weightings are evaluated. The user can specify these weighting sets in the strategy keyword block using a `multi_objective_weight_sets` list, a number of `random_weight_sets`, or both (see the DAKOTA Reference Manual [30] for additional specification details). Figure 10.4 shows the input commands from the file `dakota_pareto.in` in the `Dakota/test` directory.

DAKOTA performs one multiobjective optimization problem for each set of multiobjective weights. The collection of computed optimal solutions form a Pareto set, which can be useful in making trade-off decisions in engineering design. Since solutions for different multiobjective weights are independent, parallel computing may be used to concurrently execute the multiobjective optimization problems.

Figure 10.5 shows the results summary for the Pareto-set optimization strategy. For the four multiobjective weighting sets (as identified by the `w1`, `w2`, `w3` headers), the local optima (as identified by the `x1`, `x2` headers) are all different and correspond to individual objective function values of (f_1, f_2, f_3) = (0.0,0.5,0.5), (13.1,-1.2,8.16), (532.,33.6,-2.9), and (0.125,0.0,0.0) (note: the composite objective function is tabulated under the `obj_fn` header). The first three solutions reflect exclusive optimization of each of the individual objective functions in turn, whereas the final solution reflects a balanced weighting and the lowest sum of the three objectives. Plotting these (f_1, f_2, f_3) triplets on a 3-dimensional plot results in a Pareto surface (not shown), which is useful for visualizing the trade-offs in the competing objectives.

10.5 Mixed Integer Nonlinear Programming (MINLP)

For DAKOTA 4.2, branch and bound is currently inoperative due to ongoing restructuring of PICO and its incorporation into COLINY. This will be supported again in future releases.

Many nonlinear optimization problems involve a combination of discrete and continuous variables. These are known as mixed integer nonlinear programming (MINLP) problems. A typical MINLP optimization problem is

```
strategy,
        pareto_set
          graphics
          opt_method_pointer = 'NLP'
          multi_objective_weight_sets =
                                  1.   0.   0.
                                  0.   1.   0.
                                  0.   0.   1.
                                .333 .333 .333


method,
        id_method = 'NLP'
        dot_bfgs



model,
        single

variables,
        continuous_design = 2
          initial_point     0.9    1.1
          upper_bounds      5.8    2.9
          lower_bounds      0.5   -2.9
          descriptors       'x1'   'x2'

interface,
        system #asynchronous
          analysis_driver = 'text_book'

responses,
        num_objective_functions = 3
        analytic_gradients
        no_hessians
```

Figure 10.4: DAKOTA input file for the Pareto optimization strategy.

```
<<<<< Results summary:
   set_id            w1            w2            w3            x1            x2        obj_fn
        1             1             0             0   0.9996554048    0.997046351 7.612301561e-11
        2             0             1             0           0.5           2.9          -1.2
        3             0             0             1           5.8 1.12747589e-11          -2.9
        4         0.333         0.333         0.333           0.5   0.5000000041      0.041625
```

Figure 10.5: DAKOTA results summary for the Pareto-set optimization strategy.

formulated as follows:

$$\begin{aligned} \text{minimize:} \quad & f(\mathbf{x}, \mathbf{d}) \\ \text{subject to:} \quad & \mathbf{g}_L \leq \mathbf{g}(\mathbf{x}, \mathbf{d}) \leq \mathbf{g}_U \\ & \mathbf{h}(\mathbf{x}, \mathbf{d}) = \mathbf{h}_t \\ & \mathbf{x}_L \leq \mathbf{x} \leq \mathbf{x}_U \\ & \mathbf{d} \in \{-2, -1, 0, 1, 2\} \end{aligned} \tag{10.1}$$

where $\mathbf{d}$ is a vector whose elements are integer values. In situations where the discrete variables can be temporarily relaxed (i.e., noncategorical discrete variables, see Section 12.2.2), the branch-and-bound algorithm can be applied. Categorical variables (e.g., true/false variables, feature counts, etc.) that are not relaxable cannot be used with the branch and bound strategy. During the branch and bound process, the discrete variables are treated as continuous variables and the integrality conditions on these variables are incrementally enforced through a sequence of optimization subproblems. By the end of this process, an optimal solution that is feasible with respect to the integrality conditions is computed.

DAKOTA's branch and bound strategy (keyword: `branch_and_bound`) can solve optimization problems having either discrete or mixed continuous/discrete variables. This strategy uses the parallel branch-and-bound algorithm from the PICO software package [27, 28] to generate a series of optimization subproblems ("branches"). These subproblems are solved as continuous variable problems using any of DAKOTA's nonlinear optimization algorithms (e.g., DOT, NPSOL). When a solution to a branch is feasible with respect to the integrality constraints, it provides an upper bound on the optimal objective function, which can be used to prune branches with higher objective functions that are not yet feasible. Since solutions for different branches are independent, parallel computing may be used to concurrently execute the optimization subproblems.

PICO, by itself, targets the solution of mixed integer linear programming (MILP) problems, and through coupling with DAKOTA's nonlinear optimizers, is extended to solution of MINLP problems. In the case of MILP problems, the upper bound obtained with a feasible solution is an exact bound and the branch and bound process is provably convergent to the global minimum. For nonlinear problems which may exhibit nonconvexity or multimodality, the process is heuristic in general, since there may be good solutions that are missed during the solution of a particular branch. However, the process still computes a series of locally optimal solutions, and is therefore a natural extension of the results from local optimization techniques for continuous domains. Only with rigorous global optimization of each branch can a global minimum be guaranteed when performing branch and bound on nonlinear problems of unknown structure.

In cases where there are only a few discrete variables and when the discrete values are drawn from a small set, then it may be reasonable to perform a separate optimization problem for all of the possible combinations of the discrete variables. However, this brute force approach becomes computationally intractable if these conditions are not met. The branch-and-bound algorithm will generally require solution of fewer subproblems than the brute force method, although it will still be significantly more expensive than solving a purely continuous design problem.

10.5.1 Example MINLP Problem

As an example, consider the following MINLP problem [41]:

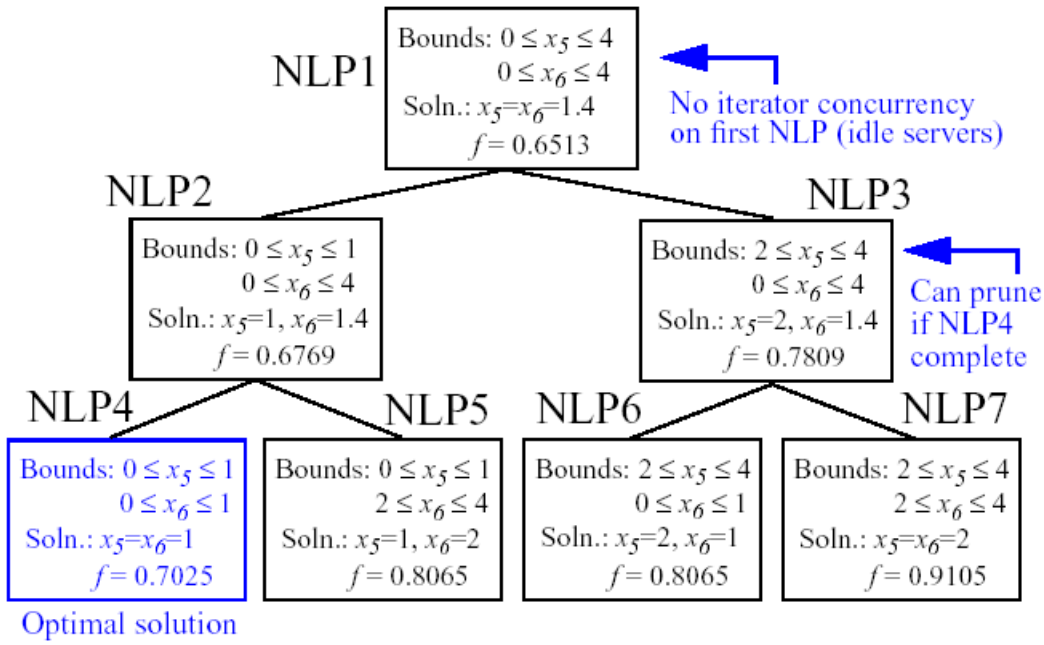


Figure 10.6: Branching history for example MINLP optimization problem.

$$
\begin{aligned}
\text{minimize:} \quad & f(\mathbf{x}) = \sum_{i=1}^{6} (x_i - 1.4)^4 \\
& g_1 = x_1^2 - \frac{x_2}{2} \leq 0 \\
& g_2 = x_2^2 - \frac{x_1}{2} \leq 0 \\
& -10 \leq x_1, x_2, x_3, x_4 \leq 10 \\
& x_5, x_6 \in \{0, 1, 2, 3, 4\}
\end{aligned}
\tag{10.2}
$$

This problem is a variant of the textbook test problem described in Section 22.1. In addition to the introduction of two integer variables, a modified value of 1.4 is used inside the quartic sum to render the continuous solution a non-integral solution.

Figure 10.6 shows the sequence of branches generated for this problem. The first optimization subproblem relaxes the integrality constraint on parameters x_5 and x_6, so that $0 \leq x_5 \leq 4$ and $0 \leq x_6 \leq 4$. The values for x_5 and x_6 at the solution to this first subproblem are $x_5 = x_6 = 1.4$. Since x_5 and x_6 must be integers, the next step in the solution process "branches" on parameter x_5 to create two new optimization subproblems; one with $0 \leq x_5 \leq 1$ and the other with $2 \leq x_5 \leq 4$. Note that, at this first branching, the bounds on x_6 are still $0 \leq x_6 \leq 4$. Next, the two new optimization subproblems are solved. Since they are independent, they can be performed in parallel. The branch-and-bound process continues, operating on both x_5 and x_6 , until a optimization subproblem is solved where x_5 and x_6 are integer-valued. At the solution to this problem, the optimal values for x_5 and x_6 are $x_5 = x_6 = 1$.

In this example problem, the branch-and-bound algorithm executes as few as five and no more than seven opti-

mization subproblems to reach the solution. For comparison, the brute force approach would require 25 optimization problems to be solved (i.e., five possible values for each of x_5 and x_6).

In the example given above, the discrete variables are integer-valued. In some cases, the discrete variables may be real-valued, such as $x \in \{0.0, 0.5, 1.0, 1.5, 2.0\}$. The branch-and-bound algorithm is restricted to work with integer values. Therefore, it is up to the user to perform a transformation between the discrete integer values from DAKOTA and the discrete real values that are passed to the simulation code (see Section 12.2.2). When integrality is not being relaxed, a common mapping is to use the integer value from DAKOTA as the index into a vector of discrete real values. However, when integrality is relaxed, additional logic for interpolating between the discrete real values is needed.

Chapter 11

Models

11.1 Overview

Chapters 4 through 8 have presented the different "iterators" available in DAKOTA. An iterator iterates on a model in order to map a set of variables into a set of responses. This model may involve a simple mapping involving a single interface, or it may involve recursions using sub-iterator and sub-models. These recursion capabilities were developed in order to provide mechanisms for "nesting," "layering," and "recasting" of software components, which allows the use of these components as building blocks to accomplish more sophisticated studies, such as surrogate-based optimization or optimization under uncertainty. In a nested relationship, a sub-iterator is executed using its sub-model for every evaluation of the nested model. In a layered relationship, on the other hand, sub-iterators and sub-models are used only for periodic updates and verifications. And in a recast relationship, the input variable and output response definitions in a sub-model are reformulated in order to support new problem definitions. In each of these cases, the sub-model is of arbitrary type, such that model recursions can be chained together in as long of a sequence as needed (e.g., layered containing nested contained layered containing single in Section 11.6.2). Figure 11.1 displays the model class hierarchy from the DAKOTA Developers Manual [31], with derived classes for single models, nested models, recast models, and two types of surrogate models: data fit and hierarchical/multifidelity. A third type of derived surrogate model supporting reduced-order models (ROM) is planned for future releases.

Section 11.2 describes single models; Section 11.3 describes recast models; Section 11.4 describes surrogate models of the data fit, multifidelity, and ROM type; and Section 11.5 describes nested models. Finally, Section 11.6 presents a number of advanced examples demonstrating model recursion.

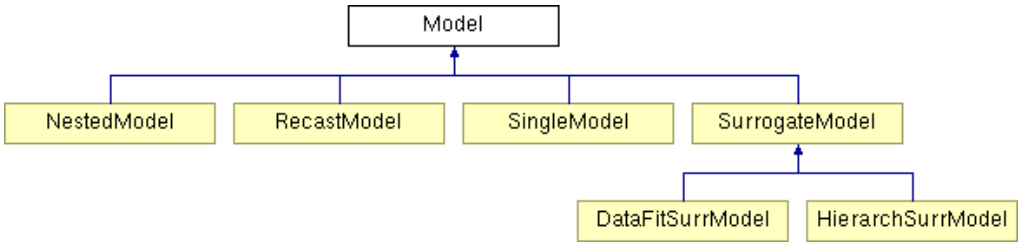


Figure 11.1: The DAKOTA model class hierarchy.

11.2 Single Models

The single model is the simplest model type. It uses a single interface instance (see Chapter 13) to map variables (see Chapter 12) into responses (see Chapter 14). There is no recursion in this case. Refer to the Models chapter in the DAKOTA Reference Manual [30] for additional information on the single model specification.

11.3 Recast Models

The recast model is not directly visible to the user within the input specification. Rather, it is used "behind the scenes" to recast the inputs and outputs of a sub-model for the purposes of reformulating the problem posed to an iterator. Examples include variable and response scaling (see Section 7.3.2), transformations of uncertain variables and associated response derivatives to employ standardized random variables (see Sections 6.3.2 and 6.4.3), multiobjective optimization (see Section 7.3.1), merit functions (see Section 9.2), and expected improvement/feasibility (see Sections 9.4 and 6.3.3). Refer to the DAKOTA Developers Manual [31] for additional details on the mechanics of recasting problem formulations.

11.4 Surrogate Models

Surrogate models provide an approximation to an original, high fidelity "truth" model. A number of surrogate model selections are possible, which are categorized as data fits, multifidelity models, and reduced-order models.

Each of the surrogate model types supports the use of correction factors that improve the local accuracy of the surrogate models. The correction factors force the surrogate models to match the true function values and possibly true function derivatives at the center point of each trust region. Currently, DAKOTA supports either zeroth-, first-, or second-order accurate correction methods, each of which can be applied using either an additive, multiplicative, or combined correction function. For each of these correction approaches, the correction is applied to the surrogate model and the corrected model is then interfaced with whatever algorithm is being employed. The default behavior is that no correction factor is applied.

The simplest correction approaches are those that enforce consistency in function values between the surrogate and original models at a single point in parameter space through use of a simple scalar offset or scaling applied to the surrogate model. First-order corrections such as the first-order multiplicative correction (also known as beta correction [15]) and the first-order additive correction [88] also enforce consistency in the gradients and provide a much more substantial correction capability that is sufficient for ensuring provable convergence in SBO algorithms (see Section 9.2). SBO convergence rates can be further accelerated through the use of second-order corrections which also enforce consistency in the Hessians [35], where the second-order information may involve analytic, finite-difference, or quasi-Newton Hessians.

Correcting surrogate models with additive corrections involves

$$\hat{f_{hi_\alpha}}(\mathbf{x}) = f_{lo}(\mathbf{x}) + \alpha(\mathbf{x}) \tag{11.1}$$

where multifidelity notation has been adopted for clarity. For multiplicative approaches, corrections take the form

$$\hat{f_{hi_\beta}}(\mathbf{x}) = f_{lo}(\mathbf{x})\beta(\mathbf{x}) \tag{11.2}$$

where, for local corrections, $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ are first or second-order Taylor series approximations to the exact

correction functions:

$$\alpha(\mathbf{x}) = A(\mathbf{x_c}) + \nabla A(\mathbf{x_c})^T(\mathbf{x}-\mathbf{x_c}) + \frac{1}{2}(\mathbf{x}-\mathbf{x_c})^T\nabla^2 A(\mathbf{x_c})(\mathbf{x}-\mathbf{x_c}) \tag{11.3}$$

$$\beta(\mathbf{x}) = B(\mathbf{x_c}) + \nabla B(\mathbf{x_c})^T(\mathbf{x}-\mathbf{x_c}) + \frac{1}{2}(\mathbf{x}-\mathbf{x_c})^T\nabla^2 B(\mathbf{x_c})(\mathbf{x}-\mathbf{x_c}) \tag{11.4}$$

where the exact correction functions are

$$A(\mathbf{x}) = f_{hi}(\mathbf{x}) - f_{lo}(\mathbf{x}) \tag{11.5}$$

$$B(\mathbf{x}) = \frac{f_{hi}(\mathbf{x})}{f_{lo}(\mathbf{x})} \tag{11.6}$$

Refer to [35] for additional details on the derivations.

A combination of additive and multiplicative corrections can provide for additional flexibility in minimizing the impact of the correction away from the trust region center. In other words, both additive and multiplicative corrections can satisfy local consistency, but through the combination, global accuracy can be addressed as well. This involves a convex combination of the additive and multiplicative corrections:

$$\hat{f_{hi_\gamma}}(\mathbf{x}) = \gamma \hat{f_{hi_\alpha}}(\mathbf{x}) + (1-\gamma)\hat{f_{hi_\beta}}(\mathbf{x}) \tag{11.7}$$

where γ is calculated to satisfy an additional matching condition, such as matching values at the previous design iterate.

11.4.1 Data Fit Surrogate Models

A surrogate of the *data fit* type is a non-physics-based approximation typically involving interpolation or regression of a set of data generated from the original model. Data fit surrogates can be further characterized by the number of data points used in the fit, where a local approximation (e.g., first or second-order Taylor series) uses data from a single point, a multipoint approximation (e.g., two-point exponential approximations (TPEA) or two-point adaptive nonlinearity approximations (TANA)) uses a small number of data points often drawn from the previous iterates of a particular algorithm, and a global approximation (e.g., polynomial response surfaces, kriging, neural networks, radial basis functions, splines) uses a set of data points distributed over the domain of interest, often generated using a design of computer experiments.

DAKOTA contains several types of surface fitting methods that can be used with optimization and uncertainty quantification methods and strategies such as surrogate-based optimization and optimization under uncertainty. These are: polynomial models (linear, quadratic, and cubic), first-order Taylor series expansion, kriging spatial interpolation, artificial neural networks, multivariate adaptive regression splines, radial basis functions, and moving least squares. With the exception of Taylor series methods, all of the above methods listed in the previous sentence are accessed in DAKOTA through the Surfpack library. All of these surface fitting methods can be applied to problems having an arbitrary number of design parameters. However, surface fitting methods usually are practical only for problems where there are a small number of parameters (e.g., a maximum of somewhere in the range of 30-50 design parameters). The mathematical models created by surface fitting methods have a variety of names in the engineering community. These include surrogate models, meta-models, approximation models, and response surfaces. For this manual, the terms surface fit model and surrogate model are used.

The data fitting methods in DAKOTA include software developed by Sandia researchers and by various researchers in the academic community.

Procedures for Surface Fitting

The surface fitting process consists of three steps: (1) selection of a set of design points, (2) evaluation of the true response quantities (e.g., from a user-supplied simulation code) at these design points, and (3) using the response data to solve for the unknown coefficients (e.g., polynomial coefficients, neural network weights, kriging correlation factors) in the surface fit model. In cases where there is more than one response quantity (e.g., an objective function plus one or more constraints), then a separate surface is built for each response quantity. Currently, the surface fit models are built using only 0th-order information (function values only), although extensions to using higher-order information (gradients and Hessians) are possible. Each surface fitting method employs a different numerical method for computing its internal coefficients. For example, the polynomial surface uses a least-squares approach that employs a singular value decomposition to compute the polynomial coefficients, whereas the kriging surface uses Maximum Likelihood Estimation to compute its correlation coefficients. More information on the numerical methods used in the surface fitting codes is provided in the DAKOTA Developers Manual [31].

The set of design points that is used to construct a surface fit model is generated using either the DDACE software package [127] or the LHS software package [80]. These packages provide a variety of sampling methods including Monte Carlo (random) sampling, Latin hypercube sampling, orthogonal array sampling, central composite design sampling, and Box-Behnken sampling. More information on these software packages is provided in Chapter 5.

Taylor Series

The Taylor series model is purely a local approximation method. That is, it provides local trends in the vicinity of a single point in parameter space. The first-order Taylor series expansion is:

$$\hat{f}(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla_{\mathbf{x}} f(\mathbf{x}_0)^T (\mathbf{x} - \mathbf{x}_0) \tag{11.8}$$

and the second-order expansion is:

$$\hat{f}(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla_{\mathbf{x}} f(\mathbf{x}_0)^T (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \nabla^2_{\mathbf{x}} f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) \tag{11.9}$$

where $\mathbf{x}_0$ is the expansion point in n-dimensional parameter space and $f(\mathbf{x}_0)$, $\nabla_{\mathbf{x}} f(\mathbf{x}_0)$, and $\nabla^2_{\mathbf{x}} f(\mathbf{x}_0)$ are the computed response value, gradient, and Hessian at the expansion point, respectively. As dictated by the responses specification used in building the local surrogate, the gradient may be analytic or numerical and the Hessian may be analytic, numerical, or based on quasi-Newton secant updates.

In general, the Taylor series model is accurate only in the region of parameter space that is close to $\mathbf{x}_0$. While the accuracy is limited, the first-order Taylor series model reproduces the correct value and gradient at the point $\mathbf{x}_0$, and the second-order Taylor series model reproduces the correct value, gradient, and Hessian. This consistency is useful in provably-convergent surrogate-based optimization. The other surface fitting methods do not use gradient information directly in their models, and these methods rely on an external correction procedure in order to satisfy the consistency requirements of provably-convergent SBO.

Two Point Adaptive Nonlinearity Approximation

The TANA-3 method [146] is a multipoint approximation method based on the two point exponential approximation [43]. This approach involves a Taylor series approximation in intermediate variables where the powers used for the intermediate variables are selected to match information at the current and previous expansion points. The

form of the TANA model is:

$$\hat{f}(\mathbf{x}) \approx f(\mathbf{x}_2) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{x}_2) \frac{x_{i,2}^{1-p_i}}{p_i}(x_i^{p_i} - x_{i,2}^{p_i}) + \frac{1}{2}\epsilon(\mathbf{x}) \sum_{i=1}^{n}(x_i^{p_i} - x_{i,2}^{p_i})^2 \tag{11.10}$$

where n is the number of variables and:

$$p_i = 1 + \ln\left[\frac{\frac{\partial f}{\partial x_i}(\mathbf{x}_1)}{\frac{\partial f}{\partial x_i}(\mathbf{x}_2)}\right] \Big/ \ln\left[\frac{x_{i,1}}{x_{i,2}}\right] \tag{11.11}$$

$$\epsilon(\mathbf{x}) = \frac{H}{\sum_{i=1}^{n}(x_i^{p_i} - x_{i,1}^{p_i})^2 + \sum_{i=1}^{n}(x_i^{p_i} - x_{i,2}^{p_i})^2} \tag{11.12}$$

$$H = 2\left[f(\mathbf{x}_1) - f(\mathbf{x}_2) - \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{x}_2) \frac{x_{i,2}^{1-p_i}}{p_i}(x_{i,1}^{p_i} - x_{i,2}^{p_i})\right] \tag{11.13}$$

and $\mathbf{x}_2$ and $\mathbf{x}_1$ are the current and previous expansion points. Prior to the availability of two expansion points, a first-order Taylor series is used.

Linear, Quadratic, and Cubic Polynomial Models

Linear, quadratic, and cubic polynomial models are available in DAKOTA. The form of the linear polynomial model is

$$\hat{f}(\mathbf{x}) \approx c_0 + \sum_{i=1}^{n} c_i x_i \tag{11.14}$$

the form of the quadratic polynomial model is:

$$\hat{f}(\mathbf{x}) \approx c_0 + \sum_{i=1}^{n} c_i x_i + \sum_{i=1}^{n}\sum_{j \geq i}^{n} c_{ij} x_i x_j \tag{11.15}$$

and the form of the cubic polynomial model is:

$$\hat{f}(\mathbf{x}) \approx c_0 + \sum_{i=1}^{n} c_i x_i + \sum_{i=1}^{n}\sum_{j \geq i}^{n} c_{ij} x_i x_j + \sum_{i=1}^{n}\sum_{j \geq i}^{n}\sum_{k \geq j}^{n} c_{ijk} x_i x_j x_k \tag{11.16}$$

In all of the polynomial models, $\hat{f}(\mathbf{x})$ is the response of the polynomial model; the x_i, x_j, x_k terms are the components of the n-dimensional design parameter values; the c_0 , c_i , c_{ij} , c_{ijk} terms are the polynomial coefficients, and n is the number of design parameters. The number of coefficients, n_c, depends on the order of polynomial model and the number of design parameters. For the linear polynomial:

$$n_{c_{linear}} = n + 1 \tag{11.17}$$

for the quadratic polynomial:

$$n_{c_{quad}} = \frac{(n+1)(n+2)}{2} \tag{11.18}$$

and for the cubic polynomial:

$$n_{c_{cubic}} = \frac{(n^3 + 6n^2 + 11n + 6)}{6} \tag{11.19}$$

There must be at least n_c data samples in order to form a fully determined linear system and solve for the polynomial coefficients. In DAKOTA, a least-squares approach involving a singular value decomposition numerical method is applied to solve the linear system.

The utility of the polynomial models stems from two sources: (1) over a small portion of the parameter space, a low-order polynomial model is often an accurate approximation to the true data trends, and (2) the least-squares procedure provides a surface fit that smooths out noise in the data. For this reason, the surrogate-based optimization strategy often is successful when using polynomial models, particularly quadratic models. However, a polynomial surface fit may not be the best choice for modeling data trends over the entire parameter space, unless it is known a priori that the true data trends are close to linear, quadratic, or cubic. See [94] for more information on polynomial models.

Kriging Spatial Interpolation Models

In DAKOTA 4.2, we have 2 versions of spatial interpolation models. They are denoted by `kriging` and `gaussian_process`, respectively. We are in the process of migrating from the `kriging` model to the `gaussian_process` model. For now, both are supported. They are very similar: the differences are explained in more detail below.

The kriging method uses techniques developed in the geostatistics and spatial statistics communities ([19], [84]) to produce smooth, C^2-continuous surface fit models of the response values from a set of data points. The form of the kriging model is

$$\hat{f}(\mathbf{x}) \approx \beta + \mathbf{r}^T \mathbf{R}^{-1}(\mathbf{f} - \beta \mathbf{e}) \tag{11.20}$$

where $\mathbf{x}$ is the current point in n-dimensional parameter space; β is the estimate of the mean response value, r is the correlation vector of terms between $\mathbf{x}$ and the data points, $\mathbf{R}$ is the correlation matrix for all of the data points, $\mathbf{f}$ is the vector of response values, and $\mathbf{e}$ is a vector with all values set to one. The terms in the correlation vector and matrix are computed using a Gaussian correlation function and are dependent on an n-dimensional vector of correlation parameters, $\Theta = \{\theta_1, \ldots, \theta_n\}$. In DAKOTA, a Maximum Likelihood Estimation procedure is performed to compute the correlation parameters for the kriging model. More detail on this kriging approach may be found in [62].

The kriging interpolation model is a nonparametric surface fitting approach. That is, the kriging surface does not assume that there is an underlying trend in the response data. This is in contrast to the quadratic polynomial model and the linear Taylor series model. Since the kriging model is nonparametric, it can be used to model surfaces with slope discontinuities along with multiple local minima and maxima. Kriging interpolation is useful for both SBO and OUU, as well as for studying the global response value trends in the parameter space. This surface fitting method can be constructed using a minimum of $n_{c_{linear}}$ design points, but it is recommended to use at least $n_{c_{quad}}$ design points when possible (refer to Section 11.4.1 for n_c definitions).

The kriging model is guaranteed to pass through all of the response data values that are used to construct the model. Generally, this is a desirable feature. However, if there is considerable numerical noise in the response data, then a surface fitting method that provides some data smoothing (e.g., quadratic polynomial, MARS) may be a better choice for SBO and OUU applications. Another feature of the kriging model is that the predicted response values, $\hat{f}(\mathbf{x})$, decay to the mean value, β, when $\mathbf{x}$ is far from any of the data points from which the kriging model was constructed (i.e., when the model is used for extrapolation). This is neither a positive nor a negative aspect of kriging, but rather a different behavior than is exhibited by the other surface fitting methods. One drawback to the kriging model is that data points in close proximity lead to ill-conditioning in the numerical procedure and the kriging software will terminate if such a situation occurs. For this reason, the user is advised to avoid sample reuse (`reuse_samples = region` and `reuse_samples = all` specifications) when performing surrogate-based optimization.

As mentioned above, there are two surrogate models in DAKOTA 4.2 which provide Gaussian process surrogates, the `kriging` model and the `gaussian_process` model. More details on the `gaussian_process` model can be found in [90]. The differences between these models are as follows:

- Trend Function: In general, a GP model may incorporate a parametric trend function whose purpose is to capture large-scale variations. The trend function in the `kriging` model is a constant, whereas the trend function in the `gaussian_process` model can be specified as a constant, linear,or quadratic trend function. This is specified by the keyword `trend` followed by one of `constant`, `linear`, or `quadratic`.

- Correlation Parameter Determination: Both the kriging and Gaussian process model use a Maximum Likelihood Estimation (MLE) approach to find the optimal values of the hyper-parameters governing the mean and correlation functions. The `kriging` model using a CONMIN local optimizer while the `gaussian_process` model uses a global optimization method called DIRECT. Also, the `gaussian_process` model has an iterative solution approach where the coefficients defining the trend function and the process variance are determined, then these values are used when determining the optimal values for the correlation coefficients, and the process iterates. We have found the optimization of the correlation parameters to be more robust in the `gaussian_process` model. Note that one can pre-define the correlation parameters in `kriging` but not in the `gaussian_process`.

- Ill-conditioning. One of the major problems in determining the governing values for a Gaussian process or kriging model is the fact that the correlation matrix can easily become ill-conditioned when there are too many input points close together. Since the predictions from the Gaussian process model involve inverting the correlation matrix, ill-conditioning can lead to poor predictive capability and should be avoided. The `gaussian_process` model has two features to overcome ill-conditioning. The first is that the algorithm will add a small amount of noise to the diagonal elements of the matrix (this is often referred to as a "nugget") and sometimes this is enough to improve the conditioning. The second is that the user can specify to build the GP based only on a subset of points. The algorithm chooses an optimal subset of points (with respect to predictive capability on the remaining unchosen points) using a greedy heuristic. This option is specified with the keyword `point_selection` in the input file. The `kriging` model does not have provisions to deal with ill-conditioning: it stops if the problem is too ill-conditioned.

Artificial Neural Network (ANN) Models

The ANN surface fitting method in DAKOTA employs a stochastic layered perceptron (SLP) artificial neural network based on the direct training approach of Zimmerman [147]. The SLP ANN method is designed to have a lower training cost than traditional ANNs. This is a useful feature for SBO and OUU where new ANNs are constructed many times during the optimization process (i.e., one ANN for each response function, and new

ANNs for each optimization iteration). The form of the SLP ANN model is

$$\hat{f}(\mathbf{x}) \approx \tanh(\tanh((\mathbf{x}\mathbf{A}_0 + \theta_0)\mathbf{A}_1 + \theta_1)) \tag{11.21}$$

where $\mathbf{x}$ is the current point in n-dimensional parameter space, and the terms $\mathbf{A}_0, \theta_0, \mathbf{A}_1, \theta_1$ are the matrices and vectors that correspond to the neuron weights and offset values in the ANN model. These terms are computed during the ANN training process, and are analogous to the polynomial coefficients in a quadratic surface fit. A singular value decomposition method is used in the numerical methods that are employed to solve for the weights and offsets.

The SLP ANN is a non parametric surface fitting method. Thus, along with kriging and MARS, it can be used to model data trends that have slope discontinuities as well as multiple maxima and minima. However, unlike kriging, the ANN surface is not guaranteed to exactly match the response values of the data points from which it was constructed. This ANN can be used with SBO and OUU strategies. As with kriging, this ANN can be constructed from fewer than $n_{c_{quad}}$ data points, however, it is a good rule of thumb to use at least $n_{c_{quad}}$ data points when possible.

Multivariate Adaptive Regression Spline (MARS) Models

This surface fitting method uses multivariate adaptive regression splines from the MARS3.5 package [48] developed at Stanford University.

The form of the MARS model is based on the following expression:

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^{M} a_m B_m(\mathbf{x}) \tag{11.22}$$

where the a_m are the coefficients of the truncated power basis functions B_m, and M is the number of basis functions. The MARS software partitions the parameter space into subregions, and then applies forward and backward regression methods to create a local surface model in each subregion. The result is that each subregion contains its own basis functions and coefficients, and the subregions are joined together to produce a smooth, C^2-continuous surface model.

MARS is a nonparametric surface fitting method and can represent complex multimodal data trends. The regression component of MARS generates a surface model that is not guaranteed to pass through all of the response data values. Thus, like the quadratic polynomial model, it provides some smoothing of the data. The MARS reference material does not indicate the minimum number of data points that are needed to create a MARS surface model. However, in practice it has been found that at least $n_{c_{quad}}$, and sometimes as many as 2 to 4 times $n_{c_{quad}}$, data points are needed to keep the MARS software from terminating. Provided that sufficient data samples can be obtained, MARS surface models can be useful in SBO and OUU applications, as well as in the prediction of global trends throughout the parameter space.

Radial Basis Functions

Radial basis functions are functions whose value typically depends on the distance from a center point, called the centroid, $\mathbf{c}$. The surrogate model approximation is then built up as the sum of K weighted radial basis functions:

$$\hat{f}(\mathbf{x}) = \sum_{k=1}^{K} w_k \phi(\| \mathbf{x} - \mathbf{c_k} \|) \tag{11.23}$$

where the ϕ are the individual radial basis functions. These functions can be of any form, but often a Gaussian bell-shaped function or splines are used. Our implementation uses a Gaussian radial basis function. The weights are determined via a linear least squares solution approach. See [101] for more details.

Moving Least Squares

Moving Least Squares can be considered a more specialized version of linear regression models. In linear regression, one usually attempts to minimize the sum of the squared residuals, where the residual is defined as the difference between the surrogate model and the true model at a fixed number of points. In weighted least squares, the residual terms are weighted so the determination of the optimal coefficients governing the polynomial regression function, denoted by $\hat{f}(\mathbf{x})$, are obtained by minimizing the weighted sum of squares at N data points:

$$\sum_{n=1}^{N} w_n (\| \hat{f}(\mathbf{x_n}) - f(\mathbf{x_n}) \|) \tag{11.24}$$

Moving least squares is a further generalization of weighted least squares where the weighting is "moved" or recalculated for every new point where a prediction is desired. [95] The implementation of moving least squares is still under development. We have found that it works well in trust region methods where the surrogate model is constructed in a constrained region over a few points. It does not appear to be working as well globally, at least at this point in time.

11.4.2 Multifidelity Surrogate Models

A second type of surrogate is the *model hierarchy* type (also called multifidelity, variable fidelity, variable complexity, etc.). In this case, a model that is still physics-based but is of lower fidelity (e.g., coarser discretization, reduced element order, looser convergence tolerances, omitted physics) is used as the surrogate in place of the high-fidelity model. For example, an inviscid, incompressible Euler CFD model on a coarse discretization could be used as a low-fidelity surrogate for a high-fidelity Navier-Stokes model on a fine discretization.

11.4.3 Reduced Order Models

A third type of surrogate model involves *reduced-order modeling* techniques such as proper orthogonal decomposition (POD) in computational fluid dynamics (also known as principal components analysis or Karhunen-Loeve in other fields) or spectral decomposition (also known as modal analysis) in structural dynamics. These surrogate models are generated directly from a high-fidelity model through the use of a reduced basis (e.g., eigenmodes for modal analysis or left singular vectors for POD) and projection of the original high-dimensional system down to a

small number of generalized coordinates. These surrogates are still physics-based (and may therefore have better predictive qualities than data fits), but do not require multiple system models of varying fidelity (as required for model hierarchy surrogates).

11.5 Nested Models

Nested models utilize a sub-iterator and a sub-model to perform a complete iterative study as part of every evaluation of the model. This sub-iteration accepts variables from the outer level, performs the sub-level analysis, and computes a set of sub-level responses which are passed back up to the outer level. As described in the Models chapter of the Reference Manual [30], mappings are employed for both the variable inputs to the sub-model and the response outputs from the sub-model.

In the variable mapping case, primary and secondary variable mapping specifications are used to map from the top-level variables into the sub-model variables. These mappings support three possibilities in any combination: (1) insertion of an active top-level variable value into an identified sub-model distribution parameter for an identified active sub-model variable, (2) insertion of an active top-level variable value into an identified active sub-model variable value, and (3) addition of an active top-level variable value as an inactive sub-model variable, augmenting the active sub-model variables.

In the response mapping case, primary and secondary response mapping specifications are used to map from the sub-model responses back to the top-level responses. These specifications provide real-valued multipliers that are applied to the sub-iterator response results to define the outer level response set. These nested data results may be combined with non-nested data through use of the "optional interface" component within nested models.

Several examples of nested model usage are provided in the following section.

11.6 Advanced Examples

The surrogate and nested model constructs admit a wide variety of multi-iterator, multi-model solution approaches. For example, optimization within optimization (for hierarchical multidisciplinary optimization), uncertainty quantification within uncertainty quantification (for second-order probability), uncertainty quantification within optimization (for optimization under uncertainty), and optimization within uncertainty quantification (for uncertainty of optima) are all supported, with and without surrogate model indirection. Three important examples are highlighted: second-order probability, optimization under uncertainty, and surrogate-based uncertainty quantification.

11.6.1 Second-order probability

Second-order probability approaches employ nested models to embed one uncertainty quantification (UQ) within another. The outer level UQ is commonly linked to epistemic uncertainties (also known as reducible uncertainties) resulting from a lack of knowledge, and the inner UQ is commonly linked to aleatory uncertainties (also known as irreducible uncertainties) that are inherent in nature. The outer level generates sets of realizations, typically from sampling within interval distributions. These realizations define values for distribution parameters used in a probabilistic analysis for the inner level UQ. The term "second-order" derives from this use of distributions on distributions and the generation of statistics on statistics. These approaches can be considered to be a special case of imprecise probability theory.

A sample input file is shown in Figure 11.2, in which the outer epistemic level variables are defined as intervals. Samples will be generated from these intervals to select means for X and Y that are employed in an inner level reliability analysis of the cantilever problem (see Section 22.9). Figure 11.3 shows excerpts from the resulting output. In this particular example, the outer loop generates 50 possible realizations of epistemic variables, which are then sent to the inner loop to calculate statistics such as the mean weight, and cumulative distribution function for the stress and displacement reliability indices. Thus, the outer loop has 50 possible values for the mean weight but there is no distribution structure on these 50 samples. So, only the minimum and maximum value are reported. Similarly, the minimum and maximum values of the CCDF for the stress and displacement reliability indices are reported.

When performing an epistemic analysis, response levels and probability levels should only be defined in the inner loop. For example, if one wants to generate an interval around possible CDFs or CCDFS, we suggest defining a number of probability levels in the inner loop (0.1, 0.2, 0.3, etc). For each epistemic instance, these will be calculated during the inner loop and reported back to the outer loop. In this way, there will be an ensemble of CDF percentiles (for example) and one will have interval bounds for each of these percentile levels defined.

Also note that it is possible to define the epistemic outer loop using uniform variables instead of interval variables. The process of generating the epistemic values is essentially the same in both cases. However, if the outer loop variables are defined to be uniform, the outer loop results will be reported as statistics (such as mean and standard deviation) and not merely intervals. It is important to note that these outer level statistics are only meaningful to the extent that the outer level probabilities are meaningful (which would not be the case for sampling from epistemic intervals, since the actual probabilities would not be known to be uniform). Finally, although the epistemic variables are often values defining distribution parameters for the inner loop, they are not required to be: they can just be separate uncertain variables in the problem.

11.6.2 Optimization Under Uncertainty (OUU)

Optimization under uncertainty (OUU) approaches incorporate an uncertainty quantification method within the optimization process. This is often needed in engineering design problems when one must include the effect of input parameter uncertainties on the response functions of interest. A typical engineering example of OUU would minimize the probability of failure of a structure for a set of applied loads, where there is uncertainty in the loads and/or material properties of the structural components.

In OUU, a nondeterministic method is used to evaluate the effect of uncertain variable distributions on response functions of interest (refer to Chapter 6 for additional information on nondeterministic analysis). Statistics on these response functions are then included in the objective and constraint functions of an optimization process. Different UQ methods can have very different features from an optimization perspective, leading to the tailoring of optimization under uncertainty approaches to particular underlying UQ methodologies.

If the UQ method is sampling based, then three approaches are currently supported: nested OUU, surrogate-based OUU, and trust-region surrogate-based OUU. Additional details and computational results are provided in [36].

Another class of OUU algorithms is called reliability-based design optimization (RBDO). RBDO methods are used to perform design optimization accounting for reliability metrics. The reliability analysis capabilities described in Section 6.3 provide a rich foundation for exploring a variety of RBDO formulations. [32] investigated bi-level, fully-analytic bi-level, and first-order sequential RBDO approaches employing underlying first-order reliability assessments. [33] investigated fully-analytic bi-level and second-order sequential RBDO approaches employing underlying second-order reliability assessments.

When using stochastic expansions for UQ, analytic moments and analytic design sensitivities can be exploited as described in [42]. Several approaches for obtaining design sensitivities of statistical metrics are discussed in Section 11.6.2.

```
strategy,
        single_method
          method_pointer = 'EPISTEMIC'

method,
        id_method = 'EPISTEMIC'
        model_pointer = 'EPIST_M'
        nond_sampling
          samples = 50 seed = 12347

model,
        id_model = 'EPIST_M'
        nested
          variables_pointer  = 'EPIST_V'
          sub_method_pointer = 'ALEATORY'
          responses_pointer  = 'EPIST_R'
          primary_variable_mapping   = 'X'    'Y'
          secondary_variable_mapping = 'mean' 'mean'
          primary_response_mapping   = 1. 0. 0. 0. 0. 0. 0. 0.
                                       0. 0. 0. 0. 1. 0. 0. 0.
                                       0. 0. 0. 0. 0. 0. 0. 1.

variables,
        id_variables = 'EPIST_V'
        interval_uncertain =2
          num_intervals = 1 1
          interval_probs = 1.0 1.0
          interval_bounds = 400. 600. 800. 1200.
          descriptors       'X_mean' 'Y_mean'

responses,
        id_responses = 'EPIST_R'
        num_response_functions = 3
        response_descriptors = 'mean_wt' 'ccdf_beta_s' 'ccdf_beta_d'
        no_gradients
        no_hessians

method,
        id_method = 'ALEATORY'
        model_pointer = 'ALEAT_M'
        nond_local_reliability
          mpp_search no_approx
          num_response_levels = 0 1 1
          response_levels = 0.0 0.0
          compute reliabilities
          complementary distribution

model,
        id_model = 'ALEAT_M'
        single
          variables_pointer = 'ALEAT_V'
          interface_pointer = 'ALEAT_I'
          responses_pointer = 'ALEAT_R'

variables,
        id_variables = 'ALEAT_V'
        continuous_design = 2
          initial_point    2.4522 3.8826
          descriptors'beam_width' 'beam_thickness'
        normal_uncertain = 4
          means             =  40000. 29.E+6 500. 1000.
          std_deviations    =  2000. 1.45E+6 100. 100.
          descriptors       =  'R' 'E' 'X' 'Y'

interface,
        id_interface = 'ALEAT_I'
        direct
          analysis_driver = 'cantilever'
          deactivate evaluation_cache restart_file

responses,
        id_responses = 'ALEAT_R'
        num_response_functions = 3
        response_descriptors = 'weight' 'stress' 'displ'
        analytic_gradients
        no_hessians
```

Figure 11.2: DAKOTA input file for the second-order probability example.

```
Statistics based on 50 samples:

Min and Max values for each response function:
mean_wt:  Min = 9.52091e+00  Max = 9.52091e+00
ccdf_beta_s:  Min = 1.80013e+00  Max = 4.07440e+00
ccdf_beta_d:  Min = 1.94032e+00  Max = 3.76281e+00
```

Figure 11.3: Second-order statistics on reliability indices for cantilever problem.

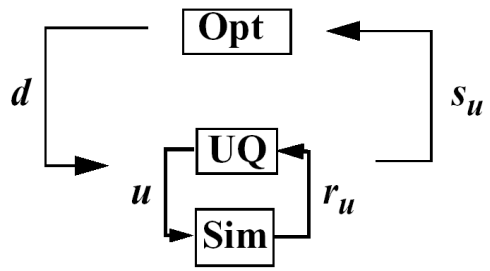


Figure 11.4: Formulation 1: Nested OUU.

Finally, when employing epistemic methods for UQ, the set of statistics available for use within optimization are interval-based. Robustness metrics typically involve the width of the intervals, and reliability metrics typically involve the worst case upper or lower bound of the interval.

Each of these OUU methods is overviewed in the following sections.

Nested OUU

In the case of a nested approach, the optimization loop is the outer loop which seeks to optimize a nondeterministic quantity (e.g., minimize probability of failure). The uncertainty quantification (UQ) inner loop evaluates this nondeterministic quantity (e.g., computes the probability of failure) for each optimization function evaluation. Figure 11.4 depicts the nested OUU iteration where $\mathbf{d}$ are the design variables, $\mathbf{u}$ are the uncertain variables characterized by probability distributions, $\mathbf{r_u}(\mathbf{d}, \mathbf{u})$ are the response functions from the simulation, and $\mathbf{s_u}(\mathbf{d})$ are the statistics generated from the uncertainty quantification on these response functions.

Figure 11.5 shows a DAKOTA input file for a nested OUU example problem that is based on the textbook test problem. This input file is named `dakota_ouu1_tb.in` in the `Dakota/test` directory. In this example, the objective function contains two probability of failure estimates, and an inequality constraint contains another probability of failure estimate. For this example, failure is defined to occur when one of the textbook response functions exceeds its threshold value. The strategy keyword block at the top of the input file identifies this as an OUU problem. The strategy keyword block is followed by the optimization specification, consisting of the optimization method, the continuous design variables, and the response quantities that will be used by the optimizer. The mapping matrices used for incorporating UQ statistics into the optimization response data are described in the DAKOTA Reference Manual [30]. The uncertainty quantification specification includes the UQ method, the uncertain variable probability distributions, the interface to the simulation code, and the UQ response attributes. As with other complex DAKOTA input files, the identification tags given in each keyword block can be used to follow the relationships among the different keyword blocks.

```
strategy,
        single_method
          method_pointer = 'OPTIM'

method,
        id_method = 'OPTIM'
        model_pointer = 'OPTIM_M'
        npsol_sqp
          convergence_tolerance = 1.e-8

model,
        id_model = 'OPTIM_M'
        nested
          variables_pointer  = 'OPTIM_V'
          sub_method_pointer = 'UQ'
          responses_pointer  = 'OPTIM_R'
          primary_response_mapping   = 0. 0. 1. 0. 0. 1. 0. 0. 0.
          secondary_response_mapping = 0. 0. 0. 0. 0. 0. 0. 0. 1.

variables,
        id_variables = 'OPTIM_V'
        continuous_design = 2
          initial_point    1.8    1.0
          upper_bounds     2.164  4.0
          lower_bounds     1.5    0.0
          descriptors      'd1'   'd2'

responses,
        id_responses = 'OPTIM_R'
        num_objective_functions = 1
        num_nonlinear_inequality_constraints = 1
        nonlinear_inequality_upper_bounds = .1
        numerical_gradients
          method_source dakota
          interval_type central
          fd_gradient_step_size = 1.e-1
        no_hessians

method,
        id_method = 'UQ'
        model_pointer = 'UQ_M'
        nond_sampling,
          samples = 50 seed = 1 sample_type lhs
          response_levels = 3.6e+11 1.2e+05 3.5e+05
          complementary distribution

model,
        id_model = 'UQ_M'
        single
          variables_pointer = 'UQ_V'
          interface_pointer = 'UQ_I'
          responses_pointer = 'UQ_R'

variables,
        id_variables = 'UQ_V'
        continuous_design = 2
        normal_uncertain = 2
          means          =  248.89, 593.33
          std_deviations =   12.4,  29.7
          descriptors    =  'nuv1'  'nuv2'
        uniform_uncertain = 2
          lower_bounds =  199.3,  474.63
          upper_bounds =  298.5,  712.
          descriptors  =  'uuv1'  'uuv2'
        weibull_uncertain = 2
          alphas       =    12., 30.
          betas        =   250., 590.
          descriptors  =  'wuv1'  'wuv2'

interface,
        id_interface = 'UQ_I'
        system asynch evaluation_concurrency = 5
          analysis_driver = 'text_book_ouu'

responses,
        id_responses = 'UQ_R'
        num_response_functions = 3
        no_gradients
        no_hessians
```

Figure 11.5: DAKOTA input file for the nested OUU example.

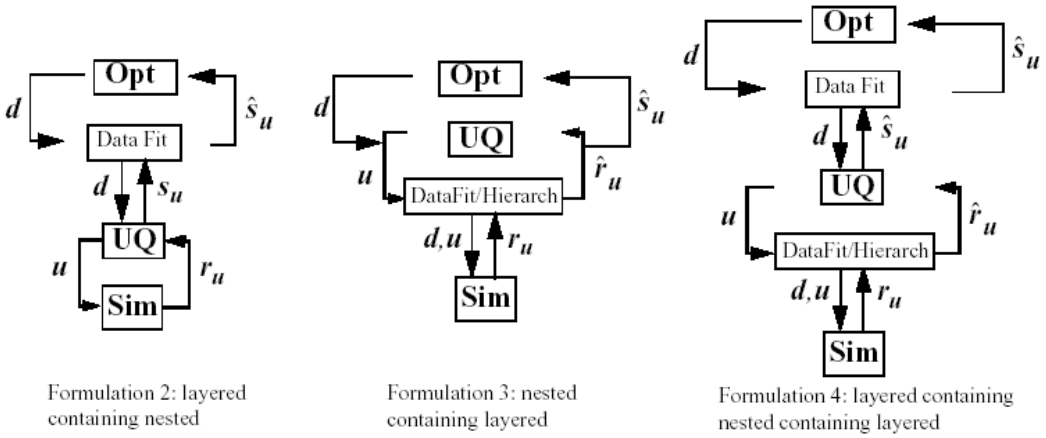


Figure 11.6: Formulations 2, 3, and 4 for Surrogate-based OUU.

Latin hypercube sampling is used as the UQ method in this example problem. Thus, each evaluation of the response functions by the optimizer entails 50 Latin hypercube samples. In general, nested OUU studies can easily generate several thousand function evaluations and gradient-based optimizers may not perform well due to noisy or insensitive statistics resulting from under-resolved sampling. These observations motivate the use of surrogate-based approaches to OUU.

Other nested OUU examples in the `Dakota/test` directory include `dakota_ouu1_tbch.in`, which adds an additional interface for including deterministic data in the textbook OUU problem, and `dakota_ouu1_cantilever.in`, which solves the cantilever OUU problem (see Section 22.9) with a nested approach. For each of these files, the "`1`" identifies formulation 1, which is short-hand for the nested approach.

Surrogate-Based OUU (SBOUU)

Surrogate-based optimization under uncertainty strategies can be effective in reducing the expense of OUU studies. Possible formulations include use of a surrogate model at the optimization level, at the uncertainty quantification level, or at both levels. These surrogate models encompass both data fit surrogates (at the optimization or UQ level) and model hierarchy surrogates (at the UQ level only). Figure 11.6 depicts the different surrogate-based formulations where $\hat{\mathbf{r}}_\mathbf{u}$ and $\hat{\mathbf{s}}_\mathbf{u}$ are approximate response functions and approximate response statistics, respectively, generated from the surrogate models.

SBOUU examples in the `Dakota/test` directory include `dakota_sbouu2_tbch.in`, `dakota_sbouu3_tbch.in`, and `dakota_sbouu4_tbch.in`, which solve the textbook OUU problem, and `dakota_sbouu2_cantilever.in`, `dakota_sbouu3_cantilever.in`, and `dakota_sbouu4_cantilever.in`, which solve the cantilever OUU problem (see Section 22.9). For each of these files, the "`2`," "`3`," and "`4`" identify formulations 2, 3, and 4, which are short-hand for the "layered containing nested," "nested containing layered," and "layered containing nested containing layered" surrogate-based formulations, respectively. In general, the use of surrogates greatly reduces the computational expense of these OUU study. However, without restricting and verifying the steps in the approximate optimization cycles, weaknesses in the data fits can be exploited and poor solutions may be obtained. The need to maintain accuracy of results leads to the use of trust-region surrogate-based approaches.

Trust-Region Surrogate-Based OUU (TR-SBOUU)

The TR-SBOUU approach applies the trust region logic of deterministic SBO (see Section 9.2) to SBOUU. Trust-region verifications are applicable when surrogates are used at the optimization level, i.e., formulations 2 and 4. As a result of periodic verifications and surrogate rebuilds, these techniques are more expensive than SBOUU; however they are more reliable in that they maintain the accuracy of results. Relative to nested OUU (formulation 1), TR-SBOUU tends to be less expensive and less sensitive to initial seed and starting point.

TR-SBOUU examples in the `Dakota/test` directory include `dakota_trsbouu2_tbch.in` and `dakota_trsbouu4_tbch.in`, which solve the textbook OUU problem, and `dakota_trsbouu2_cantilever.in` and `dakota_trsbouu4_cantilever.in`, which solve the cantilever OUU problem (see Section 22.9).

Computational results for several example problems are available in [36].

Bi-level RBDO

The simplest and most direct RBDO approach is the bi-level approach in which a full reliability analysis is performed for every optimization function evaluation. This involves a nesting of two distinct levels of optimization within each other, one at the design level and one at the MPP search level.

Since an RBDO problem will typically specify both the $\bar{z}$ level and the $\bar{p}/\bar{\beta}$ level, one can use either the RIA or the PMA formulation for the UQ portion and then constrain the result in the design optimization portion. In particular, RIA reliability analysis maps $\bar{z}$ to p/β, so RIA RBDO constrains p/β:

$$\begin{aligned} \text{minimize} \quad & f \\ \text{subject to} \quad & \beta \geq \bar{\beta} \\ \text{or} \quad & p \leq \bar{p} \end{aligned} \tag{11.25}$$

And PMA reliability analysis maps $\bar{p}/\bar{\beta}$ to z, so PMA RBDO constrains z:

$$\begin{aligned} \text{minimize} \quad & f \\ \text{subject to} \quad & z \geq \bar{z} \end{aligned} \tag{11.26}$$

where $z \geq \bar{z}$ is used as the RBDO constraint for a cumulative failure probability (failure defined as $z \leq \bar{z}$) but $z \leq \bar{z}$ would be used as the RBDO constraint for a complementary cumulative failure probability (failure defined as $z \geq \bar{z}$). It is worth noting that DAKOTA is not limited to these types of inequality-constrained RBDO formulations; rather, they are convenient examples. DAKOTA supports general optimization under uncertainty mappings [36] which allow flexible use of statistics within multiple objectives, inequality constraints, and equality constraints.

In `Dakota/test`, the `dakota_rbdo_cantilever.in`, `dakota_rbdo_short_column.in`, and `dakota_rbdo_steel_column.in` input files solve the cantilever (see Section 22.9), short column (see Section 22.8), and steel column (see Section 22.10) OUU problems using a bi-level RBDO approach employing numerical design gradients.

An important performance enhancement for bi-level methods is the use of sensitivity analysis to analytically compute the design gradients of probability, reliability, and response levels. When design variables are separate from the uncertain variables (i.e., they are not distribution parameters), then the following first-order expressions

may be used [75, 82, 3]:

$$\nabla_{\mathbf{d}} z = \nabla_{\mathbf{d}} g \tag{11.27}$$

$$\nabla_{\mathbf{d}} \beta_{cdf} = \frac{1}{\| \nabla_{\mathbf{u}} G \|} \nabla_{\mathbf{d}} g \tag{11.28}$$

$$\nabla_{\mathbf{d}} p_{cdf} = -\phi(-\beta_{cdf}) \nabla_{\mathbf{d}} \beta_{cdf} \tag{11.29}$$

where it is evident from Eqs. 6.12-6.13 that $\nabla_{\mathbf{d}} \beta_{ccdf} = -\nabla_{\mathbf{d}} \beta_{cdf}$ and $\nabla_{\mathbf{d}} p_{ccdf} = -\nabla_{\mathbf{d}} p_{cdf}$. In the case of second-order integrations, Eq. 11.29 must be expanded to include the curvature correction. For Breitung's correction (Eq. 6.39),

$$\nabla_{\mathbf{d}} p_{cdf} = \left[\Phi(-\beta_p) \sum_{i=1}^{n-1} \left(\frac{-\kappa_i}{2(1+\beta_p \kappa_i)^{\frac{3}{2}}} \prod_{\substack{j=1 \\ j \neq i}}^{n-1} \frac{1}{\sqrt{1+\beta_p \kappa_j}} \right) - \phi(-\beta_p) \prod_{i=1}^{n-1} \frac{1}{\sqrt{1+\beta_p \kappa_i}} \right] \nabla_{\mathbf{d}} \beta_{cdf} \tag{11.30}$$

where $\nabla_{\mathbf{d}} \kappa_i$ has been neglected and $\beta_p \geq 0$ (see Section 6.3.2). Other approaches assume the curvature correction is nearly independent of the design variables [107], which is equivalent to neglecting the first term in Eq. 11.30.

To capture second-order probability estimates within an RIA RBDO formulation using well-behaved β constraints, a generalized reliability index can be introduced where, similar to Eq. 6.10,

$$\beta^*_{cdf} = -\Phi^{-1}(p_{cdf}) \tag{11.31}$$

for second-order p_{cdf}. This reliability index is no longer equivalent to the magnitude of $\mathbf{u}$, but rather is a convenience metric for capturing the effect of more accurate probability estimates. The corresponding generalized reliability index sensitivity, similar to Eq. 11.29, is

$$\nabla_{\mathbf{d}} \beta^*_{cdf} = -\frac{1}{\phi(-\beta^*_{cdf})} \nabla_{\mathbf{d}} p_{cdf} \tag{11.32}$$

where $\nabla_{\mathbf{d}} p_{cdf}$ is defined from Eq. 11.30. Even when $\nabla_{\mathbf{d}} g$ is estimated numerically, Eqs. 11.27-11.32 can be used to avoid numerical differencing across full reliability analyses.

When the design variables are distribution parameters of the uncertain variables, $\nabla_{\mathbf{d}} g$ is expanded with the chain rule and Eqs. 11.27 and 11.28 become

$$\nabla_{\mathbf{d}} z = \nabla_{\mathbf{d}} \mathbf{x} \nabla_{\mathbf{x}} g \tag{11.33}$$

$$\nabla_{\mathbf{d}} \beta_{cdf} = \frac{1}{\| \nabla_{\mathbf{u}} G \|} \nabla_{\mathbf{d}} \mathbf{x} \nabla_{\mathbf{x}} g \tag{11.34}$$

where the design Jacobian of the transformation ($\nabla_{\mathbf{d}} \mathbf{x}$) may be obtained analytically for uncorrelated $\mathbf{x}$ or semi-analytically for correlated $\mathbf{x}$ ($\nabla_{\mathbf{d}} \mathbf{L}$ is evaluated numerically) by differentiating Eqs. 6.15 and 6.16 with respect to the distribution parameters. Eqs. 11.29-11.32 remain the same as before. For this design variable case, all required information for the sensitivities is available from the MPP search.

Since Eqs. 11.27-11.34 are derived using the Karush-Kuhn-Tucker optimality conditions for a converged MPP, they are appropriate for RBDO using AMV+, AMV^2+, TANA, FORM, and SORM, but not for RBDO using MVFOSM, MVSOSM, AMV, or AMV^2.

In `Dakota/test`, the `dakota_rbdo_cantilever_analytic.in` and `dakota_rbdo_short_column_analytic.in` input files solve the cantilever and short column OUU problems using a bi-level RBDO approach with analytic design gradients and first-order limit state approximations. The `dakota_rbdo_cantilever_analytic2.in`, `dakota_rbdo_short_column_analytic2.in`, and `dakota_rbdo_steel_column_analytic2.in` input files also employ analytic design gradients, but are extended to employ second-order limit state approximations and integrations.

Sequential/Surrogate-based RBDO

An alternative RBDO approach is the sequential approach, in which additional efficiency is sought through breaking the nested relationship of the MPP and design searches. The general concept is to iterate between optimization and uncertainty quantification, updating the optimization goals based on the most recent probabilistic assessment results. This update may be based on safety factors [140] or other approximations [26].

A particularly effective approach for updating the optimization goals is to use the $p/\beta/z$ sensitivity analysis of Eqs. 11.27-11.34 in combination with local surrogate models [148]. In [32] and [33], first-order and second-order Taylor series approximations were employed within a trust-region model management framework [60] in order to adaptively manage the extent of the approximations and ensure convergence of the RBDO process. Surrogate models were used for both the objective function and the constraints, although the use of constraint surrogates alone is sufficient to remove the nesting.

In particular, RIA trust-region surrogate-based RBDO employs surrogate models of f and p/β within a trust region Δ^k centered at $\mathbf{d}_c$. For first-order surrogates:

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{d}_c) + \nabla_d f(\mathbf{d}_c)^T(\mathbf{d}-\mathbf{d}_c) \\ \text{subject to} \quad & \beta(\mathbf{d}_c) + \nabla_d \beta(\mathbf{d}_c)^T(\mathbf{d}-\mathbf{d}_c) \geq \bar{\beta} \\ \text{or} \quad & p(\mathbf{d}_c) + \nabla_d p(\mathbf{d}_c)^T(\mathbf{d}-\mathbf{d}_c) \leq \bar{p} \\ & \| \mathbf{d}-\mathbf{d}_c \|_\infty \leq \Delta^k \end{aligned} \tag{11.35}$$

and for second-order surrogates:

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{d}_c) + \nabla_{\mathbf{d}} f(\mathbf{d}_c)^T(\mathbf{d}-\mathbf{d}_c) + \tfrac{1}{2}(\mathbf{d}-\mathbf{d}_c)^T\nabla^2_{\mathbf{d}} f(\mathbf{d}_c)(\mathbf{d}-\mathbf{d}_c) \\ \text{subject to} \quad & \beta(\mathbf{d}_c) + \nabla_{\mathbf{d}} \beta(\mathbf{d}_c)^T(\mathbf{d}-\mathbf{d}_c) + \tfrac{1}{2}(\mathbf{d}-\mathbf{d}_c)^T\nabla^2_{\mathbf{d}} \beta(\mathbf{d}_c)(\mathbf{d}-\mathbf{d}_c) \geq \bar{\beta} \\ \text{or} \quad & p(\mathbf{d}_c) + \nabla_{\mathbf{d}} p(\mathbf{d}_c)^T(\mathbf{d}-\mathbf{d}_c) + \tfrac{1}{2}(\mathbf{d}-\mathbf{d}_c)^T\nabla^2_{\mathbf{d}} p(\mathbf{d}_c)(\mathbf{d}-\mathbf{d}_c) \leq \bar{p} \\ & \| \mathbf{d}-\mathbf{d}_c \|_\infty \leq \Delta^k \end{aligned} \tag{11.36}$$

For PMA trust-region surrogate-based RBDO, surrogate models of f and z are employed within a trust region Δ^k centered at $\mathbf{d}_c$. For first-order surrogates:

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{d}_c) + \nabla_d f(\mathbf{d}_c)^T(\mathbf{d}-\mathbf{d}_c) \\ \text{subject to} \quad & z(\mathbf{d}_c) + \nabla_d z(\mathbf{d}_c)^T(\mathbf{d}-\mathbf{d}_c) \geq \bar{z} \\ & \| \mathbf{d}-\mathbf{d}_c \|_\infty \leq \Delta^k \end{aligned} \tag{11.37}$$

and for second-order surrogates:

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{d}_c) + \nabla_{\mathbf{d}} f(\mathbf{d}_c)^T(\mathbf{d}-\mathbf{d}_c) + \tfrac{1}{2}(\mathbf{d}-\mathbf{d}_c)^T\nabla^2_{\mathbf{d}} f(\mathbf{d}_c)(\mathbf{d}-\mathbf{d}_c) \\ \text{subject to} \quad & z(\mathbf{d}_c) + \nabla_{\mathbf{d}} z(\mathbf{d}_c)^T(\mathbf{d}-\mathbf{d}_c) + \tfrac{1}{2}(\mathbf{d}-\mathbf{d}_c)^T\nabla^2_{\mathbf{d}} z(\mathbf{d}_c)(\mathbf{d}-\mathbf{d}_c) \geq \bar{z} \\ & \| \mathbf{d}-\mathbf{d}_c \|_\infty \leq \Delta^k \end{aligned} \tag{11.38}$$

where the sense of the z constraint may vary as described previously. The second-order information in Eqs. 11.36 and 11.38 will typically be approximated with quasi-Newton updates.

In `Dakota/test`, the `dakota_rbdo_cantilever_trsb.in` and `dakota_rbdo_short_column_trsb.in` input files solve the cantilever and short column OUU problems using a first-order sequential RBDO approach with analytic design gradients and first-order limit state approximations. The `dakota_rbdo_cantilever_trsb2.in`, `dakota_rbdo_short_column_trsb2.in`, and `dakota_rbdo_steel_column_trsb2.in` input files utilize second-order sequential RBDO approaches that employ second-order limit state approximations and integrations (from analytic limit state Hessians with respect to the uncertain variables) and quasi-Newton approximations to the reliability metric Hessians with respect to design variables.

Stochastic Expansion-Based Design Optimization

Section 6.4.8 describes sensitivity analysis of the polynomial chaos expansion with respect to the random variables. Here we extend this analysis to include sensitivity analysis with respect to design variables. With the introduction of design variables $\boldsymbol{s}$, a polynomial chaos expansion only over the random variables $\boldsymbol{\xi}$ has the functional relationship:

$$R(\boldsymbol{\xi}, \boldsymbol{s}) \cong \sum_{j=0}^{P} \alpha_j(\boldsymbol{s}) \Psi_j(\boldsymbol{\xi}) \tag{11.39}$$

In this case, design sensitivities for the mean and variance in Eqs. 6.63 and 6.64 are as follows:

$$\frac{d\mu_R}{ds} = \frac{d\alpha_0}{ds} = \frac{d}{ds}\langle R \rangle = \langle \frac{dR}{ds} \rangle \tag{11.40}$$

$$\frac{d\sigma_R^2}{ds} = \sum_{j=1}^{P} \langle \Psi_j^2 \rangle \frac{d\alpha_j^2}{ds} = 2 \sum_{j=1}^{P} \alpha_j \langle \frac{dR}{ds}, \Psi_j \rangle \tag{11.41}$$

since

$$\frac{d\alpha_j}{ds} = \frac{\langle \frac{dR}{ds}, \Psi_j \rangle}{\langle \Psi_j^2 \rangle} \tag{11.42}$$

The coefficients calculated in Eq. 11.42 may be interpreted as either the design sensitivities of the chaos coefficients for the response expansion or the chaos coefficients of an expansion for the response design sensitivities. The evaluation of integrals involving $\frac{dR}{ds}$ extends the data requirements for the PCE approach to include response sensitivities at each of the sampling points for the quadrature, sparse grid, sampling, or point collocation coefficient estimation approaches. The resulting expansions are valid only for a particular set of design variables and must be recalculated each time the design variables are modified.

Similarly for stochastic collocation,

$$R(\boldsymbol{\xi}, \boldsymbol{s}) \cong \sum_{j=1}^{N_p} r_j(\boldsymbol{s}) \boldsymbol{L}_j(\boldsymbol{\xi}) \tag{11.43}$$

leads to

$$\frac{d\mu_R}{ds} = \frac{d}{ds}\langle R \rangle = \sum_{j=1}^{N_p} \frac{dr_j}{ds} \langle \boldsymbol{L}_j \rangle = \sum_{j=1}^{N_p} w_j \frac{dr_j}{ds} \tag{11.44}$$

$$\frac{d\sigma_R^2}{ds} = \sum_{j=1}^{N_p} 2 w_j r_j \frac{dr_j}{ds} - 2\mu_R \frac{d\mu_R}{ds} = \sum_{j=1}^{N_p} 2 w_j (r_j - \mu_R) \frac{dr_j}{ds} \tag{11.45}$$

based on differentiation of Eqs. 6.65-6.66.

Alternatively, a stochastic expansion can be formed over both $\boldsymbol{\xi}$ and $\boldsymbol{s}$. Assuming a bounded design domain $\boldsymbol{s}_L \leq \boldsymbol{s} \leq \boldsymbol{s}_U$ (with no implied probability content), a Legendre chaos basis would be appropriate for each of the dimensions in $\boldsymbol{s}$ within a polynomial chaos expansion.

$$R(\boldsymbol{\xi}, \boldsymbol{s}) \cong \sum_{j=0}^{P} \alpha_j \Psi_j(\boldsymbol{\xi}, \boldsymbol{s}) \tag{11.46}$$

In this case, design sensitivities for the mean and variance do not require response sensitivity data, but this comes at the cost of forming the PCE over additional dimensions. For this combined variable expansion, the mean

and variance are evaluated by evaluating the expectations over only the random variables, which eliminates the polynomial dependence on $\boldsymbol{\xi}$, leaving behind the desired polynomial dependence on $\boldsymbol{s}$:

$$\mu_R(\boldsymbol{s}) = \sum_{j=0}^{P} \alpha_j \langle \Psi_j(\boldsymbol{\xi}, \boldsymbol{s}) \rangle_{\boldsymbol{\xi}} \tag{11.47}$$

$$\sigma_R^2(\boldsymbol{s}) = \sum_{j=0}^{P} \sum_{k=0}^{P} \alpha_j \alpha_k \langle \Psi_j(\boldsymbol{\xi}, \boldsymbol{s}) \Psi_k(\boldsymbol{\xi}, \boldsymbol{s}) \rangle_{\boldsymbol{\xi}} - \mu_R^2(\boldsymbol{s}) \tag{11.48}$$

The remaining polynomials may then be differentiated with respect to $\boldsymbol{s}$. In this approach, the combined PCE is valid for the full design variable range ($\boldsymbol{s}_L \leq \boldsymbol{s} \leq \boldsymbol{s}_U$) and does not need to be updated for each change in design variables, although adaptive localization techniques (i.e., trust region model management approaches) can be employed when improved local accuracy of the sensitivities is required.

Similarly for stochastic collocation,

$$R(\boldsymbol{\xi}, \boldsymbol{s}) \cong \sum_{j=1}^{N_p} r_j \boldsymbol{L}_j(\boldsymbol{\xi}, \boldsymbol{s}) \tag{11.49}$$

leads to

$$\mu_R(\boldsymbol{s}) = \sum_{j=1}^{N_p} r_j \langle \boldsymbol{L}_j(\boldsymbol{\xi}, \boldsymbol{s}) \rangle_{\boldsymbol{\xi}} \tag{11.50}$$

$$\sigma_R^2(\boldsymbol{s}) = \sum_{j=1}^{N_p} \sum_{k=1}^{N_p} r_j r_k \langle \boldsymbol{L}_j(\boldsymbol{\xi}, \boldsymbol{s}) \boldsymbol{L}_k(\boldsymbol{\xi}, \boldsymbol{s}) \rangle_{\boldsymbol{\xi}} - \mu_R^2(\boldsymbol{s}) \tag{11.51}$$

where the remaining polynomials not eliminated by the expectation over $\boldsymbol{\xi}$ are again differentiated with respect to $\boldsymbol{s}$.

Given the capability to compute analytic statistics of the response along with design sensitivities of these statistics, DAKOTA supports bi-level, sequential, and multifidelity approaches for optimization under uncertainty (OUU). The latter two approaches apply surrogate modeling approaches (data fits and multifidelity modeling) to the uncertainty analysis and then apply trust region model management to the optimization process.

The simplest and most direct approach is to employ these analytic statistics and their design derivatives directly within an optimization loop. This approach is known as bi-level OUU, since there is an inner level uncertainty analysis nested within an outer level optimization.

Consider the common reliability-based design example of a deterministic objective function with a reliability constraint:

$$\begin{array}{ll} \text{minimize} & f \\ \text{subject to} & \beta \geq \bar{\beta} \end{array} \tag{11.52}$$

where β is computed relative to a prescribed threshold response value $\bar{z}$ and is constrained by a prescribed reliability level $\bar{\beta}$. Another common example is robust design in which the constraint enforcing a reliability lower-bound has been replaced with a constraint enforcing a variance upper-bound:

$$\begin{array}{ll} \text{minimize} & f \\ \text{subject to} & \sigma^2 \leq \bar{\sigma}^2 \end{array} \tag{11.53}$$

Solving these problems using a bi-level approach involves computing $\beta(\boldsymbol{s})$ and $\frac{d\beta}{d\boldsymbol{s}}$ for Eq. 11.52 or σ^2 and $\frac{d\sigma^2}{d\boldsymbol{s}}$ for Eq. 11.53 for each set of design variables $\boldsymbol{s}$ passed from the optimizer. This approach is supported for both Uncertain and Combined expansions using PCE and SC.

An alternative OUU approach is the sequential approach, in which additional efficiency is sought through breaking the nested relationship of the UQ and optimization loops. The general concept is to iterate between optimization and uncertainty quantification, updating the optimization goals based on the most recent uncertainty assessment results. This approach is common with the reliability methods community, for which the updating strategy may be based on safety factors [140] or other approximations [26].

A particularly effective approach for updating the optimization goals is to use data fit surrogate models, and in particular, local Taylor series models allow direct insertion of stochastic sensitivity analysis capabilities. In Ref. [32], first-order Taylor series approximations were explored, and in Ref. [33], second-order Taylor series approximations are investigated. In both cases, a trust-region model management framework [34] is used to adaptively manage the extent of the approximations and ensure convergence of the OUU process. Surrogate models are used for both the objective and the constraint functions, although the use of surrogates is only required for the functions containing statistical results; deterministic functions may remain explicit is desired.

In particular, trust-region surrogate-based optimization for reliability-based design employs surrogate models of f and β within a trust region Δ^k centered at $\mathbf{s}_c$:

$$\begin{array}{rc} \text{minimize} & f(\mathbf{s}_c) + \nabla_s f(\mathbf{s}_c)^T(\mathbf{s}-\mathbf{s}_c) \\ \text{subject to} & \beta(\mathbf{s}_c) + \nabla_s \beta(\mathbf{s}_c)^T(\mathbf{s}-\mathbf{s}_c) \geq \bar{\beta} \\ & \| \mathbf{s}-\mathbf{s}_c \|_\infty \leq \Delta^k \end{array} \tag{11.54}$$

and trust-region surrogate-based optimization for robust design employs surrogate models of f and σ^2 within a trust region Δ^k centered at $\mathbf{s}_c$:

$$\begin{array}{rc} \text{minimize} & f(\mathbf{s}_c) + \nabla_s f(\mathbf{s}_c)^T(\mathbf{s}-\mathbf{s}_c) \\ \text{subject to} & \sigma^2(\mathbf{s}_c) + \nabla_s \sigma^2(\mathbf{s}_c)^T(\mathbf{s}-\mathbf{s}_c) \leq \bar{\sigma}^2 \\ & \| \mathbf{s}-\mathbf{s}_c \|_\infty \leq \Delta^k \end{array} \tag{11.55}$$

Second-order local surrogates may also be employed, where the Hessians are typically approximated with quasi-Newton updates. The sequential approach is available for Uncertain expansions using PCE and SC.

The multifidelity OUU approach is another trust-region surrogate-based approach. Instead of the surrogate UQ model being a simple data fit (in particular, first-/second-order Taylor series model) of the truth UQ model results, we now employ distinct UQ models of differing fidelity. This differing UQ fidelity could stem from the fidelity of the underlying simulation model, the fidelity of the UQ algorithm, or both. In this paper, we focus on the fidelity of the UQ algorithm. For reliability methods, this could entail varying fidelity in approximating assumptions (e.g., Mean Value for low fidelity, SORM for high fidelity), and for stochastic expansion methods, it could involve differences in selected levels of p and k refinement.

Here, we define UQ fidelity as point-wise accuracy in the design space and take the low fidelity model, whose validity over the design space will be adaptively controlled, to be the Combined expansion PCE/SC model, and the high fidelity truth model to be the Uncertain expansion PCE/SC model, with validity only at a single design point. This will allow us to take advantage of the design space spanning and lower cost of the Combined expansion approach to the extent possible, with fallback to the greater accuracy and higher expense of the Uncertain expansion approach when needed. The Combined expansion approach will span only the current trust region of the design space and will need to be reconstructed for each new trust region. The design derivatives of each model provide the necessary data to correct the low fidelity model to first-order consistency with the high fidelity model at the center of each trust region.

Multifidelity optimization for reliability-based design can be formulated as:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{s}) \\ \text{subject to} & \hat{\beta_{hi}}(\mathbf{s}) \geq \bar{\beta} \\ & \| \mathbf{s} - \mathbf{s}_c \|_\infty \leq \Delta^k \end{array} \tag{11.56}$$

and multifidelity optimization for robust design can be formulated as:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{s}) \\ \text{subject to} & \hat{\sigma_{hi}}^2(\mathbf{s}) \leq \bar{\sigma}^2 \\ & \| \mathbf{s} - \mathbf{s}_c \|_\infty \leq \Delta^k \end{array} \tag{11.57}$$

where the deterministic objective function is not approximated and $\hat{\beta_{hi}}$ and $\hat{\sigma_{hi}}^2$ are the approximated high-fidelity UQ results resulting from correction of the low-fidelity UQ results. In the case of an additive correction function:

$$\hat{\beta_{hi}}(\mathbf{s}) = \beta_{lo}(\mathbf{s}) + \alpha_\beta(\mathbf{s}) \tag{11.58}$$

$$\hat{\sigma_{hi}}^2(\mathbf{s}) = \sigma^2_{lo}(\mathbf{s}) + \alpha_{\sigma^2}(\mathbf{s}) \tag{11.59}$$

where correction functions $\alpha(\mathbf{s})$ enforcing first-order consistency [35] are typically employed.

In `Dakota/test`, the `dakota_pcbdo_cantilever.in`, `dakota_pcbdo_rosenbrock.in`, `dakota_pcbdo_short_column.in`, and `dakota_pcbdo_steel_column.in` input files solve cantilever (see Section 22.9), Rosenbrock, short column (see Section 22.8), and steel column (see Section 22.10) OUU problems using a bi-level polynomial chaos-based approach, where the statistical design metrics are reliability indices based on moment projection (i.e., Eqs. 6.3-6.4). The test matrix in the former three input files evaluate design gradients of these reliability indices using several different approaches: analytic design gradients based on a PCE formed over only over the random variables (Eqs. 11.40-11.41), analytic design gradients based on a PCE formed over all variables (differentiation of Eqs. 11.47-11.48), numerical design gradients based on a PCE formed only over the random variables, and numerical design gradients based on a PCE formed over all variables. In the cases where the expansion is formed over all variables, only a single PCE construction is required for the complete PCBDO process, whereas the expansions only over the random variables must be recomputed for each change in design variables. Sensitivities for "augmented" design variables (which are separate from and augment the random variables) may be handled using either analytic approach; however, sensitivities for "inserted" design variables (which define distribution parameters for the random variables) must be handled using Eqs. 11.40-11.41 where $\frac{dR}{ds}$ is calculated as $\frac{dR}{dx}\frac{dx}{ds}$. Additional test input files include:

- `dakota_scbdo_cantilever.in`, `dakota_scbdo_rosenbrock.in`, `dakota_scbdo_short_column.in`, and `dakota_scbdo_steel_column.in` input files solve cantilever, Rosenbrock, short column, and steel column OUU problems using a bi-level stochastic collocation-based approach.
- `dakota_pcbdo_cantilever_trsb.in`, `dakota_pcbdo_rosenbrock_trsb.in`, `dakota_pcbdo_short_column_trsb.in`, `dakota_pcbdo_steel_column_trsb.in`, `dakota_scbdo_cantilever_trsb.in`, `dakota_scbdo_rosenbrock_trsb.in`, `dakota_scbdo_short_column_trsb.in`, and `dakota_scbdo_steel_column_trsb.in` input files solve cantilever, Rosenbrock, short column, and steel column OUU problems using sequential polynomial chaos-based and stochastic collocation-based approaches.
- `dakota_pcbdo_cantilever_mf.in`, `dakota_pcbdo_rosenbrock_mf.in`, `dakota_pcbdo_short_column_mf.in`, `dakota_scbdo_cantilever_mf.in`, `dakota_scbdo_rosenbrock_mf.in`, and `dakota_scbdo_short_column_mf.in` input files solve cantilever, Rosenbrock, and short column OUU problems using multifidelity polynomial chaos-based and stochastic collocation-based approaches.

Epistemic OUU

An emerging capability is optimization under epistemic uncertainty. As described in the Nested Model section of the Reference Manual [30], epistemic and mixed aleatory/epistemic uncertainty quantification methods generate lower and upper interval bounds for all requested response, probability, reliability, and generalized reliability level mappings. Design for robustness in the presence of epistemic uncertainty could simply involve minimizing the range of these intervals (subtracting lower from upper using the nested model response mappings), and design for reliability in the presence of epistemic uncertainty could involve controlling the worst case upper or lower bound of the interval.

All facilities are in place to conduct these types of studies; however, no input file examples are provided at this time.

11.6.3 Surrogate-Based Uncertainty Quantification

Many uncertainty quantification (UQ) methods are computationally costly. For example, sampling often requires many function evaluations to obtain accurate estimates of moments or percentile values of an output distribution. One approach to overcome the computational cost of sampling is to evaluate the true function (e.g. run the analysis driver) on a fixed, small set of samples, use these sample evaluations to create a response surface approximation (e.g. a surrogate model or meta-model) of the underlying "true" function, then perform random sampling (using thousands or millions of samples) on the approximation to obtain estimates of the mean, variance, and percentiles of the response.

This approach, called "surrogate-based uncertainty quantification" is easy to do in DAKOTA, and one can set up input files to compare the results using no approximation (e.g. determine the mean, variance, and percentiles of the output directly based on the initial sample values) with the results obtained by sampling a variety of surrogate approximations. Example input files of a standard UQ analysis based on sampling alone vs. sampling a surrogate are shown in the `dakota_uq_sampling.in` and `dakota_surr_uq.in` in the `Dakota/examples/methods` directory.

Note that one must exercise some caution when using surrogate-based methods for uncertainty quantification. In general, there is not a single, straightforward approach to incorporate the error of the surrogate fit into the uncertainty estimates of the output produced by sampling the surrogate. Two references which discuss some of the related issues are [61] and [124]. The first reference shows that statistics of a response based on a surrogate model were less accurate, and sometimes biased, for surrogates constructed on very small sample sizes. In many cases, however, [61] shows that surrogate-based UQ performs well and sometimes generates more accurate estimates of statistical quantities on the output. The second reference goes into more detail about the interaction between sample type and response surface type (e.g., are some response surfaces more accurate when constructed on a particular sample type such as LHS vs. an orthogonal array?) In general, there is not a strong dependence of the surrogate performance with respect to sample type, but some sample types perform better with respect to some metrics and not others (for example, a Hammersley sample may do well at lowering root mean square error of the surrogate fit but perform poorly at lowering the maximum absolute deviation of the error). Much of this work is empirical and application dependent. If you choose to use surrogates in uncertainty quantification, we strongly recommend trying a variety of surrogates and examining diagnostic goodness-of-fit metrics.

Chapter 12

Variables

12.1 Overview

The `variables` specification in a DAKOTA input file specifies the parameter set to be iterated by a particular method. In the case of an optimization study, these variables are adjusted in order to locate an optimal design; in the case of parameter studies/sensitivity analysis/design of experiments, these parameters are perturbed to explore the parameter space; and in the case of uncertainty analysis, the variables are associated with distribution/interval characterizations which are used to compute corresponding distribution/interval characterizations for response functions. To accommodate these and other types of studies, DAKOTA supports design, uncertain, and state variable types for continuous and discrete variable domains.

This chapter will present a brief overview of the types of variables and their uses, as well as cover some user issues relating to integer/discrete conversions, file formats, and the active set vector. For a detailed description of variables section syntax and example specifications, refer to the Variables Commands chapter in the DAKOTA Reference Manual [30].

12.2 Design Variables

Design variables are those variables which are modified for the purposes of computing an optimal design. These variables may be continuous (real-valued) or discrete (integer-valued).

12.2.1 Continuous Design Variables

The most common type of design variables encountered in engineering applications are of the continuous type. These variables may assume any real value (e.g., `12.34`, `-1.735e+07`) within their bounds. All but a handful of the optimization algorithms in DAKOTA support continuous design variables exclusively.

12.2.2 Discrete Design Variables

Engineering design problems may contain discrete variables such as material types, feature counts, stock gauge selections, etc. These variables may assume only a fixed number of values within their bounds. While the general

discrete variable case would allow this fixed set of values to include real numbers (e.g., x_1 can only assume the values `4.2`, `6.4`, and `8.5`), DAKOTA assumes that the discrete variables can be specified as a sequence of integers (e.g., x_1 can be `1`, `2`, or `3`) and that a mapping from the integer sequence to the discrete values can be applied if necessary within the user's interface. A common mapping is to use the integer value from DAKOTA as the index into a vector of discrete real values.

Discrete variables may be classified as either "noncategorical" or "categorical" discrete variables. In the former noncategorical case, the integrality condition can be relaxed during the solution process since the model can still compute meaningful response functions for non-integer values. For example, a discrete variable representing the thickness of a structure is generally a noncategorical variable since it can assume a continuous range of values during the algorithm iterations, even if it is desired to have a stock gauge thickness in the end. In the latter categorical case, the integrality cannot be relaxed since the model cannot obtain a solution for a non-integer value. For example, feature counts are generally categorical variables, since most computational models will not support a non-integer value for the number of instances of some feature (e.g., number of support brackets).

Gradient-based optimization methods cannot be directly applied to problems with discrete variables. For problems with noncategorical variables, branch and bound techniques can be used to relax the integrality conditions and apply gradient-based methods to a series of generated subproblems. For problems with categorical variables, nongradient-based methods (e.g., `coliny_ea`) are commonly used. Branch and bound techniques are discussed in Section 10.5 and nongradient-based methods are further described in Chapter 7.

In addition to engineering applications, many non-engineering applications in the fields of scheduling, logistics, and resource allocation contain discrete design parameters. Within the Department of Energy, solution techniques for these problems impact programs in stockpile evaluation and management, production planning, nonproliferation, transportation (routing, packing, logistics), infrastructure analysis and design, energy production, environmental remediation, and tools for massively parallel computing such as domain decomposition and meshing.

12.3 Uncertain Variables

Deterministic variables (i.e., those with a single known value) do not capture the behavior of the input variables in all situations. In many cases, the exact value of a model parameter is not precisely known. An example of such an input variable is the thickness of a heat treatment coating on a structural steel I-beam used in building construction. Due to variabilities and tolerances in the coating process, the thickness of the layer is known to follow a normal distribution with a certain mean and standard deviation as determined from experimental data. The inclusion of the uncertainty in the coating thickness is essential to accurately represent the resulting uncertainty in the response of the building.

Currently, uncertain variables in DAKOTA are modeled as continuous random variables, or in the case of histogram, with an empirical histogram representation. If a problem contains discrete random variables, then these variables can be modeled using the point-based histogram representation. The following types of uncertain variables are available:

- Normal: a probability distribution characterized by a mean and standard deviation. Also referred to as Gaussian. Bounded normal is also supported by some methods with an additional specification of lower and upper bounds.

- Lognormal: a probability distribution characterized by a mean and either a standard deviation or an error factor. The natural logarithm of a lognormal variable has a normal distribution. Bounded lognormal is also supported by some methods with an additional specification of lower and upper bounds.

- Uniform: a probability distribution characterized by a lower bound and an upper bound. Probability is constant between the bounds.
- Loguniform: a probability distribution characterized by a lower bound and an upper bound. The natural logarithm of a loguniform variable has a uniform distribution.
- Triangular: a probability distribution characterized by a mode, a lower bound, and an upper bound.
- Exponential: a probability distribution characterized by a beta parameter.
- Beta: a flexible probability distribution characterized by a lower bound and an upper bound and alpha and beta parameters. The uniform distribution is a special case.
- Gamma: a flexible probability distribution characterized by alpha and beta parameters. The exponential distribution is a special case.
- Gumbel: the Type I Largest Extreme Value probability distribution. Characterized by alpha and beta parameters.
- Frechet: the Type II Largest Extreme Value probability distribution. Characterized by alpha and beta parameters.
- Weibull: the Type III Smallest Extreme Value probability distribution. Characterized by alpha and beta parameters.
- Histogram: an empirically-based probability distribution characterized by a set of (x, y) pairs that either map out histogram bins (a continuous interval with associated bin count) or histogram points (a discrete point value with associated count).
- Interval: an interval-based specification characterized by sets of lower and upper bounds and Basic Probability Assignments (BPAs) associated with each interval. This is not a probability distribution, as the exact structure of the probabilities within each interval is not known. It is commonly used with epistemic uncertainty methods.

DAKOTA also supports a user-supplied correlation matrix to provide correlations among the uncertain input variables. By default, the correlation matrix is set to the identity matrix, i.e., no correlation among the uncertain variables.

For additional information on random variable probability distributions, refer to [70] and [125]. Refer to the DAKOTA Reference Manual [30] for more detail on the uncertain variable specifications and to Chapter 6 for a description of methods available to quantify the uncertainty in the response.

12.4 State Variables

State variables consist of "other" variables which are to be mapped through the simulation interface, in that they are not to be used for design and they are not modeled as being uncertain. State variables provide a convenient mechanism for parameterizing additional model inputs which, in the case of a numerical simulator, might include solver convergence tolerances, time step controls, or mesh fidelity parameters. For additional model parameterizations involving strings (e.g., "mesh1.exo"), refer to the analysis components specification described in Section 12.6.1 and in the Interface Commands chapter of the DAKOTA Reference Manual [30]. Similar to the

design variables discussed in Section 12.2, state variables can be continuous (real-valued) or discrete (integer-valued). For discrete variables which are not a sequence of integers, a mapping can be applied between the integer and discrete values in the user's interface.

State variables, as with other types of variables, are viewed differently depending on the method in use. Since these variables are neither design nor uncertain variables, algorithms for optimization, least squares, and uncertainty quantification do not iterate on these variables; i.e., they are not active and are hidden from the algorithm. However, DAKOTA still maps these variables through the user's interface where they affect the computational model in use. This allows optimization, least squares, and uncertainty quantification studies to be executed under different simulation conditions (which will result, in general, in different results). Parameter studies and design of experiments methods, on the other hand, are general-purpose iterative techniques which do not draw a distinction between variable types. They include state variables in the set of variables to be iterated, which allows these studies to explore the effect of state variable values on the response data of interest.

In the future, state variables might be used in direct coordination with an optimization, least squares, or uncertainty quantification algorithm. For example, state variables could be used to enact model adaptivity through the use of a coarse mesh or loose solver tolerances in the initial stages of an optimization with continuous model refinement as the algorithm nears the optimal solution.

12.5 Mixed Variables

As alluded to in the previous section, the iterative method selected for use in DAKOTA determines what subset, or view, of the variables data is active in the iteration. The general case of having a mixture of various different types of variables is supported within all of the DAKOTA methods even though certain methods will only modify certain types of variables (e.g., optimizers and least squares methods only modify design variables, and uncertainty quantification methods, with the exception of `all_variables` mode, only utilize uncertain variables). This implies that variables which are not under the direct control of a particular iterator will be mapped through the interface in an unmodified state. This allows for a variety of parameterizations within the model in addition to those which are being used by a particular iterator, which can provide the convenience of consolidating the control over various modeling parameters in a single file (the DAKOTA input file). An important related point is that the variable set that is active with a particular iterator is the same variable set for which derivatives are typically computed (see Section 14.3).

12.6 DAKOTA Parameters File Data Format

Simulation interfaces which employ system calls and forks to create separate simulation processes must communicate with the simulation code through the file system. This is accomplished through the reading and writing of parameters and results files. DAKOTA uses a particular format for this data input/output. Depending on the user's interface specification, DAKOTA will write the parameters file in either standard or APREPRO format (future XML formats are planned). The former option uses a simple "`value tag`" format, whereas the latter option uses a "`{ tag = value }`" format for compatibility with the APREPRO utility [120] (as well as DPrePro, BPREPRO, and JPrePost variants).

```
<int> variables
<double> <var_tag_cdv_1>
...
<double> <var_tag_cdv_n>
<int> <var_tag_ddv_1>
...
<int> <var_tag_ddv_n>
<double> <var_tag_uv_1>
...
<double> <var_tag_uv_n>
<double> <var_tag_csv_1>
...
<double> <var_tag_csv_n>
<int> <var_tag_dsv_1>
...
<int> <var_tag_dsv_n>
<int> functions
<int> ASV_1
...
<int> ASV_m
<int> derivative_variables
<int> DVV_1
...
<int> DVV_p
<int> analysis_components
<string> AC_1
...
<string> AC_q
```

Figure 12.1: Parameters file data format - standard option.

12.6.1 Parameters file format (standard)

Prior to invoking a simulation, DAKOTA creates a parameters file which contains the current parameter values and a set of function requests. The standard format for this parameters file is shown in Figure 12.1.

where "`<int>`" denotes an integer value, "`<double>`" denotes a double precision value, "`<string>`" denotes a string value, and "`...`" indicates omitted lines for brevity. Each of the colored blocks (black for variables, blue for active set vector, red for derivative variables vector, and green for analysis components) denotes an array which begins with an array length and a descriptive tag. These array lengths are useful for dynamic memory allocation within a simulator or filter program.

The first array for variables begins with the total number of variables (`n`) with its identifier string "`variables`." The next `n` lines specify the current values and descriptors of all of the variables within the parameter set *in the following order*: continuous design, discrete design, uncertain, continuous state, and discrete state variables, where the uncertain variables break out using the following order: normal uncertain, lognormal uncertain, uniform uncertain, loguniform uncertain, triangular uncertain, exponential uncertain, beta uncertain, gamma uncertain, gumbel uncertain, frechet uncertain, weibull uncertain, histogram uncertain (bin histograms followed by point

histograms), and interval uncertain. This ordering is consistent with the specification order in dakota.input.txt. The lengths of these vectors add to a total of n (that is, $n_{nuv} + n_{lnuv} + n_{uuv} + n_{luuv} + n_{tuv} + n_{euv} + n_{buv} + n_{gauv} + n_{guuv} + n_{fuv} + n_{wuv} + n_{huv} + n_{iuv} = n_{uv}$ and $n_{cdv} + n_{ddv} + n_{uv} + n_{csv} + n_{dsv} = n$). If any of the variable types are not present in the problem, then its block is omitted entirely from the parameters file. The tags are the variable descriptors specified in the user's DAKOTA input file, or if no descriptors have been specified, default descriptors are used.

The second array for the active set vector (ASV) begins with the total number of functions (`m`) and its identifier string "`functions`." The next `m` lines specify the request vector for each of the `m` functions in the response data set followed by the tags "`ASV_i`." These integer codes indicate what data is required on the current function evaluation and are described further in Section 12.7.

The third array for the derivative variables vector (DVV) begins with the number of derivative variables (`p`) and its identifier string "`derivative_variables`." The next `p` lines specify integer variable identifiers followed by the tags "`DVV_i`." These integer identifiers are used to identify the subset of variables that are active for the calculation of derivatives (gradient vectors and Hessian matrices), and correspond to the list of variables in the first array (e.g., an identifier of 2 indicates that the second variable in the list is active for derivatives).

The final array for the analysis components (AC) begins with the number of analysis components (`q`) and its identifier string "`analysis_components`." The next `q` lines provide additional strings for use in specializing a simulation interface followed by the tags "`AC_i`." These strings are specified in a user's input file for a set of `analysis_drivers` using the `analysis_components` specification. The subset of the analysis components used for a particular analysis driver is the set passed in a particular parameters file.

Several standard-format parameters file examples are shown in Section 13.6.

12.6.2 Parameters file format (APREPRO)

For the APREPRO format option, the same data is present and the same ordering is used as in the standard format. The only difference is that values are associated with their tags within "`{ tag = value }`" constructs as shown in Figure 12.2. An APREPRO-format parameters file example is shown in Section 13.6.

The use of the APREPRO format option allows direct usage of these parameters files by the APREPRO utility, which is a file pre-processor that can significantly simplify model parameterization. Similar pre-processors include DPrePro, BPREPRO, and JPrePost. *[Note: APREPRO is a Sandia-developed pre-processor that is not currently distributed with DAKOTA. DPrePro is a Perl script distributed with DAKOTA that performs many of the same functions as APREPRO, and is optimized for use with DAKOTA parameters files in either format. BPREPRO and JPrePost are additional Perl and JAVA tools, respectively, in use at other sites.]* When a parameters file in APREPRO format is included within a template file (using an include directive), the APREPRO utility recognizes these constructs as variable definitions which can then be used to populate targets throughout the template file [120]. DPrePro, conversely, does not require the use of includes since it processes the DAKOTA parameters file and template simulation file separately to create a simulation input file populated with the variables data.

12.7 The Active Set Vector

The active set vector contains a set of integer codes, one per response function, which describe the data needed on a particular execution of an interface. Integer values of 0 through 7 denote a 3-bit binary representation of all possible combinations of value, gradient, and Hessian requests for a particular function, with the most significant bit denoting the Hessian, the middle bit denoting the gradient, and the least significant bit denoting the value. The specific translations are shown in Table 12.1.

```
{ DAKOTA_VARS = <int> }
{ <var_tag_cdv1> = <double> }
...
{ <var_tag_cdvn> = <double> }
{ <var_tag_ddv1> = <int> }
...
{ <var_tag_ddvn> = <int> }
{ <var_tag_uv1> = <double> }
...
{ <var_tag_uvn> = <double> }
{ <var_tag_csv1> = <double> }
...
{ <var_tag_csvn> = <double> }
{ <var_tag_dsv1> = <int> }
...
{ <var_tag_dsvn> = <int> }
{ DAKOTA_FNS = <int> }
{ ASV_1 = <int> }
...
{ ASV_m = <int> }
{ DAKOTA_DER_VARS = <int> }
{ DVV_1 = <int> }
...
{ DVV_p = <int> }
{ DAKOTA_AN_COMPS = <int> }
{ AC_1 = <int> }
...
{ AC_q = <int> }
```

Figure 12.2: Parameters file data format - APREPRO option.

Table 12.1: Active set vector integer codes.

Integer Code	Binary representation	Meaning
7	111	Get Hessian, gradient, and value
6	110	Get Hessian and gradient
5	101	Get Hessian and value
4	100	Get Hessian
3	011	Get gradient and value
2	010	Get gradient
1	001	Get value
0	000	No data required, function is inactive

The active set vector in DAKOTA gets its name from managing the active set, i.e., the set of functions that are active on a particular function evaluation. However, it also manages the type of data that is needed for functions that are active, and in that sense, has an extended meaning beyond that typically used in the optimization literature.

12.7.1 Active set vector control

Active set vector control may be turned off to allow the user to simplify the supplied interface by removing the need to check the content of the active set vector on each evaluation. The Interface Commands chapter in the DAKOTA Reference Manual [30] provides additional information on this option (`deactivate active_set_vector`). Of course, this option trades some efficiency for simplicity and is most appropriate for those cases in which only a relatively small penalty occurs when computing and returning more data than may be needed on a particular function evaluation.

Chapter 13

Interfaces

13.1 Overview

The `interface` specification in a DAKOTA input file controls how function evaluations will be performed. The mechanisms currently in place for performing function evaluations involve interfacing with one or more computational simulation codes, computing algebraic mappings, or a combination of the two.

This chapter will describe algebraic mappings in Section 13.2, followed by discussion of a variety of mechanisms for simulation code invocation in Section 13.3. It also provides an overview of simulation interface components, covers issues relating to file management and presents a number of example data mappings.

For a detailed description of interface specification syntax, refer to the interface commands chapter in the DAKOTA Reference Manual [30].

13.2 Algebraic Mappings

If desired, one can define algebraic input-output mappings using the AMPL code [46] and save these mappings in 3 files: `stub.nl`, `stub.col`, and `stub.row`, where `stub` is a particular root name describing a particular problem. These files names can be communicated to DAKOTA using the `algebraic_mappings` input.

DAKOTA will employ `stub.col` and `stub.row` to extract the input and output identifier strings, respectively, and employs the AMPL solver library [51] to process a directed acyclic graph (DAG) specification in `stub.nl`.

As a simple example (from `Dakota/test/ampl/fma`), consider simple algebraic mappings based on Newton's law $F = ma$. The following file is the AMPL model file showing the variable declarations and output metric definitions:

```
var mass;
var a;
var v;
minimize force: mass*a;
minimize energy: 0.5 * mass * v^2;
option auxfiles rc;     # generate stub.row and stub.col
```

When processed by an AMPL executable, three files are created (as requested by the auxfiles command). The first is the `fma.nl` file containing problem statistics, expression graphs, bounds, etc.:

```
g3 0 1 0        # problem fma
 3 0 2 0 0      # vars, constraints, objectives, ranges, eqns
 0 2    # nonlinear constraints, objectives
 0 0    # network constraints: nonlinear, linear
 0 3 0  # nonlinear vars in constraints, objectives, both
 0 0 0 1        # linear network variables; functions; arith, flags
 0 0 0 0 0      # discrete variables: binary, integer, nonlinear (b,c,o)
 0 4    # nonzeros in Jacobian, gradients
 6 4    # max name lengths: constraints, variables
 0 0 0 0 0      # common exprs: b,c,o,c1,o1
O0 0
o2
v0
v1
O1 0
o2
o2
n0.5
v0
o5
v2
n2
b
3
3
3
k2
0
0
G0 2
0 0
1 0
G1 2
0 0
2 0
```

Next, the `fma.col` file contains the set of variable descriptor strings:

```
mass
a
v
```

and the `fma.row` file contains the set of response descriptor strings:

```
force
energy
```

The variable and objective function names declared within AMPL should be a subset of the variable descriptors and response descriptors used by DAKOTA (see the DAKOTA Reference Manual [30] for information on DAKOTA variable and response descriptors). Ordering of the inputs and outputs within the AMPL declaration is not important, as DAKOTA will reorder data as needed. The following listing shows an excerpt from `Dakota/test/dakota_ampl.in`, which demonstrates a combined algebraic/simulation-based mapping in which algebraic mappings from the `fma` definition are overlaid with simulation-based mappings from `text_book`:

```
variables,
        continuous_design = 5
          descriptor    'x1' 'mass' 'a' 'x4' 'v'
          initial_point  0.0  2.0  1.0  0.0  3.0
          lower_bounds  -3.0  0.0 -5.0 -3.0 -5.0
          upper_bounds   3.0 10.0  5.0  3.0  5.0

interface,
        algebraic_mappings = 'ampl/fma.nl'
        system
          analysis_driver = 'text_book'
          parameters_file = 'tb.in'
          results_file    = 'tb.out'
          file_tag

responses,
        response_descriptors = 'force' 'ineq1' 'energy'
        num_objective_functions = 1
        num_nonlinear_inequality_constraints = 1
        num_nonlinear_equality_constraints = 1
        nonlinear_equality_targets = 20.0
        analytic_gradients
        no_hessians
```

Note that the algebraic inputs and outputs are a subset of the total inputs and outputs and that DAKOTA will track the algebraic contributions to the total response set using the order of the descriptor strings. In the case where both the algebraic and simulation-based components contribute to the same function, they are overlaid using a simple summation.

To solve `text_book` algebraically (refer to Section 2.2 for definition), the following AMPL model file could be used

```
#   Problem : Textbook problem used in DAKOTA testing
#             Constrained quartic, 2 continuous variables
#   Solution: x=(0.5, 0.5), obj = .125, c1 = 0, c2 = 0
#

# continuous variables
var x1 >=  0.5 <= 5.8 := 0.9;
var x2 >= -2.9 <= 2.9 := 1.1;

# objective function
minimize obj: (x1 - 1)^4 + (x2 - 1)^4;

# constraints (current required syntax for DAKOTA/AMPL interface)
minimize c1: x1^2 - 0.5*x2;
minimize c2: x2^2 - 0.5*x1;

# required for output of *.row and *.col files
option auxfiles rc;
```

Note that the nonlinear constraints should not currently be declared as constraints within AMPL. Since the DAKOTA variable bounds and constraint bounds/targets currently take precedence over any AMPL specification, the current approach is to declare all AMPL outputs as objective functions and then map them into the appropriate response function type (objectives, least squares terms, nonlinear inequality/equality constraints, or generic response functions) within the DAKOTA input specification.

13.3 Simulation Interfaces

The invocation of a simulation code is performed using either system calls, forks, or direct function invocations. In the system call and fork cases, a separate process is created for the simulation and communication between DAKOTA and the simulation occurs through parameter and response files. For system call and fork interfaces, then, the interface section must also specify the details of this data transfer. In the direct function case, a separate process is not created and communication occurs directly through the function parameter list. Sections 13.3.1 through 13.3.5 provide information on the simulation interfacing approaches.

13.3.1 The Direct Function Simulation Interface

The direct function interface may be used to invoke simulations that are linked into the DAKOTA executable. This interface eliminates overhead from process creation and file I/O and can simplify operations on massively parallel computers. These advantages are balanced with the practicality of converting an existing simulation code into a library with a subroutine interface. Sandia codes for structural dynamics (Salinas), computational fluid dynamics (Sage), and circuit simulation (Xyce) and external codes such as Phoenix Integration's ModelCenter framework and The Mathworks' Matlab have been linked in this way, and a direct interface to Sandia's SIERRA multiphysics framework is under development. In the latter case, the additional effort is particularly justified since SIERRA unifies an entire suite of physics codes. [*Note: the "sandwich implementation" of combining a direct interface plug-in with DAKOTA's library mode is discussed in the DAKOTA Developers Manual [31]*].

In addition to direct linking with simulation codes, the direct interface also provides access to internal polynomial test functions that are used for algorithm performance and regression testing. The following test functions

are available: `cantilever`, `cyl_head`, `log_ratio`, `rosenbrock`, `short_column`, and `text_book` (including `text_book1`, `text_book2`, `text_book3`, and `text_book_ouu`). While these functions are also available as external programs in the `Dakota/test` directory, maintaining internally linked versions allows more rapid testing. See Chapter 22 for additional information on several of these test problems. An example input specification for a direct interface follows:

```
interface,
        direct
          analysis_driver = 'rosenbrock'
```

Additional specification examples are provided in Section 2.4 and additional information on asynchronous usage of the direct function interface is provided in Section 18.2.1. Guidance for usage of some particular direct simulation interfaces is in Section 17.3 and the details of adding a simulation code to the direct interface are provided in Section 17.2.

13.3.2 The System Call Simulation Interface

The system call approach invokes a simulation code or simulation driver by using the `system` function from the standard C library [83]. In this approach, the system call creates a new process which communicates with DAKOTA through parameter and response files. The system call approach allows the simulation to be initiated via its standard invocation procedure (as a "black box") and then coordinated with any variety of tools for pre- and post-processing. This approach has been widely used in previous studies [38, 40, 29]. The system call approach involves more process creation and file I/O overhead than the direct function approach; however, this is most often of very little significance relative to the expense of the simulations. An example of a system call interface specification follows:

```
interface,
        system
          analysis_driver = 'text_book'
          parameters_file = 'text_book.in'
          results_file    = 'text_book.out'
          file_tag file_save
```

More detailed examples of using the system call interface are provided in Section 2.4.10 and in Section 17.1, and information on asynchronous usage of the system call interface is provided in Section 18.2.1.

13.3.3 The Fork Simulation Interface

The fork simulation interface uses the `fork`, `exec`, and `wait` families of functions to manage simulation codes or simulation drivers. Calls to `fork` or `vfork` create a copy of the DAKOTA process, `execvp` replaces this copy with the simulation code or driver process, and then DAKOTA uses the `wait` or `waitpid` functions to wait for completion of the new process. Transfer of variables and response data between DAKOTA and the simulator code or driver occurs through the file system in exactly the same manner as for the system call interface. An example of a fork interface specification follows:

```
interface,
        fork
          input_filter    = 'test_3pc_if'
          output_filter   = 'test_3pc_of'
          analysis_driver = 'test_3pc_ac'
```

```
parameters_file = 'tb.in'
results_file    = 'tb.out'
file_tag
```

Information on asynchronous usage of the fork interface is provided in Section 18.2.1.

13.3.4 Syntax for Filter and Driver Strings

With the fork interface, and on most systems, with the system interface as well, the string values supplied for `input_filter`, `output_fileter`, and `analysis_driver` can involve simple Bourne-shell syntax for specifying environment values that the filter or driver will see. For example,

```
analysis_driver = 'opfile=myspec outlev=2 mydriver'
```

would cause `mydriver` to be invoked with environment variables `opfile` and `outlev` having the values "myspec" and "2", respectively. If the driver is a shell script, it can access these values as `$opfile` and `$outlev`; a compiled driver can obtain these values from a function; drivers written in C or C++ can use the standard `getenv` function (e.g., invoking `getenv("opfile")`).

Both the values assigned to environment variables and name of the file invoked as filter or driver can contain spaces, provided that the values in question are quoted. Within strings delimited by single quotes, you can use double quotes for quoting, and vice versa. For instance,

```
analysis_driver = 'opfile="my spec" "my driver"'
```

and

```
analysis_driver = "opfile='my spec' 'my driver'"
```

both specify a driver named "`my driver`" and value "`my spec`" for `$opfile`.

13.3.5 Fork or System Call: Which to Use?

The primary operational difference between the fork and system call simulation interfaces is that, in the fork interface, the `fork`/`exec` functions return a UNIX process identifier that the `wait`/`waitpid` functions can use to detect the completion of a simulation for either synchronous or asynchronous operations. The system call simulation interface, on the other hand, must use a response file detection scheme for this purpose in the asynchronous case. Thus, an important advantage of the fork interface over the system call interface is that it avoids the potential of a file race condition when employing asynchronous local parallelism (refer to Section 18.2.1). This condition can occur when the responses file has been created but the writing of the response data set to this file has not been completed (see Section 18.2.1). While significant care has been taken to manage this file race condition in the system call case, the fork interface still has the potential to be more robust when performing function evaluations asynchronously.

Another advantage of the fork interface is that it has additional asynchronous capabilities when a function evaluation involves multiple analyses. As shown in Table 18.1, the fork interface supports asynchronous local and hybrid parallelism modes for managing concurrent analyses within function evaluations, whereas the system call interface does not. These additional capabilities again stem from the ability to track child processes by their UNIX process identifiers.

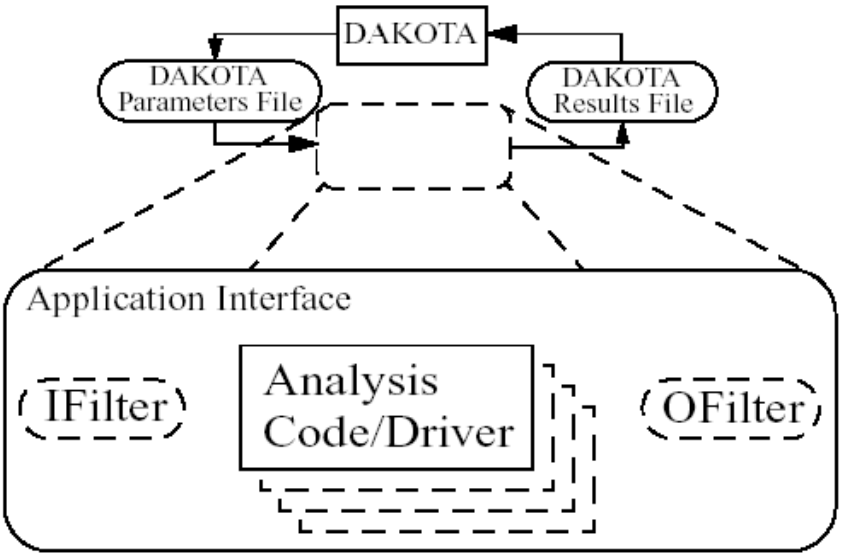


Figure 13.1: Components of the simulation interface

The only disadvantage to the fork interface compared with the system interface is that the `fork`/`exec`/`wait` functions are not part of the standard C library, whereas the `system` function is. As a result, support for implementations of the `fork`/`exec`/`wait` functions can vary from platform to platform. At one time, these commands were not available on some of Sandia's massively parallel computers. However, in the more mainstream UNIX environments, availability of `fork`/`exec`/`wait` should not be an issue.

In summary, the system call interface has been a workhorse for many years and is well tested and proven, but the fork interface supports additional capabilities and is recommended when managing asynchronous simulation code executions. Having both interfaces available has proven to be useful on a number of occasions and they will both continue to be supported for the foreseeable future.

13.4 Simulation Interface Components

Figure 13.1 is an extension of Figure 1.1 that adds details of the components that make up each of the simulation interfaces (system call, fork, and direct). These components include an `input_filter` ("IFilter"), one or more `analysis_drivers` ("Analysis Code/Driver"), and an `output_filter` ("OFilter"). The input and output filters provide optional facilities for managing simulation pre- and post-processing, respectively. More specifically, the input filter can be used to insert the DAKOTA parameters into the input files required by the simulator program, and the output filter can be used to recover the raw data from the simulation results and compute the desired response data set. If there is a single analysis code, it is often convenient to combine these pre- and post-processing functions into a single simulation driver script, and the separate input and output filter facilities are rarely used in this case. If there are multiple analysis drivers, however, the input and output filter facilities provide a convenient means for managing *nonrepeated* portions of the pre- and post-processing for multiple analyses. That is, pre- and post-processing tasks that must be performed for each analysis can be performed within the individual analysis drivers, and shared pre- and post-processing tasks that are only performed once for the set of analyses can be performed within the input and output filters.

When spawning function evaluations using system calls or forks, DAKOTA must communicate parameter and response data with the analysis drivers and filters through use of the file system. This is accomplished by passing the names of the parameters and results files on the command line when executing an analysis driver or filter. The input filter or analysis driver read data from the parameters file and the output filter or analysis driver write the appropriate data to the responses file. While not essential when the file names are fixed, the file names must be retrieved from the command line when DAKOTA is changing the file names from one function evaluation to the next (i.e., using UNIX temporary files or root names tagged with numerical identifiers). In the case of a UNIX C-shell script, the two command line arguments are retrieved using `$argv[1]` and `$argv[2]` (see [4]). Similarly, Bourne shell scripts retrieve the two command line arguments using `$1` and `$2`, and Perl scripts retrieve the two command line arguments using `@ARGV[0]` and `@ARGV[1]`. In the case of a C or C++ program, command line arguments are retrieved using `argc` (argument count) and `argv` (argument vector) [83], and for Fortran 77, the `iargc` function returns the argument count and the `getarg` subroutine returns command line arguments.

13.4.1 Single analysis driver without filters

If a single `analysis_driver` is selected in the interface specification and filters are not needed (as indicated by omission of the `input_filter` and `output_filter` specifications), then only one process will appear in the execution syntax of the simulation interface. An example of this syntax in the system call case is:

```
(driver params.in results.out)
```

where "`driver`" is the user-specified analysis driver and "`params.in`" and "`results.out`" are the names of the parameters and results files, respectively, passed on the command line. In this case, the user need not retrieve the command line arguments since the same file names will be employed each time.

For the same mapping, the fork simulation interface echoes the following syntax:

```
blocking fork: driver params.in results.out
```

for which only a single blocking fork is needed to perform the evaluation.

Executing the same mapping with the direct simulation interface results in an echo of the following syntax:

```
Direct function: invoking driver
```

where this analysis driver must be linked as a function within DAKOTA's direct interface (see Section 17.2). Note that no files are involved for communication of parameter and response data, since this data is passed directly through the function parameter lists.

Both the system call and fork interfaces support asynchronous operations. The asynchronous system call execution syntax involves executing the system call in the background:

```
(driver params.in.1 results.out.1) &
```

and the asynchronous fork execution syntax involves use of a nonblocking fork:

```
nonblocking fork: driver params.in.1 results.out.1
```

where file tagging (see Section 13.5.2) has been user-specified in both cases to prevent conflicts between concurrent analysis drivers. In these cases, the user must retrieve the command line arguments since the file names change on each evaluation. Execution of the direct interface must currently be performed synchronously since multithreading is not yet supported (see Section 18.2.1).

13.4.2 Single analysis driver with filters

When filters are used, the syntax of the system call that DAKOTA performs is:

```
(ifilter params.in results.out; driver params.in results.out;
     ofilter params.in results.out)
```

in which the input filter ("`ifilter`"), analysis driver ("`driver`"), and output filter ("`ofilter`") processes are combined into a single system call through the use of semi-colons and parentheses (see [4]). All three portions are passed the names of the parameters and results files on the command line.

For the same mapping, the fork simulation interface echoes the following syntax:

```
blocking fork: ifilter params.in results.out;
     driver params.in results.out; ofilter params.in results.out
```

where a series of three blocking forks is used to perform the evaluation.

Executing the same mapping with the direct simulation interface results in an echo of the following syntax:

```
Direct function: invoking { ifilter driver ofilter }
```

where each of the three components must be linked as a function within DAKOTA's direct interface. Since asynchronous operations are not yet supported, execution simply involves invocation of each of the three linked functions in succession. Again, no files are involved since parameter and response data are passed directly through the function parameter lists.

Asynchronous executions would appear as follows for the system call interface:

```
(ifilter params.in.1 results.out.1; driver params.in.1 results.out.1;
     ofilter params.in.1 results.out.1) &
```

and, for the fork interface, as:

```
nonblocking fork: ifilter params.in.1 results.out.1;
     driver params.in.1 results.out.1; ofilter params.in.1 results.out.1
```

where file tagging of evaluations has again been user-specified in both cases. For the system call simulation interface, use of parentheses and semi-colons to bind the three processes into a single system call simplifies asynchronous process management compared to an approach using separate system calls. The fork simulation interface, on the other hand, does not rely on parentheses and accomplishes asynchronous operations by first forking an intermediate process. This intermediate process is then reforked for the execution of the input filter, analysis driver, and output filter. The intermediate process can be blocking or nonblocking (nonblocking in this case), and the second level of forks can be blocking or nonblocking (blocking in this case). The fact that forks can be reforked multiple times using either blocking or nonblocking approaches provides the enhanced flexibility to support a variety of local parallelism approaches (see Chapter 18).

13.4.3 Multiple analysis drivers without filters

If a list of `analysis_drivers` is specified and filters are not needed (i.e., neither `input_filter` nor `output_filter` appears), then the system call syntax would appear as:

```
(driver1 params.in results.out.1; driver2 params.in results.out.2;
     driver3 params.in results.out.3)
```

where "`driver1`", "`driver2`", and "`driver3`" are the user-specified analysis drivers and "`params.in`" and "`results.out`" are the user-selected names of the parameters and results files. Note that the results files for the different analysis drivers have been automatically tagged to prevent overwriting. This automatic tagging of *analyses* (see Section 13.5.4) is a separate operation from user-selected tagging of *evaluations* (see Section 13.5.2).

For the same mapping, the fork simulation interface echoes the following syntax:

```
blocking fork: driver1 params.in results.out.1;
     driver2 params.in results.out.2; driver3 params.in results.out.3
```

for which a series of three blocking forks is needed (no reforking of an intermediate process is required).

Executing the same mapping with the direct simulation interface results in an echo of the following syntax:

```
Direct function: invoking { driver1 driver2 driver3 }
```

where, again, each of these components must be linked within DAKOTA's direct interface and no files are involved for parameter and response data transfer.

Both the system call and fork interfaces support asynchronous function evaluations. The asynchronous system call execution syntax would be reported as

```
(driver1 params.in.1 results.out.1.1; driver2 params.in.1 results.out.1.2;
     driver3 params.in.1 results.out.1.3) &
```

and the nonblocking fork execution syntax would be reported as

```
nonblocking fork: driver1 params.in.1 results.out.1.1;
     driver2 params.in.1 results.out.1.2; driver3 params.in.1 results.out.1.3
```

where, in both cases, file tagging of evaluations has been user-specified to prevent conflicts between concurrent analysis drivers and file tagging of the results files for multiple analyses is automatically used. In the fork interface case, an intermediate process is forked to allow a non-blocking function evaluation, and this intermediate process is then reforked for the execution of each of the analysis drivers.

13.4.4 Multiple analysis drivers with filters

Finally, when combining filters with multiple `analysis_drivers`, the syntax of the system call that DAKOTA performs is:

```
(ifilter params.in.1 results.out.1;
     driver1 params.in.1 results.out.1.1;
     driver2 params.in.1 results.out.1.2;
     driver3 params.in.1 results.out.1.3;
     ofilter params.in.1 results.out.1)
```

in which all processes have again been combined into a single system call through the use of semi-colons and parentheses. Note that the secondary file tagging for the results files is only used for the analysis drivers and

not for the filters. This is consistent with the filters' defined purpose of managing the non-repeated portions of analysis pre- and post-processing (e.g., overlay of response results from individual analyses; see Section 13.5.4 for additional information).

For the same mapping, the fork simulation interface echoes the following syntax:

```
blocking fork: ifilter params.in.1 results.out.1;
     driver1 params.in.1 results.out.1.1;
     driver2 params.in.1 results.out.1.2;
     driver3 params.in.1 results.out.1.3;
     ofilter params.in.1 results.out.1
```

for which a series of five blocking forks is used (no reforking of an intermediate process is required).

Executing the same mapping with the direct simulation interface results in an echo of the following syntax:

```
Direct function: invoking { ifilter driver1 driver2 driver3 ofilter }
```

where each of these components must be linked as a function within DAKOTA's direct interface. Since asynchronous operations are not supported, execution simply involves invocation of each of the five linked functions in succession. Again, no files are involved for parameter and response data transfer since this data is passed directly through the function parameter lists.

Asynchronous executions would appear as follows for the system call interface:

```
(ifilter params.in.1 results.out.1;
     driver1 params.in.1 results.out.1.1;
     driver2 params.in.1 results.out.1.2;
     driver3 params.in.1 results.out.1.3;
     ofilter params.in.1 results.out.1) &
```

and for the fork interface:

```
nonblocking fork: ifilter params.in.1 results.out.1;
     driver1 params.in.1 results.out.1.1;
     driver2 params.in.1 results.out.1.2;
     driver3 params.in.1 results.out.1.3;
     ofilter params.in.1 results.out.1
```

where, again, user-selected file tagging of evaluations is combined with automatic file tagging of analyses. In the fork interface case, an intermediate process is forked to allow a non-blocking function evaluation, and this intermediate process is then reforked for the execution of the input filter, each of the analysis drivers, and the output filter.

13.5 Simulation File Management

This section describes some of the file management features that are employed during an execution of DAKOTA when file transfer of data is used for the communication between DAKOTA and the simulation code (i.e., when the system call or fork interfaces are used). These features can be used for generating unique filenames when utilizing DAKOTA's parallel execution capabilities and for debugging purposes when troubleshooting the interface between DAKOTA and the simulation code.

13.5.1 File Saving

The `file_save` option in the interface specification allows the user to control whether parameters and results files are retained or removed from the working directory. DAKOTA's default behavior is to remove files once their use is complete in order to reduce clutter. If the method output setting is verbose, a file remove notification will follow the function evaluation echo, e.g.,

```
(driver /usr/tmp/aaaa20305 /usr/tmp/baaa20305)
Removing /usr/tmp/aaaa20305 and /usr/tmp/baaa20305
```

However, if `file_save` appears in the interface specification, these files will not be removed. This latter behavior is often useful for debugging communication between DAKOTA and simulator programs. An example of a `file_save` specification is shown in the file tagging example below.

13.5.2 File Tagging for Evaluations

When a user provides `parameters_file` and `results_file` specifications, the `file_tag` option in the interface specification allows the user to render the names of these parameters and results files unique by appending the function evaluation number to the root file names. Default behavior is to not tag these files, which has the advantage of allowing the user to ignore command line argument passing and always read to and write from the same file names. However, it has the disadvantage that files may be overwritten from one function evaluation to the next. When `file_tag` appears in the interface specification, the file names become unique through the appended evaluation number. This uniqueness makes it necessary for the user's interface to retrieve the names of these files from the command line. The file tagging feature is most often used when concurrent simulations are running in a common disk space, since it can prevent conflicts between the simulations. An example specification of `file_tag` and `file_save` is shown below:

```
interface,
        system
          analysis_driver = 'text_book'
          parameters_file = 'text_book.in'
          results_file    = 'text_book.out'
          file_tag file_save
```

Special case: When a user specifies names for the parameters and results files and `file_save` is used without `file_tag`, untagged files are used in the function evaluation but are then moved to tagged files after the function evaluation is complete, to prevent overwriting files for which a `file_save` request has been given. If the output control is set to verbose, then a notification similar to the following will follow the function evaluation echo:

```
(driver params.in results.out)
Files with nonunique names will be tagged to enable file_save:
Moving params.in to params.in.1
Moving results.out to results.out.1
```

13.5.3 UNIX Temporary Files

If `parameters_file` and `results_file` are not specified by the user, then the default mechanisms for file communication are UNIX temporary files. For example, a system call to a single analysis driver would appear as:

```
(driver /usr/tmp/aaaa20305 /usr/tmp/baaa20305)
```

and a system call to an analysis driver with filter programs would appear as:

```
(ifilter /usr/tmp/aaaa22490 usr/tmp/baaa22490;
     driver /usr/tmp/aaaa22490 usr/tmp/baaa22490;
     ofilter /usr/tmp/aaaa22490 /usr/tmp/baaa22490)
```

These files have unique names as created by the `tmpnam` utility from the C standard library [83]. This uniqueness makes it a requirement for the user's interface to retrieve the names of these files from the command line. File tagging with evaluation number is unnecessary with UNIX temporary files (since they are already unique); thus, `file_tag` requests will be ignored. A `file_save` request will be honored, but it should be used with care since the temporary file directory could easily become cluttered without the user noticing.

13.5.4 File Tagging for Analysis Drivers

When multiple analysis drivers are involved in performing a function evaluation with either the system call or fork simulation interface, a secondary file tagging is *automatically* used to distinguish the results files used for the individual analyses. This applies to both the case of user-specified names for the parameters and results files and the default UNIX temporary file case. Examples for the former case were shown previously in Section 13.4.3 and Section 13.4.4. The following examples demonstrate the latter UNIX temporary file case. Even though Unix temporary files have unique names for a particular function evaluation, tagging is still needed to manage the individual contributions of the different analysis drivers to the response results, since the same root results filename is used for each component. For the system call interface, the syntax would be similar to the following:

```
(ifilter /var/tmp/aaawkaOKZ /var/tmp/baaxkaOKZ;
     driver1 /var/tmp/aaawkaOKZ /var/tmp/baaxkaOKZ.1;
     driver2 /var/tmp/aaawkaOKZ /var/tmp/baaxkaOKZ.2;
     driver3 /var/tmp/aaawkaOKZ /var/tmp/baaxkaOKZ.3;
     ofilter /var/tmp/aaawkaOKZ /var/tmp/baaxkaOKZ)
```

and, for the fork interface, similar to:

```
blocking fork:
     ifilter /var/tmp/aaawkaOKZ /var/tmp/baaxkaOKZ;
     driver1 /var/tmp/aaawkaOKZ /var/tmp/baaxkaOKZ.1;
     driver2 /var/tmp/aaawkaOKZ /var/tmp/baaxkaOKZ.2;
     driver3 /var/tmp/aaawkaOKZ /var/tmp/baaxkaOKZ.3;
     ofilter /var/tmp/aaawkaOKZ /var/tmp/baaxkaOKZ
```

Tagging of results files with an analysis identifier is needed since each analysis driver must contribute a user-defined subset of the total response results for the evaluation. If an output filter is not supplied, DAKOTA will combine these portions through a simple overlaying of the individual contributions (i.e., summing the results in `/var/tmp/baaxkaOKZ.1`, `/var/tmp/baaxkaOKZ.2`, and `/var/tmp/baaxkaOKZ.3`). If this simple approach is inadequate, then an output filter should be supplied to perform the combination. This is the reason why the results file for the output filter does not use analysis tagging; it is responsible for the results combination (i.e., combining `/var/tmp/baaxkaOKZ.1`, `/var/tmp/baaxkaOKZ.2`, and `/var/tmp/baaxkaOKZ.3` into `/var/tmp/baaxkaOKZ`). In this case, DAKOTA will read only the results file from the output filter (i.e., `/var/tmp/baaxkaOKZ`) and interpret it as the total response set for the evaluation.

Parameters files are not currently tagged with an analysis identifier. This reflects the fact that DAKOTA does not attempt to subdivide the requests in the active set vector for different analysis portions. Rather, the total active set vector is passed to each analysis driver and the appropriate subdivision of work *must be defined by the user*. This

allows the division of labor to be very flexible. In some cases, this division might occur across response functions, with different analysis drivers managing the data requests for different response functions. And in other cases, the subdivision might occur within response functions, with different analysis drivers contributing portions to each of the response functions. The only restriction is that each of the analysis drivers must follow the response format dictated by the total active set vector. For response data for which an analysis driver has no contribution, 0's must be used as placeholders.

13.6 Parameter to Response Mappings

In this section, interface mapping examples are presented through the discussion of several parameters files and their corresponding results files. A typical input file for 2 variables ($n = 2$) and 3 functions ($m = 3$) using the standard parameters file format (see Section 12.6.1) is as follows:

```
                     2 variables
1.500000000000000e+00 cdv_1
1.500000000000000e+00 cdv_2
                     3 functions
                     1 ASV_1
                     1 ASV_2
                     1 ASV_3
                     2 derivative_variables
                     1 DVV_1
                     2 DVV_2
                     0 analysis_components
```

where numerical values are associated with their tags within "`value tag`" constructs. The number of design variables (n) and the string "`variables`" are followed by the values of the design variables and their tags, the number of functions (m) and the string "`functions`", the active set vector (ASV) and its tags, the number of derivative variables and the string "`derivative_variables`", the derivative variables vector (DVV) and its tags, the number of analysis components and the string "`analysis_components`", and the analysis components array and its tags. The descriptive tags for the variables are always present and they are either the descriptors specified in the user's variables specification or are default descriptors if none were provided. The length of the active set vector is equal to the number of functions (m). In the case of an optimization data set with an objective function and two nonlinear constraints (three response functions total), the first ASV value is associated with the objective function and the remaining two are associated with the constraints (in whatever consistent constraint order has been defined by the user). The DVV defines a subset of the variables used for computing derivatives. Its identifiers are 1-based and correspond to the full set of variables listed in the first array. Finally, the analysis components pass additional strings from the user's `analysis_components` specification in a DAKOTA input file through to the simulator. They allow the development of simulation drivers that are more flexible, by allowing them to be passed additional specifics at run time, e.g., the names of model files such as a particular mesh to use.

For the APREPRO format option (see Section 12.6.2), the same set of data appears as follows:

```
{ DAKOTA_VARS     =                     2 }
{ cdv_1           =  1.500000000000000e+00 }
{ cdv_2           =  1.500000000000000e+00 }
{ DAKOTA_FNS      =                     3 }
{ ASV_1           =                     1 }
{ ASV_2           =                     1 }
{ ASV_3           =                     1 }
```

```
{ DAKOTA_DER_VARS =                    2 }
{ DVV_1           =                    1 }
{ DVV_2           =                    2 }
{ DAKOTA_AN_COMPS =                    0 }
```

where the numerical values are associated with their tags within "`{ tag = value }`" constructs.

The user-supplied simulation interface, comprised of a simulator program or driver and (optionally) filter programs, is responsible for reading the parameters file and creating a results file that contains the response data requested in the ASV. This response data is written in the format described in Section 14.2. Since the ASV contains all ones in this case, the response file corresponding to the above input file would contain values for the three functions:

```
1.250000000000000e-01 f
1.500000000000000e+00 c1
1.500000000000000e+00 c2
```

Since function tags are optional, the following would be equally acceptable:

```
1.250000000000000e-01
1.500000000000000e+00
1.500000000000000e+00
```

For the same parameters with different ASV components,

```
                     2 variables
1.500000000000000e+00 cdv_1
1.500000000000000e+00 cdv_2
                     3 functions
                     3 ASV_1
                     3 ASV_2
                     3 ASV_3
                     2 derivative_variables
                     1 DVV_1
                     2 DVV_2
                     0 analysis_components
```

the following response data is required:

```
1.250000000000000e-01 f
1.500000000000000e+00 c1
1.500000000000000e+00 c2
[ 5.000000000000000e-01 5.000000000000000e-01 ]
[ 3.000000000000000e+00 -5.000000000000000e-01 ]
[ -5.000000000000000e-01 3.000000000000000e+00 ]
```

Here, we need not only the function values, but also each of their gradients. The derivatives are computed with respect to `cdv_1` and `cdv_2` as indicated by the DVV values. Another modification to the ASV components yields the following parameters file:

```
                     2 variables
1.500000000000000e+00 cdv_1
```

```
1.500000000000000e+00 cdv_2
                    3 functions
                    2 ASV_1
                    0 ASV_2
                    2 ASV_3
                    2 derivative_variables
                    1 DVV_1
                    2 DVV_2
                    0 analysis_components
```

for which the following results file is needed:

```
[ 5.000000000000000e-01 5.000000000000000e-01 ]
[ -5.000000000000000e-01 3.000000000000000e+00 ]
```

Here, we need gradients for functions `f` and `c2`, but not for `c1`, presumably since this constraint is inactive.

A full Newton optimizer might make the following request:

```
                    2 variables
1.500000000000000e+00 cdv_1
1.500000000000000e+00 cdv_2
                    1 functions
                    7 ASV_1
                    2 derivative_variables
                    1 DVV_1
                    2 DVV_2
                    0 analysis_components
```

for which the following results file,

```
1.250000000000000e-01 f
[ 5.000000000000000e-01 5.000000000000000e-01 ]
[[ 3.000000000000000e+00 0.000000000000000e+00
   0.000000000000000e+00 3.000000000000000e+00 ]]
```

containing the objective function, its gradient vector, and its Hessian matrix, is needed. Again, the derivatives (gradient vector and Hessian matrix) are computed with respect to `cdv_1` and `cdv_2` as indicated by the DVV values.

Lastly, a more advanced example could have multiple types of variables present; in this example, 2 continuous and 3 discrete design, 2 normal uncertain, and 3 continuous and 2 discrete state variables. When a mixture of variable types is present, the content of the DVV (and therefore the required length of gradient vectors and Hessian matrices) depends upon the type of study being performed (see Section 14.3). For a reliability analysis problem, the uncertain variables are the active continuous variables and the following parameters file would be typical:

```
                   12 variables
1.500000000000000e+00 cdv_1
1.500000000000000e+00 cdv_2
                    2 ddv_1
                    2 ddv_2
                    2 ddv_3
```

```
5.000000000000000e+00 nuv_1
5.000000000000000e+00 nuv_2
3.500000000000000e+00 csv_1
3.500000000000000e+00 csv_2
3.500000000000000e+00 csv_3
                    4 dsv_1
                    4 dsv_2
                    3 functions
                    3 ASV_1
                    3 ASV_2
                    3 ASV_3
                    2 derivative_variables
                    6 DVV_1
                    7 DVV_2
                    2 analysis_components
            mesh1.exo AC_1
              db1.xml AC_2
```

Gradients are requested with respect to variable entries 6 and 7, which correspond to normal uncertain variables nuv_1 and nuv_2. The following response data would be appropriate:

```
7.943125000000000e+02 f
1.500000000000000e+00 c1
1.500000000000000e+00 c2
[ 2.560000000000000e+02 2.560000000000000e+02 ]
[ 0.000000000000000e+00 0.000000000000000e+00 ]
[ 0.000000000000000e+00 0.000000000000000e+00 ]
```

In a parameter study, however, no distinction is drawn between different types of continuous variables, and derivatives would be needed with respect to all continuous variables ($n_{dvv} = 7$ for the continuous design variables cdv_1 and cdv_2, the normal uncertain variables nuv_1 and nuv_2, and the continuous state variables csv_1, csv_2 and csv_3). The parameters file would appear as

```
                   12 variables
1.500000000000000e+00 cdv_1
1.500000000000000e+00 cdv_2
                    2 ddv_1
                    2 ddv_2
                    2 ddv_3
5.000000000000000e+00 nuv_1
5.000000000000000e+00 nuv_2
3.500000000000000e+00 csv_1
3.500000000000000e+00 csv_2
3.500000000000000e+00 csv_3
                    4 dsv_1
                    4 dsv_2
                    3 functions
                    3 ASV_1
                    3 ASV_2
                    3 ASV_3
                    7 derivative_variables
                    1 DVV_1
                    2 DVV_2
```

```
                     6 DVV_3
                     7 DVV_4
                     8 DVV_5
                     9 DVV_6
                    10 DVV_7
                     2 analysis_components
            mesh1.exo AC_1
              db1.xml AC_2
```

and the corresponding results would appear as

```
7.943125000000000e+02 f
1.500000000000000e+00 c1
1.500000000000000e+00 c2
[  5.000000000000000e-01  5.000000000000000e-01  2.560000000000000e+02
   2.560000000000000e+02  6.250000000000000e+01  6.250000000000000e+01
   6.250000000000000e+01 ]
[  3.000000000000000e+00 -5.000000000000000e-01  0.000000000000000e+00
   0.000000000000000e+00  0.000000000000000e+00  0.000000000000000e+00
   0.000000000000000e+00 ]
[ -5.000000000000000e-01  3.000000000000000e+00  0.000000000000000e+00
   0.000000000000000e+00  0.000000000000000e+00  0.000000000000000e+00
   0.000000000000000e+00 ]
```

Chapter 14

Responses

14.1 Overview

The `responses` specification in a DAKOTA input file specifies the types of data that can be returned from an interface during DAKOTA's execution. The specification includes the number and type of response functions (objective functions, nonlinear constraints, least squares terms, etc.) as well as availability of first and second derivatives (gradient vectors and Hessian matrices) for these response functions.

This chapter will present a brief overview of the response data sets and their uses, as well as cover some user issues relating to file formats and derivative vector and matrix sizing. For a detailed description of responses section syntax and example specifications, refer to the Responses Commands chapter in the DAKOTA Reference Manual [30].

14.1.1 Response function types

The types of response functions specified in the responses specification should be selected to be consistent with the type of iterative technique specified in the method specification:

- an optimization data set comprised of `num_objective_functions`, `num_nonlinear_inequality_constraints`, and `num_nonlinear_equality_constraints`. This data set is appropriate for use with optimization methods (e.g., the methods in Section 3.5).
- a least squares data set comprised of `num_least_squares_terms`, `num_nonlinear_inequality_constraints`, and `num_nonlinear_equality_constraints`. This data set is appropriate for use with nonlinear least squares algorithms (e.g., the methods in Section 3.7).
- a generic data set comprised of `num_response_functions`. This data set is appropriate for use with uncertainty quantification methods (e.g., the methods in Section 3.4).

Certain general-purpose iterative techniques, such as parameter studies and design of experiments methods, can be used with any of these data sets.

14.1.2 Gradient availability

Gradient availability for these response functions may be described by:

- `no_gradients`: gradient data is not needed.
- `numerical_gradients`: gradient data is needed and will be computed by finite differences.
- `analytic_gradients`: gradient data is needed and is available directly from the simulation code (finite differencing is not required).
- `mixed_gradients`: some gradient information is available directly from the simulation whereas the rest will have to be finite differenced.

The gradient specification also links back to the iterative method being employed. Gradient data is commonly needed when the iterative study involves gradient-based optimization, reliability analysis for uncertainty quantification, or local sensitivity analysis.

14.1.3 Hessian availability

Hessian availability for the response functions is similar to the gradient availability specifications, with the addition of support for quasi-Hessians:

- `no_hessians`: Hessian data is not needed.
- `numerical_gradients`: Hessian data is needed and will be computed by finite differences. These finite differences may be involve first-order differences of gradients (if analytic gradients are available for the response function of interest) or second-order differences of values (in all other cases).
- `quasi_hessians`: Hessian data is needed and will be accumulated using secant updates (BFGS or SR1) from a series of gradient evaluations.
- `analytic_hessians`: Hessian data is needed and is available directly from the simulation code.
- `mixed_hessians`: Hessian data is needed and will be obtained from a mix of numerical, analytic, and quasi sources.

The Hessian specification also links back to the iterative method in use, and use of Hessian data would commonly appear for gradient-based optimization using full Newton methods or for reliability analysis with second-order limit state approximations or second-order probability integrations.

14.2 DAKOTA Results File Data Format

Simulation interfaces which employ system calls and forks to create separate simulation processes must communicate with the simulation through the file system. This is accomplished through the reading and writing of parameters and results files. DAKOTA uses its own format for this data input/output. For the results file, only one format is supported (as compared to the two parameters file formats described in Section 12.6). Ordering of response functions is as listed in Section 14.1.1 (e.g., objective functions or least squares terms are first, followed by nonlinear inequality constraints, followed by nonlinear equality constraints).

```
<double> <fn_tag_1>
<double> <fn_tag_2>
...
<double> <fn_tag_m>
[ <double> <double> .. <double> ]
[ <double> <double> .. <double> ]
...
[ <double> <double> .. <double> ]
[[ <double> <double> .. <double> ]]
[[ <double> <double> .. <double> ]]
...
[[ <double> <double> .. <double> ]]
```

Figure 14.1: Results file data format.

After completion of a simulation, DAKOTA expects to read a file containing response data for the current set of parameters and corresponding to the current set of function requests in the active set vector. This response data must be in the following format:

The first block of data (shown in black) is the function values that have been requested, followed by a block of requested gradient data (shown in blue), followed by a block of requested Hessian data (shown in red). If the amount of data in the file does not match the function request vector, DAKOTA will abort with a response recovery format error message.

Function data have no bracket delimiters and one character tag per function can be *optionally* supplied. These tags are not used by DAKOTA and are only included as an optional field for consistency with the parameters file format and for backwards compatibility. The tags are rendered optional through DAKOTA's use of regular expression pattern matching to detect whether an upcoming field is numerical data or a tag. If character tags are used, then they must be separated from data by either white space or new line characters and there must not be any white space embedded within a character tag (e.g., use "`variable1`" or "`variable_1`," but not "`variable 1`").

Function gradient vectors are delimited with single brackets [...n_{dvv}-vector of doubles...]. Tags are not used and must not be present. White space separating the brackets from the data is optional.

Function Hessian matrices are delimited with double brackets [[...$n_{dvv} \times n_{dvv}$ matrix of doubles...]]. Data is listed by rows and can either be run together or broken onto multiple lines for readability. Tags are not used and must not be present. White space separating the brackets from the data is optional, although white space must not appear between the double brackets.

The format of the numeric fields may be floating point or scientific notation. In the latter case, acceptable exponent characters are "`E`" or "`e.`" A common problem when dealing with Fortran programs is that a C++ read of a numeric field using "`D`" or "`d`" as the exponent (i.e., a double precision value from Fortran) may fail or be truncated. In this case, the "`D`" exponent characters must be replaced either through modifications to the Fortran source or compiler flags or through a separate post-processing step (e.g., using the UNIX `sed` utility).

14.3 Active Variables for Derivatives

An important question for proper management of both gradient and Hessian data is: if several different types of variables are used, *for which variables are response function derivatives needed?* That is, how is n_{dvv} determined? The short answer is that the derivative variables vector (DVV) specifies the set of variables to be used for computing derivatives, and n_{dvv} is the length of this vector. The long answer is that, in most cases, the DVV is defined directly from the set of active continuous variables for the iterative method in use.

Since methods determine what subset, or view, of the variables data is active in the iteration, it is this same set of variables for which derivatives are most commonly computed (see also Section 12.5). Derivatives are never needed with respect to any discrete variables (since these derivatives do not exist) and the active continuous variables depend on the type of study being performed. For optimization and least squares problems, the active continuous variables are the *continuous design variables* ($n_{dvv} = n_{cdv}$) since this is the information used by the minimizer in computing a search direction. Similarly, for nondeterministic analysis methods which use gradient and/or Hessian information, the active continuous variables are the *uncertain variables* ($n_{dvv} = n_{uv}$), with the exception of `all_variables` mode. And lastly, parameter study methods which are cataloguing gradient and/or Hessian information do not draw a distinction among continuous variables; therefore, the active continuous variables are defined from *all continuous variables* that are specified ($n_{dvv} = n_{cdv} + n_{uv} + n_{csv}$).

In a few cases, derivatives are needed with respect to the *inactive* continuous variables. For example, when performing reliability analysis within reliability-based design optimization, derivatives of the generic response function data set may be needed with respect to the design variables, which are inactive continuous variables within the uncertainty quantification. These instances are the reason for the creation and inclusion of the DVV guidance for derivative estimation.

In all cases, if the DVV is honored, then the correct derivative components are returned. In simple cases, such as optimization and least squares studies that only specify design variables and for nondeterministic analyses that only specify uncertain variables, then derivative component subsets are not an issue and the exact content of the DVV may be safely ignored.

Chapter 15

Inputs to DAKOTA

15.1 Overview of Inputs

The DAKOTA executable supports a number of command line inputs, as described in Section 2.1.5. Among these are specifications for the DAKOTA input file and, optionally, a restart file. The syntax of the DAKOTA input file is described in detail in the DAKOTA Reference Manual [30], and the restart file is described in Chapter 20.

The DAKOTA input file may be prepared manually (e.g., using a text editor such as `xemacs` or `vi`), or it may be defined graphically using the JAGUAR graphical user interface, as described in Section 15.2 below. Once prepared, the DAKOTA input file may identify additional files for data import as described in Section 15.3.

15.2 JAGUAR

A short description of the steps for downloading, installing, and executing JAGUAR is provided below. For DAKOTA 4.2, JAGUAR is undergoing a refactoring, and not all features are fully operational in the existing beta releases.

- **Download the JAGUAR installer**. As for the DAKOTA download process described in Section 2.1.1, the JAGUAR distribution is accessed by clicking on the download link available from:

  `http://www.cs.sandia.gov/DAKOTA/software.html`

  and filling out the short online registration form.

- **Install supporting JAVA software** (if needed). If not already installed on your machine, you will need the "Java 2 Platform, Standard Edition (J2SE)" in version 1.4.2 or newer (note that Sun has recently revised its 1.x.x versioning and the recently released 5.0 is the same as 1.5.0 in the old numbering scheme).

  `http://java.sun.com/j2se/1.5.0/download.jsp` [click on "Download JRE 5.0 ..."]

  `http://www.java.com/en/` [click on download]

- **Run the installer** (either double-click the icon or execute "java -jar JaguarInstall-*version-date*.jar"). Figure 15.1 shows a screen capture of the installer.

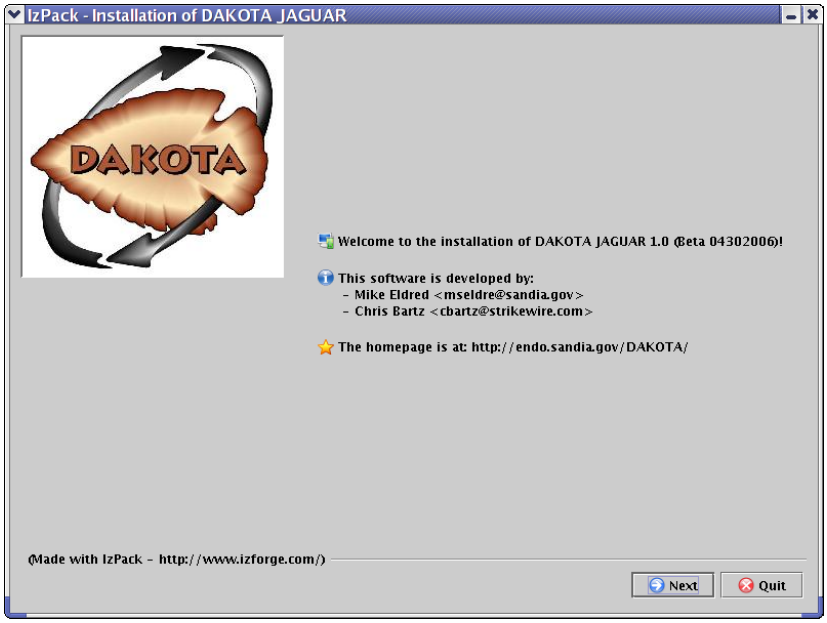


Figure 15.1: The JAGUAR installer.

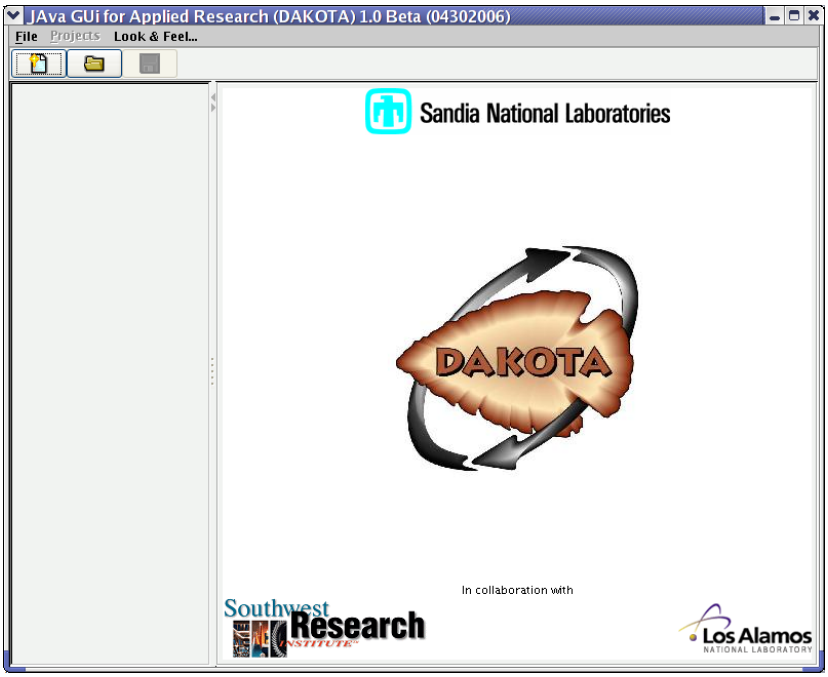


Figure 15.2: The initial JAGUAR splash screen.

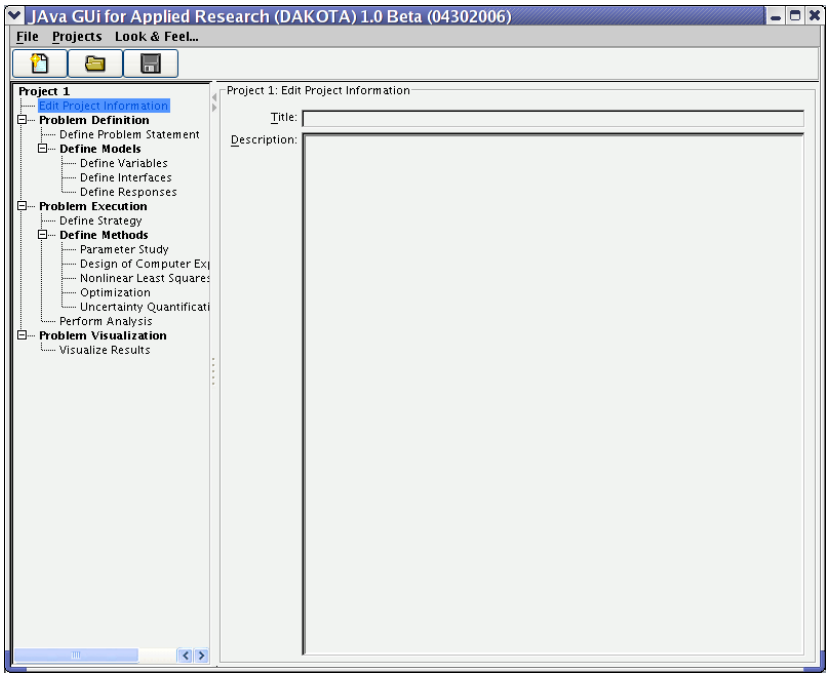


Figure 15.3: JAGUAR navigation.

- **Execute the installed GUI** (either double-click the new icon or execute "jaguar.sh" in the install directory). Figure 15.2 shows the splash screen for the JAGUAR GUI.

- **Select File → New**. Click the "File" pull-down and select "New." Observe the navigation aids on the left of the window, as shown in Figure 15.3.

- **Visit specifications**. At a minimum, select Variables, Interfaces, Responses, and Method on the left navigation. For each, select a set id (or use the default) and click OK. Select your desired settings for a particular run. Figures 15.4–15.6 show sample selections for Variables, Interfaces, and Responses. Now with the model components specified, an iterative method is selected, as shown in Figure 15.7.

- **Perform Analysis**. Visit the "Perform Analysis" view on the left navigation and review the input selections generated by JAGUAR, as shown in Figure 15.8. "Run DAKOTA" (within the Perform Analysis view) and "Problem Visualization" are not yet active, so click the File pull-down, select "Save as...," and run DAKOTA separately for now. If the DAKOTA and JAGUAR versions are not synchronized, some minor editing of this input file may be required.

15.3 Data Imports

The DAKOTA input file may identify additional files used to import data into DAKOTA.

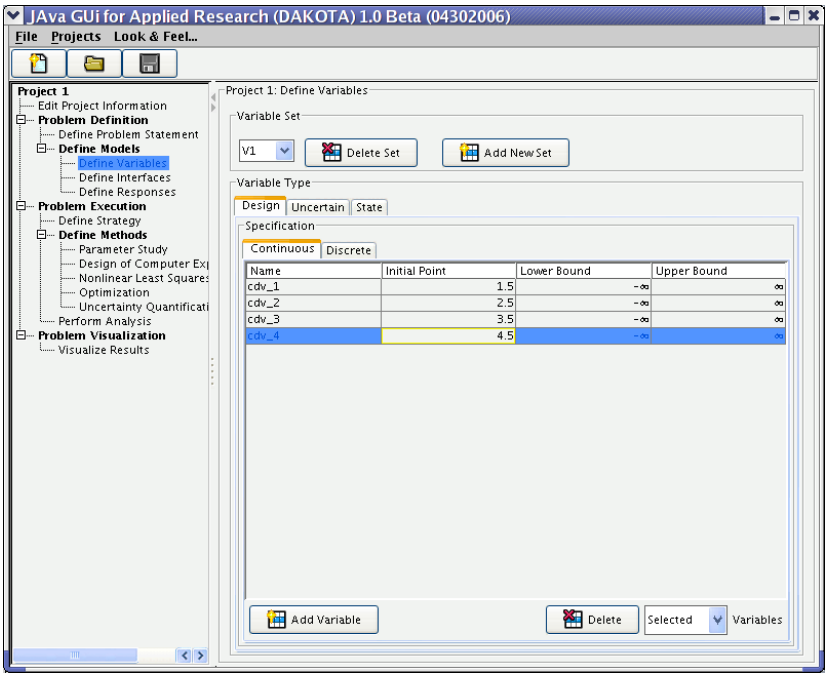


Figure 15.4: An example of JAGUAR variables inputs.

15.3.1 AMPL algebraic mappings: stub.nl, stub.row, and stub.col

As described in Section 13.2, an AMPL specification of algebraic input-to-output relationships may be imported into DAKOTA and used to define or augment the mappings of a particular interface.

15.3.2 Genetic algorithm population import

Genetic algorithms (GAs) from the JEGA and COLINY packages support a population import feature using the keywords `initialization_type flat_file = STRING`. This is useful for warm starting GAs from available data or previous runs. Refer to the Method Specification chapter in the DAKOTA Reference Manual [30] for additional information on this specification.

15.3.3 PCE coefficient import

Polynomial chaos expansion (PCE) methods compute coefficients for response expansions which employ a basis of multivariate orthogonal polynomials. Normally, the `nond_polynomial_chaos` method calculates these coefficients based either on a spectral prjection or a linear regression (see Section 6.4). However, DAKOTA also supports the option of importing a set of response PCE coefficients based on the specification `expansion_import_file = STRING`. This is useful for evaluating moments analytically or computing probabilities numerically from a known response expansion. Refer to the Method Specification chapter in the DAKOTA Reference Manual [30] for additional information on this specification.

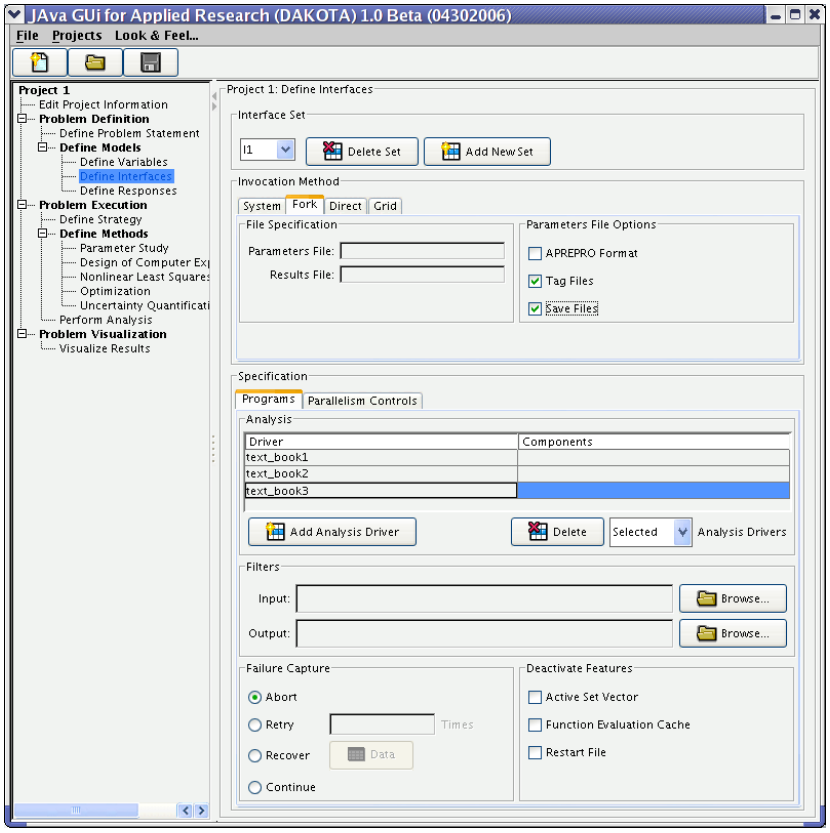


Figure 15.5: An example of JAGUAR interface inputs.

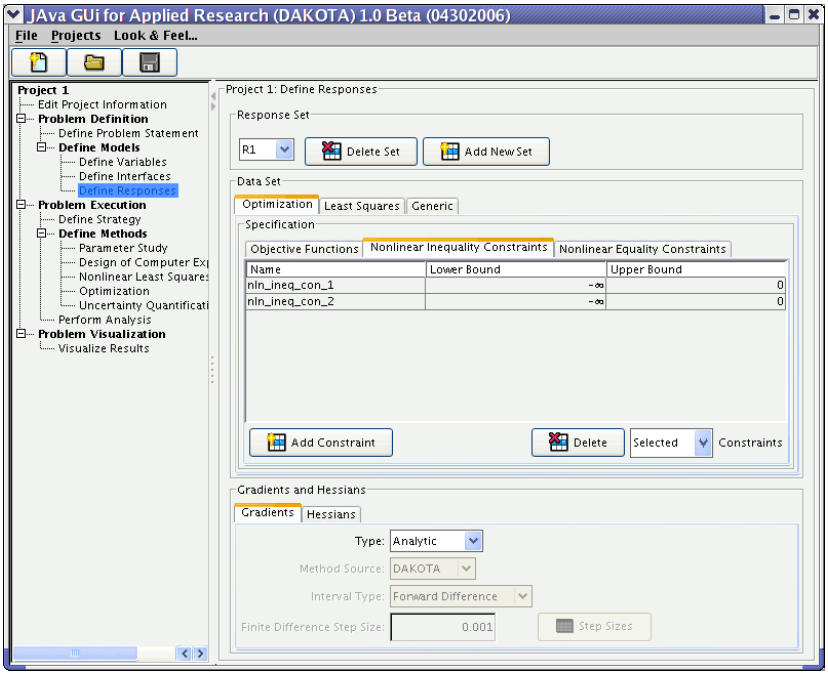


Figure 15.6: An example of JAGUAR responses inputs.

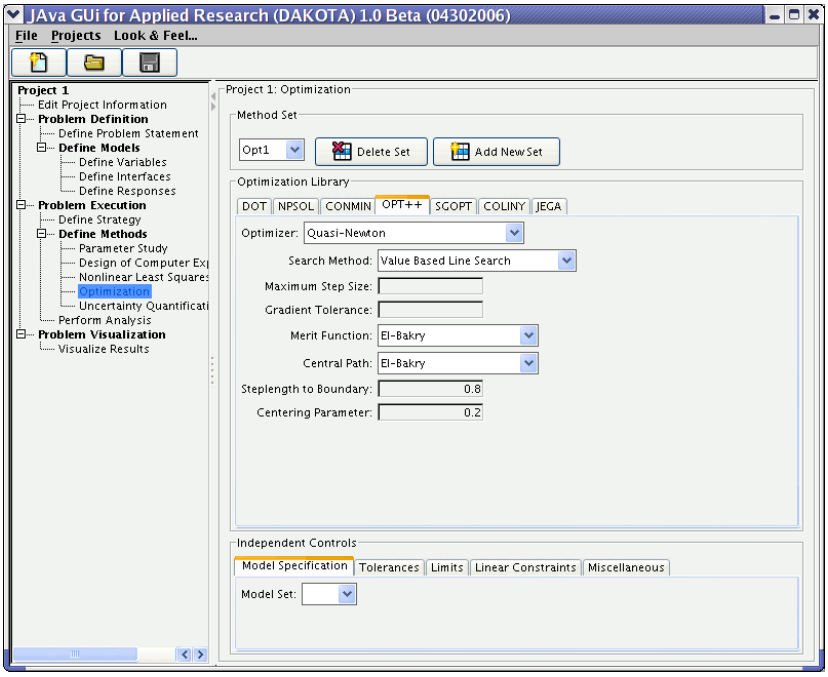


Figure 15.7: An example of JAGUAR method inputs.

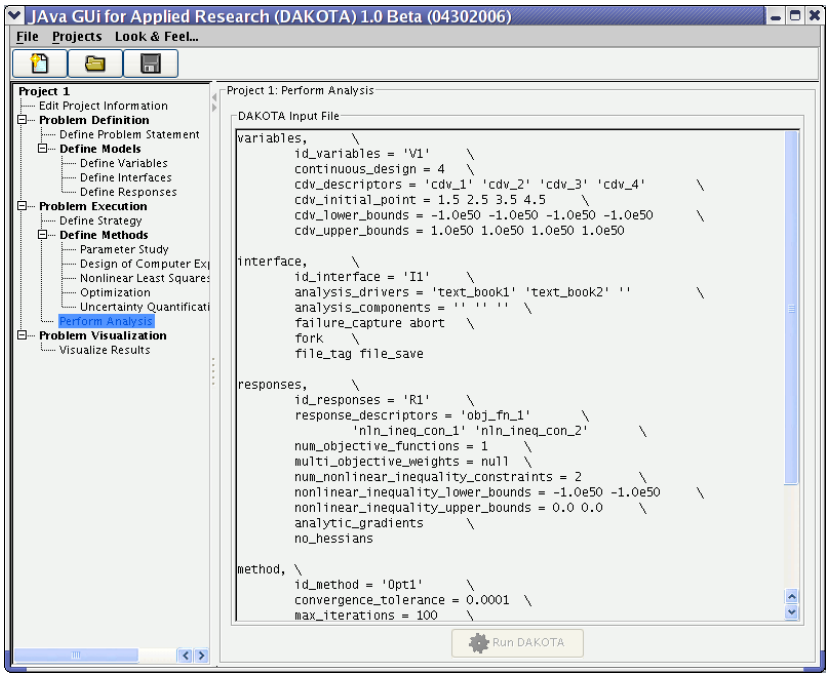


Figure 15.8: Review of the input file generated by JAGUAR.

15.3.4 Surrogate construction from data files

Global data fit surrogates may be constructed from a variety of data sources. One of these sources is an auxilliary data file, as specified by the keywords `reuse_samples samples_file = STRING`. Refer to the Model Specification chapter in the DAKOTA Reference Manual [30] for additional information on this specification.

Chapter 16

Output from DAKOTA

16.1 Overview of Output Formats

Given an emphasis on complex numerical simulation codes that run on massively parallel supercomputers, DAKOTA's output has been designed to provide a succinct, text-based reporting of the progress of the iterations and function evaluations performed by an algorithm. In addition, DAKOTA provides a tabular output format that is useful for data visualization with external tools and a basic graphical output capability that is useful as a monitoring tool. The JAGUAR graphical user interface described in Section 15.2 is also an emerging capability that will provide more advanced visualization facilities in time.

16.2 Standard Output

DAKOTA outputs basic information to "standard out" (i.e., the screen) for each function evaluation, consisting of an evaluation number, parameter values, execution syntax, the active set vector, and the response data set. To describe the standard output of DAKOTA, optimization of the "container" problem (see Chapter 22 for problem formulation) is used as an example. The input file for this example is shown in Figure 16.1. In this example, there is one equality constraint, and DAKOTA's finite difference algorithm is used to provide central difference numerical gradients to the NPSOL optimizer.

```
strategy,
        single_method
        graphics
        tabular_graphics_data

method,
        npsol_sqp

variables,
        continuous_design = 2
          descriptors'H' 'D'
          initial_point 4.5 4.5
          lower_bounds  0.0 0.0

interface,
        system
          analysis_driver = 'container'
          parameters_file = 'container.in'
          results_file    = 'container.out'
          file_tag

responses,
        num_objective_functions = 1
        num_nonlinear_equality_constraints = 1
        numerical_gradients
          method_source dakota
          interval_type central
          fd_gradient_step_size = 0.001
        no_hessians
```

Figure 16.1: DAKOTA input file for the “container” example problem.

A partial listing of the DAKOTA output for the container optimization example follows:

```
Running MPI executable in serial mode.
DAKOTA version 4.2 released 11/05/2008.
Constructing Single Method Strategy...
Writing new restart file dakota.rst
methodName = npsol_sqp
gradientType = numerical
Numerical gradients using central differences
to be calculated by the dakota finite difference routine.
hessianType = none

>>>>> Running Single Method Strategy.

>>>>> Running npsol_sqp iterator.




                     NPSOL  ---  Version 5.0-2      Sept 1995
                     ========================================

-----------------------------------------
Begin Dakota derivative estimation routine
-----------------------------------------

>>>>> Initial map for analytic portion of response:

------------------------------
Begin Function Evaluation    1
------------------------------
Parameters for function evaluation 1:
                      4.5000000000e+00 H
                      4.5000000000e+00 D

(container container.in.1 container.out.1)

Active response data for function evaluation 1:
Active set vector = { 1 1 }
                      1.0713145108e+02 obj_fn
                      8.0444076396e+00 nln_eq_con_1


>>>>> Dakota finite difference gradient evaluation for x[1] + h:

------------------------------
Begin Function Evaluation    2
------------------------------
Parameters for function evaluation 2:
                      4.5045000000e+00 H
                      4.5000000000e+00 D

(container container.in.2 container.out.2)
```

```
Active response data for function evaluation 2:
Active set vector = { 1 1 }
                      1.0719761302e+02 obj_fn
                      8.1159770472e+00 nln_eq_con_1


>>>>> Dakota finite difference gradient evaluation for x[1] - h:
------------------------------
Begin Function Evaluation    3
------------------------------
Parameters for function evaluation 3:
                      4.4955000000e+00 H
                      4.5000000000e+00 D

(container container.in.3 container.out.3)

Active response data for function evaluation 3:
Active set vector = { 1 1 }
                      1.0706528914e+02 obj_fn
                      7.9728382320e+00 nln_eq_con_1


>>>>> Dakota finite difference gradient evaluation for x[2] + h:

------------------------------
Begin Function Evaluation    4
------------------------------
Parameters for function evaluation 4:
                      4.5000000000e+00 H
                      4.5045000000e+00 D

(container container.in.4 container.out.4)

Active response data for function evaluation 4:
Active set vector = { 1 1 }
                      1.0727959301e+02 obj_fn
                      8.1876180243e+00 nln_eq_con_1


>>>>> Dakota finite difference gradient evaluation for x[2] - h:

------------------------------
Begin Function Evaluation    5
------------------------------
Parameters for function evaluation 5:
                      4.5000000000e+00 H
                      4.4955000000e+00 D

(container container.in.5 container.out.5)

Active response data for function evaluation 5:
Active set vector = { 1 1 }
                      1.0698339109e+02 obj_fn
```

```
                      7.9013403937e+00 nln_eq_con_1


>>>>> Total response returned to iterator:

Active set vector = { 3 3 } Deriv vars vector = { 1 2 }
                      1.0713145108e+02 obj_fn
                      8.0444076396e+00 nln_eq_con_1
 [  1.4702653619e+01  3.2911324639e+01 ] obj_fn gradient
 [  1.5904312809e+01  3.1808625618e+01 ] nln_eq_con_1 gradient




 Majr Minr    Step  Fun  Merit function Norm gZ  Violtn   nZ Penalty Conv
    0    1 0.0E+00    1  9.90366719E+01 1.6E+00 8.0E+00    1 0.0E+00 F FF

<SNIP>

>>>>> Dakota finite difference gradient evaluation for x[2] - h:

-------------------------------
Begin Function Evaluation   40
-------------------------------
Parameters for function evaluation 40:
                      4.9873894231e+00 H
                      4.0230575428e+00 D

(container container.in.40 container.out.40)

Active response data for function evaluation 40:
Active set vector = { 1 1 }
                      9.8301287596e+01 obj_fn
                     -1.2698647501e-01 nln_eq_con_1


>>>>> Total response returned to iterator:

Active set vector = { 3 3 } Deriv vars vector = { 1 2 }
                      9.8432498116e+01 obj_fn
                     -9.6918029158e-12 nln_eq_con_1
 [  1.3157517860e+01  3.2590159623e+01 ] obj_fn gradient
 [  1.2737124497e+01  3.1548877601e+01 ] nln_eq_con_1 gradient


    7    1 1.0E+00    8  9.84324981E+01 4.8E-11 9.7E-12    1 1.7E+02 T TT

 Exit NPSOL - Optimal solution found.

 Final nonlinear objective value =    98.43250

NPSOL exits with INFORM code = 0 (see "Interpretation of output" section in NPSO
L manual)
```

```
NOTE: see Fortran device 9 file (fort.9 or ftn09)
      for complete NPSOL iteration history.

<<<<< Iterator npsol_sqp completed.
<<<<< Function evaluation summary: 40 total (40 new, 0 duplicate)
<<<<< Best parameters          =
                     4.9873894231e+00 H
                     4.0270846274e+00 D
<<<<< Best objective function  =
                     9.8432498116e+01
<<<<< Best constraint values   =
                    -9.6918029158e-12
<<<<< Best data captured at function evaluation 36
<<<<< Single Method Strategy completed.
DAKOTA execution time in seconds:
  Total CPU        =       0.05 [parent =   0.051992, child =  -0.001992]
  Total wall clock =   0.396813
```

The first block of lines provide a report on the DAKOTA configuration and settings. The lines that follow, down to the line "`Exit NPSOL - Optimal solution found`", contain information about the function evaluations that have been requested by NPSOL and performed by DAKOTA. Evaluations 6 through 39 have been omitted from the listing for brevity.

Following the line "`Begin Function Evaluation 1`", the initial values of the design variables, the syntax of the function evaluation, and the resulting objective and constraint function values are listed. The values of the design variables are labeled with the tags `H` and `D`, respectively, according to the descriptors to these variables given in the input file, Figure 16.1. The values of the objective function and volume constraint are labeled with the tags `obj_fn` and `nln_eq_con_1`, respectively. Note that the initial design parameters are infeasible since the equality constraint is violated ($\neq 0$). However, by the end of the run, the optimizer finds a design that is both feasible and optimal for this example. Between the design variables and response values, the content of the system call to the simulator is displayed as "`(container container.in.1 container.out.1)`", with `container` being the name of the simulator and `container.in.1` and `container.out.1` being the names of the parameters and results files, respectively.

Just preceding the output of the objective and constraint function values is the line "`Active set vector =` {`1 1`}". The active set vector indicates the types of data that are required from the simulator for the objective and constraint functions, and values of "`1`" indicate that the simulator must return values for these functions (gradient and Hessian data are not required). For more information on the active set vector, see Section 12.7.

Since finite difference gradients have been specified, DAKOTA computes their values by making additional function evaluation requests to the simulator at perturbed parameter values. Examples of the gradient-related function evaluations have been included in the sample output, beginning with the line that reads "`>>>>> Dakota finite difference evaluation for x[1] + h:`". The resulting finite difference gradients are listed after function evaluation 5 beginning with the line "`>>>>> Total response returned to iterator:`". Here, another active set vector is displayed in the DAKOTA output file. The line "`Active set vector =` { `3 3` }" indicates that the total response resulting from the finite differencing contains function values and gradients.

The final lines of the DAKOTA output, beginning with the line "`<<<<< Iterator npsol_sqp completed`", summarize the results of the optimization study. The best values of the optimization parameters, objective function, and volume constraint are presented along with the function evaluation number where they occurred, total function evaluation counts, and a timing summary. In the end, the objective function has been minimized and the equality constraint has been satisfied (driven to zero within the constraint tolerance).

```
%eval_id              H              D         obj_fn   nln_eq_con_1
       1            4.5            4.5    107.1314511     8.04440764
       2    5.801246882    3.596476363    94.33737399    -4.59103645
       3    5.197920019    3.923577479     97.7797214  -0.6780884711
       4    4.932877133    4.044776216    98.28930566  -0.1410680284
       5    4.989328733    4.026133158     98.4270019 -0.005324671422
       6    4.987494493    4.027041977    98.43249058 -7.307058462e-06
       7    4.987391669     4.02708372    98.43249809 -2.032539782e-08
       8    4.987389423    4.027084627    98.43249812 -9.630143905e-12
```

Figure 16.2: DAKOTA's tabular output file showing the iteration history of the "container" optimization problem.

The DAKOTA results are intermixed with iteration information from the NPSOL library. The lines with the heading "`Majr Minr Step Fun Merit function Norm gZ Violtn nZ Penalty Conv`" come from Fortran write statements within NPSOL. The output is mixed since both DAKOTA and NPSOL are writing to the same standard output stream. The relative locations of these output contributions can vary depending on the specifics of output buffering and flushing on a particular platform and depending on whether or not the standard output is being redirected to a file. In some cases, output from the optimization library may appear on each iteration (as in this example), and in other cases, it may appear at the end of the DAKOTA output. Finally, a more detailed summary of the NPSOL iterations is written to the Fortran device 9 file (e.g., `fort.9` or `ftn09`).

16.3 Tabular Output Data

DAKOTA has the capability to print the iteration history in tabular form to a file. The keyword `tabular_graphics_data` needs to be included in the strategy specification (see Figure 16.1). The primary intent of this capability is to facilitate the transfer of DAKOTA's iteration history data to an external mathematical analysis and/or graphics plotting package (e.g., MATLAB, TECplot, Excel, S-plus, Minitab). Any evaluations from DAKOTA's internal finite differencing are suppressed, which leads to better data visualizations. This suppression of lower level data is consistent with the data that is sent to the graphics windows, as described in Section 16.4. If this data suppression is undesirable, Section 20.2.3 describes an approach where every function evaluation, even the ones from finite differencing, can be saved to a file in tabular format.

The default file name for the tabular output data is "`dakota_tabular.dat`" and the output from the "container" optimization problem is shown in Figure 16.2. This file contains the complete history of data requests from NPSOL (8 requests map into a total of 40 function evaluations when including the central finite differencing). The first column is the data request number, the second and third columns are the design parameter values (labeled in the example as "`H`" and "`D`"), the fourth column is the objective function (labeled "`obj_fn`"), and the fifth column is the nonlinear equality constraint (labeled "`nln_eq_con_1`").

16.4 Graphics Output

Graphics capabilities are available for monitoring the progress of an iterative study. The graphics option is invoked by adding the `graphics` flag in the strategy specification of the DAKOTA input file (see Figure 16.1). The graphics display the values of each response function (e.g., objective and constraint functions) and each parameter for the function evaluations in the study. As for the tabular output described in Section 16.3, internal finite

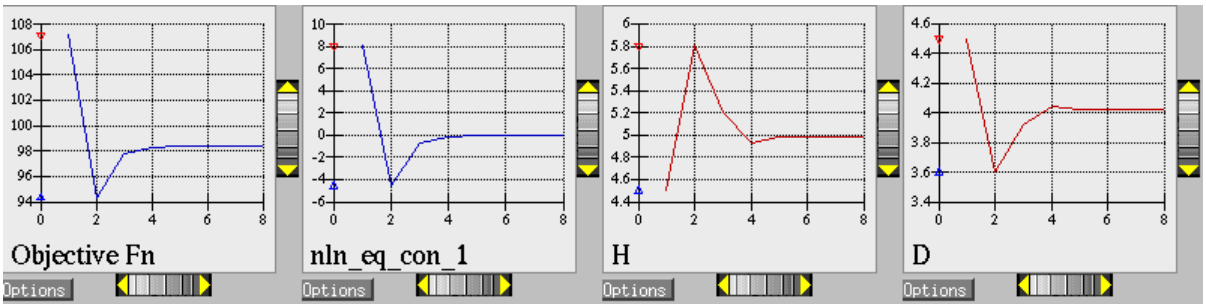


Figure 16.3: DAKOTA 2D graphics for "container" problem showing history of an objective function, an equality constraint, and two variables.

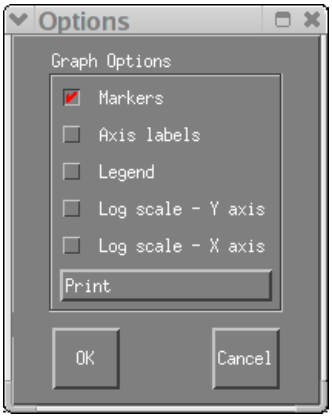


Figure 16.4: Options for DAKOTA 2D graphics.

difference evaluations are suppressed in order to omit this clutter from the graphics. Figure 16.3 shows the optimization iteration history for the container example.

If DAKOTA is executed on a remote machine, the DISPLAY variable in the user's UNIX environment [58] may need to be set to the local machine in order to display the graphics window.

The scroll bars which are located on each graph below and to the right of each plot may be operated by dragging on the bars or pressing the arrows, both of which result in expansion/contraction of the axis scale. Clicking on the "Options" button results in the window shown in Figure 16.4, which allows the user to include min/max markers on the vertical axis, vertical and horizontal axis labels, and a plot legend within the corresponding graphics plot. In addition, the values of either or both axes may be plotted using a logarithmic scale (so long as all plot values are greater than zero) and an encapsulated postscript (EPS) file, named `dakota_graphic_i.eps` where i is the plot window number, can be created using the "Print" button.

In addition to these two-dimensional iteration history plots, three-dimensional surface plots can be generated when using data fit surrogate models (see Section 11.4.1) in combination with the graphics keyword. This feature is currently available only if there are two parameters in the problem (a mechanism for selecting a two parameter subset of an n-dimensional problem is not currently available). When DAKOTA is executed, a 3-D surface plot is automatically spawned (Figure 16.5 shows an example from optimization of the Rosenbrock problem). The creation of the 3-D surface plot pauses the advance of the iterative algorithm. To continue progress, click the right mouse button or hit return while the mouse cursor is in the 3D graphics window.

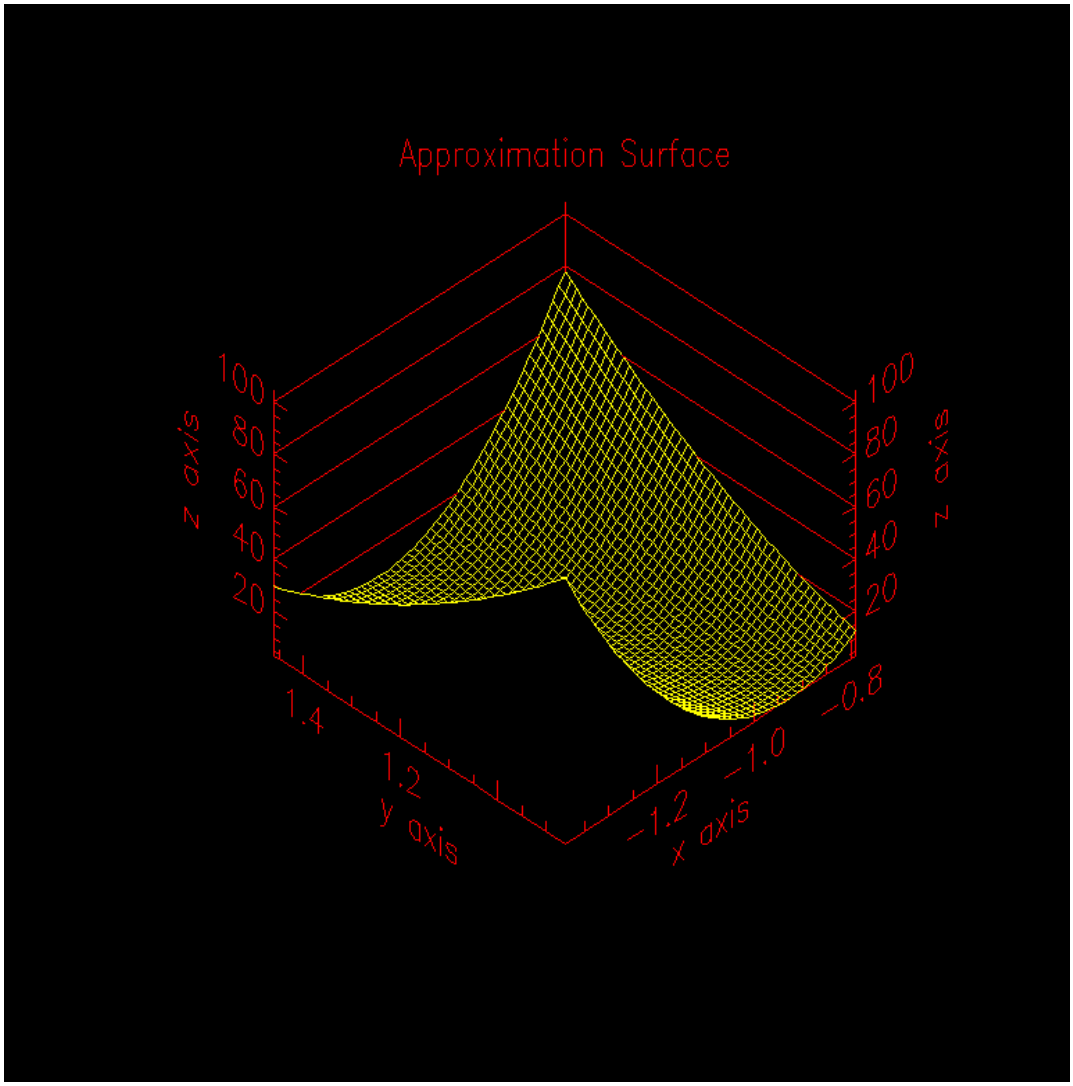


Figure 16.5: An example of the 3-D surface plotting that is available for surrogate-based optimization with two design parameters.

The 3D graphics from the PLplot library have a dependency on external font files. If the 3D graphics fail with a message similar to:

```
Cannot open library file: plstnd5.fnt
lib dir="<...some_path...>"
*** PLPLOT ERROR ***
Unable to open font file
Program aborted
```

then the solution is to locate the font files that came with your DAKOTA installation and set the `$PLPLOT_LIB` environment variable to point to them, e.g.:

```
setenv PLPLOT_LIB /home/<user_name>/Dakota/packages/plplot/data
```

16.5 Error Messages Output

A variety of error messages are printed by DAKOTA in the event that an error is detected in the input specification. Some of the more common input errors, and the associated error messages, are described below. See also the Common Specification Mistakes section in the DAKOTA Reference Manual [30].

Incorrectly spelled specifications, such as `''numericl_gradients''`, will result in error messages of the form:

```
Parser detected syntax error: unrecognized identifier 'numericl_gradients'
  within responses keyword.
Please refer to the dakota.input.txt distributed with this executable.
```

The input parser catches syntax errors, but not logic errors. The fact that certain input combinations are erroneous must be detected after parsing, at object construction time. For example, if a `no_gradients` specification for a response data set is combined with selection of a gradient-based optimization method, then this error must be detected during set-up of the optimizer (see last line of listing):

```
Running MPI executable in serial mode.
DAKOTA version 4.0 released 05/12/2006.
Writing new restart file dakota.rst
Constructing Single Method Strategy...
methodName = npsol_sqp
gradientType = none
hessianType = none

Error: gradient-based optimizers require a gradient specification.
```

Another common mistake involves a mismatch between the amount of data expected on a function evaluation and the data returned by the user's simulation code or driver. The available response data is specified in the responses keyword block, and the subset of this data needed for a particular evaluation is managed by the active set vector. For example, if DAKOTA expects function values and gradients to be returned (as indicated by an active set vector containing 3's), but the user's simulation code only returns function values, then the following error message is generated:

```
At EOF: insufficient data for functionGradient 1
```

Unfortunately, descriptive error messages are not available for all possible failure modes of DAKOTA. If you encounter core dumps, segmentation faults, or other failures, please report the problem by sending E-mail to dakota-developers@development.sandia.gov and include the information specified in the support FAQ.

Chapter 17

Advanced Simulation Code Interfaces

17.1 Building an Interface to a Engineering Simulation Code

To interface an engineering simulation package to DAKOTA using one of the black-box interfaces (system call or fork), pre- and post-processing functionality typically needs to be supplied (or developed) in order to transfer the parameters from DAKOTA to the simulator input file and to extract the response values of interest from the simulator's output file for return to DAKOTA (see Figures 1.1 and 13.1). This is often managed through the use of scripting languages, such as C-shell [4], Bourne shell [10], Perl [134], or Python [89]. While these are common and convenient choices for simulation drivers/filters, it is important to recognize that any executable file can be used. If the user prefers, the desired pre- and post-processing functionality may also be compiled or interpreted from any number of programming languages (C, C++, F77, F95, JAVA, Basic, etc.).

In the `Dakota/examples/script_interfaces/generic` directory, a simple example uses the Rosenbrock test function as a mock engineering simulation code. Several scripts have been included to demonstrate ways to accomplish the pre- and post-processing needs. Actual simulation codes will, of course, have different pre- and post-processing requirements, and as such, this example serves only to demonstrate the issues associated with interfacing a simulator. Modifications will almost surely be required for new applications.

17.1.1 Generic Script Interface Files

The `generic` directory contains four important files: `dakota_rosenbrock.in` (the DAKOTA input file), `simulator_script` (the simulation driver script), `dprepro` (a pre-processing utility), and `templatedir/ros.template` (a template simulation input file).

The `dakota_rosenbrock.in` file specifies the study that DAKOTA will perform and, in the interface section, describes the components to be used in performing function evaluations. In particular, it identifies `simulator_script` as its `analysis_driver`, as shown in Figure 17.1.

The `simulator_script` listed in Figure 17.2 is a short C-shell driver script that DAKOTA executes to perform each function evaluation. The names of the parameters and results files are passed to the script on its command line so that they can be referenced internal to the script by the variables `$argv[1]` and `$argv[2]`, respectively. The `simulator_script` is divided into five parts: set up, pre-processing, analysis, post-processing, and clean up.

The set up portion strips the function evaluation number from `$argv[1]` and assigns it to the shell variable `$num`,

```
#  DAKOTA INPUT FILE - dakota_rosenbrock.in
#  This sample Dakota input file optimizes the Rosenbrock function.
#  See p. 95 in Practical Optimization by Gill, Murray, and Wright.

method,
        npsol_sqp
# if NPSOL is not available, comment the above and try the following instead:
#       conmin_frcg

variables,
        continuous_design = 2
          cdv_initial_point   -1.0      1.0
          cdv_lower_bounds    -2.0     -2.0
          cdv_upper_bounds     2.0      2.0
          cdv_descriptor       'x1'     'x2'

interface,
        system
#         asynchronous
          analysis_driver = 'simulator_script'
          parameters_file = 'params.in'
          results_file    = 'results.out'
          file_tag
# uncomment to leave params.in.X and results.out.X files in directory
#         file_save
          aprepro

responses,
        num_objective_functions = 1
        numerical_gradients
          fd_gradient_step_size = .000001
# to instead use analytic gradients returned by the simulator comment the
# preceding two lines and uncomment the following:
#       analytic_gradients

        no_hessians
```

Figure 17.1: The `dakota_rosenbrock.in` input file.

```
#!/bin/csh -f
# Sample simulator to Dakota system call script
# See Advanced Simulation Code Interfaces chapter in Users Manual

# $argv[1] is params.in.(fn_eval_num) FROM Dakota
# $argv[2] is results.out.(fn_eval_num) returned to Dakota

# ------------------------
# Set up working directory
# ------------------------

set num = ‘echo $argv[1] | cut -c 11-‘
mkdir workdir.$num
# or for applications requiring the template files in the working directory
#cp -r templatedir workdir.$num

cd workdir.$num

# --------------
# PRE-PROCESSING
# --------------
# Incorporate the parameters from DAKOTA into the template, writing ros.in
# Use the following line if SNL’s APREPRO utility is used instead of DPrePro.
# ../aprepro -c ’*’ -q --nowarning ros.template ros.in

../dprepro ../$argv[1] ../templatedir/ros.template ros.in

# --------
# ANALYSIS
# --------

../rosenbrock_bb

# ---------------
# POST-PROCESSING
# ---------------

# extract function value from the simulation output
grep ’Function value’ ros.out | cut -c 18- >! $argv[2]
# extract gradients from the simulation output (in this case will be ignored
# by DAKOTA if not needed)
grep -i ’Function g’ ros.out | cut -c 21- >> $argv[2]

# moving $argv[2] at the end avoids any problems with read race conditions
mv $argv[2] ../.

# --------
# Clean up
# --------

cd ..
#\rm -rf workdir.$num
```

Figure 17.2: The `simulator_script` sample driver script.

which is then used to create a tagged working directory for a particular evaluation. For example, on the first evaluation, "1" is stripped from "params.in.1" in order to create "workdir.1". The primary reason for creating separate working directories is so that the files associated with one simulation do not conflict with those for another simulation. This is particularly important when executing concurrent simulations in parallel (to actually execute function evaluations concurrently, uncomment the asynchronous line in dakota_rosenbrock.in).

In the pre-processing portion, the simulator_script utilizes dprepro, which is a parsing utility used to extract the current variable values from a parameters file ($argv[1]) and then insert them into the simulator template input file (ros.template) to create a new input file (ros.in) for the simulator. Internal to Sandia, the APREPRO utility is often used for this purpose. For external sites where APREPRO is not available, the DPrePro utility mentioned above is an alternative with many of the capabilities of APREPRO that is specifically tailored for use with DAKOTA and is distributed with it (in Dakota/examples/script_interfaces/generic/dprepro, or Dakota/bin in a binary distribution). Additionally, the BPREPRO utility is another alternative to APREPRO (see [136]), and at Lockheed Martin sites, the JPrePost utility is available as a JAVA pre- and post-processor [44]. The dprepro script partially listed in Figure 17.3 will be used here for simplicity of discussion. It can use either DAKOTA's aprepro parameters file format (see Section 12.6.2) or DAKOTA's standard format (see Section 12.6.1), so either option may be selected in the interface section of the DAKOTA input file. The ros.template file listed in Figure 17.4 is a template simulation input file which contains targets for the incoming variable values, identified by the strings "{x1}" and "{x2}". These identifiers match the variable descriptors specified in dakota_rosenbrock.in. The template input file is contrived as Rosenbrock has nothing to do with finite element analysis; it only mimics a finite element code in order to demonstrate the simulator template process. The dprepro script will search the simulator template input file for fields marked with the curly brackets and then create a new file (ros.in) by replacing these targets with the corresponding numerical values for the variables. As noted in the usage information for dprepro and shown in simulator_script, the names for the DAKOTA parameters file ($argv[1]), template file (ros.template), and generated input file (ros.in) must be specified in the dprepro command line arguments.

The third part of the script executes the rosenbrock_bb simulator. The input and output file names, ros.in and ros.out, respectively, are hard-coded into the FORTRAN 77 program rosenbrock_bb.f. When the rosenbrock_bb simulator is executed, the values for x1 and x2 are read in from ros.in, the Rosenbrock function is evaluated, and the function value is written out to ros.out.

The fourth part performs the post-processing and returns the response results to DAKOTA. Using the UNIX "grep" utility, the particular response values of interest are extracted from the raw simulator output and saved to $argv[2], which in the case of the first evaluation is "results.out.1". This results file is moved up one level, out of the working directory, so that DAKOTA may retrieve it. Note that moving the completed results file up a level at the end of the evaluation avoids any problems with read race conditions (see Section 18.2.1).

Finally, in the clean up phase, the working directory is removed to reduce the amount of disk storage required to execute the study. If data from each simulation needs to be saved, this step can be commented out by inserting a "#" character before "\rm -rf".

As an example of the data flow on a particular function evaluation, consider evaluation 60. The parameters file for this evaluation (params.in.60) consists of:

```
{ DAKOTA_VARS     =                      2 }
{ x1              =  1.638248083045767e-01 }
{ x2              =  2.197300525769129e-02 }
{ DAKOTA_FNS      =                      1 }
{ ASV_1           =                      1 }
{ DAKOTA_DER_VARS =                      2 }
{ DVV_1           =                      1 }
```

```
#!/usr/bin/perl
#
# DPREPRO: A Perl pre-processor for manipulating input files with DAKOTA.
# ____________________________________________________________________
#
# Copyright (c) 2001, Sandia National Laboratories.
# This software is distributed with DAKOTA under the GNU GPL.
# For more information, see the README file in the top Dakota directory.
#
# Usage: dprepro parameters_file template_input_file new_input_file
#
# Reads the variable tags and values from the parameters_file and then
# replaces each appearance of "{tag}" in the template_input_file with
# its associated value in order to create the new_input_file.  The
# parameters_file written by DAKOTA may either be in standard format
# (using "value tag" constructs) or in "aprepro" format (using
# "{ tag = value }" constructs), and the variable tags used inside
# template_input_file must match the variable descriptors specified in
# the DAKOTA input file.  Supports assignments and numerical expressions
# in the template file, and the parameters file takes precedence in
# the case of duplicate assignments (so that template file assignments
# can be treated as defaults to be overridden).
# ____________________________________________________________________

# Check for correct number of command line arguments and store the filenames.
if( @ARGV != 3 ) {
  print STDERR "Usage: dprepro parameters_file template_input_file ",
    "new_input_file\n";
  exit(-1);
}
$params_file   = $ARGV[0]; # DAKOTA parameters file (aprepro or standard format)
$template_file = $ARGV[1]; # template simulation input file
$new_file      = $ARGV[2]; # new simulation input file with insertions

# Regular expressions for numeric fields
$e  = "-?(?:\\d+\\.?\\d*|\\.\\d+)[eEdD](?:\\+|-)?\\d+"; # exponential notation
$f  = "-?\\d+\\.\\d*|-?\\.\\d+";                        # floating point
$i  = "-?\\d+";                                         # integer
$ui = "\\d+";                                           # unsigned integer
$n  = "$e|$f|$i";                                       # numeric field

################################
# Process DAKOTA parameters file
################################

# Open parameters file for input.
open (DAKOTA_PARAMS, "<$params_file") || die "Can't open $params_file: $!";
```

Figure 17.3: Partial listing of the `dprepro` script.

```
Title of Model: Rosenbrock black box
**************************************************************************
* Description:  This is an input file to the Rosenbrock black box
*               Fortran simulator.  This simulator is structured so
*               as to resemble the input/output from an engineering
*               simulation code, even though Rosenbrock's function
*               is a simple analytic function.  The node, element,
*               and material blocks are dummy inputs.
*
* Input:  x1 and x2
* Output: objective function value
**************************************************************************
node 1 location 0.0 0.0
node 2 location 0.0 1.0
node 3 location 1.0 0.0
node 4 location 1.0 1.0
node 5 location 2.0 0.0
node 6 location 2.0 1.0
node 7 location 3.0 0.0
node 8 location 3.0 1.0
element 1 nodes 1 3 4 2
element 2 nodes 3 5 6 4
element 3 nodes 5 7 8 6
element 4 nodes 7 9 10 8
material 1 elements 1 2
material 2 elements 3 4
variable 1 {x1}
variable 2 {x2}
end
```

Figure 17.4: Listing of the `ros.template` file

```
{ DVV_2           =                        2 }
{ DAKOTA_AN_COMPS =                        0 }
```

The first portion of the file indicates that there are two variables, followed by new values for variables `x1` and `x2`, and one response function (an objective function), followed by an active set vector (ASV) value of `1`. The ASV indicates the need to return the value of the objective function for these parameters (see Section 12.7). The `dprepro` script reads the variable values from this file, namely `1.638248083045767e-01` and `2.197300525769129e-02` for `x1` and `x2` respectively, and substitutes them in the {`x1`} and {`x2`} fields of the `ros.template` file. The final three lines of the resulting input file (`ros.in`) then appear as follows:

```
variable 1 1.638248083045767e-01
variable 2 2.197300525769129e-02
end
```

where all other lines are identical to the template file. The `rosenbrock_bb` simulator accepts `ros.in` as its input file and generates the following output to the file `ros.out`:

```
Beginning execution of model: Rosenbrock black box
Set up complete.
Reading nodes.
Reading elements.
Reading materials.
Checking connectivity...OK
****************************************************

Input value for x1 =  0.1638248083045767E+00
Input value for x2 =  0.2197300525769129E-01

Computing solution...Done
***************************************************
Function value =  0.7015563211077899E+00
```

Next, the appropriate data is extracted from the raw simulator output and returned in the results file. This post-processing is relatively trivial in this case, and the `simulator_script` uses the `grep` and `cut` utilities to extract the value from the last line of the `ros.out` output file and save it to `$argv[2]`, which is the `results.out.60` file for this evaluation. This single value provides the objective function value requested by the ASV.

After 132 of these function evaluations, the following DAKOTA output shows the final solution using the `rosenbrock_bb` simulator:

```
 Exit NPSOL - Optimal solution found.

 Final nonlinear objective value =   0.1165709E-06

NPSOL exits with INFORM code = 0 (see "Interpretation of output" section in NPSOL manual)

NOTE: see Fortran device 9 file (fort.9 or ftn09)
      for complete NPSOL iteration history.

<<<<< Iterator npsol_sqp completed.
<<<<< Function evaluation summary: 132 total (132 new, 0 duplicate)
<<<<< Best parameters          =
```

```
                      9.9965861597e-01 x1
                      9.9931682059e-01 x2
<<<<< Best objective function  =
                    1.1657092064e-07
<<<<< Best data captured at function evaluation 130
<<<<< Single Method Strategy completed.
DAKOTA execution time in seconds:
  Total CPU        =       0.12 [parent =   0.109984, child =   0.010016]
  Total wall clock =    4.73601
```

17.1.2 Adapting These Scripts to Another Simulation

To adapt this approach for use with another simulator, several steps need to be performed:

1. Create a template simulation input file by identifying the fields in an existing input file that correspond to the variables of interest and then replacing them with {} identifiers (e.g. {`cdv_1`}, {`cdv_2`}, etc.) which match the DAKOTA variable descriptors. Copy this template input file to a templatedir that will be used to create working directories for the simulation.

2. Modify the `dprepro` arguments in `simulator_script` to reflect names of the DAKOTA parameters file (previously "`$argv[1]`"), template file name (previously "`ros.template`") and generated input file (previously "`ros.in`"). Alternatively, use APREPRO, BPREPRO, or JPrePost to perform this step (and adapt the syntax accordingly).

3. Modify the analysis section of `simulator_script` to replace the `rosenbrock_bb` function call with the new simulator name and command line syntax (typically including the input and output file names).

4. Change the post-processing section in `simulator_script` to reflect the revised extraction process. At a minimum, this would involve changing the `grep` command to reflect the name of the output file, the string to search for, and the characters to cut out of the captured output line. For more involved post-processing tasks, invocation of additional tools may have to be added to the script.

5. Modify the `dakota_rosenbrock.in` input file to reflect, at a minimum, updated variables and responses specifications.

These nonintrusive interfacing approaches can be used to rapidly interface with simulation codes. While generally custom for each new application, typical interface development time is on the order of an hour or two. Thus, this approach is scalable when dealing with many different application codes. Weaknesses of this approach include the potential for loss of data precision (if care is not taken to preserve precision in pre- and post-processing file I/O), a lack of robustness in post-processing (if the data capture is too simplistic), and scripting overhead (only noticeable if the simulation time is on the order of a second or less).

If the application scope at a particular site is more focused and only a small number of simulation codes are of interest, then more sophisticated interfaces may be warranted. For example, the economy of scale afforded by a common simulation framework justifies additional effort in the development of a high quality DAKOTA interface. In these cases, more sophisticated interfacing approaches could involve a more thoroughly developed black box interface with robust support of a variety of inputs and outputs, or it might involve intrusive interfaces such as the direct simulation interface discussed below in Section 17.2 or the SAND interface described in Section 1.3.

17.1.3 Additional Examples

A variety of additional examples of black-box interfaces to simulation codes are maintained in the `Dakota/examples/script_interfaces` directory in the source code distribution.

17.2 Developing a Direct Simulation Interface

If a more efficient interface to a simulation is desired (e.g., to eliminate process creation and file I/O overhead) or if a targeted computer architecture cannot accommodate separate optimization and simulation processes (e.g., due to lightweight operating systems on compute nodes of large parallel computers), then linking a simulation code directly with DAKOTA may be desirable. This is an advanced capability of DAKOTA, and it requires a user to have access to (and knowledge of) the DAKOTA source code, as well as the source code of the simulation code.

Three approaches are outlined below for developing direct linking between DAKOTA and a simulation: extension, derivation, and sandwich. For additional information, refer to "Interfacing with DAKOTA as a Library" in the DAKOTA Developers Manual [31].

Once performed, DAKOTA can bind with the new direct simulation interface using the `direct` interface specification in combination with an `analysis_driver`, `input_filter` or `output_filter` specification that corresponds to the name of the new subroutine.

17.2.1 Extension

The first approach to using the direct function capability with a new simulation (or new internal test function) involves *extension* of the existing **DirectFnApplicInterface** class to include new simulation member functions. In this case, the following steps are performed:

1. The functions to be invoked (analysis programs, input and output filters, internal testers) must have their main programs changed into callable functions/subroutines.

2. The resulting callable function can then be added directly to the private member functions in **DirectFnApplicInterface** if this function will directly access the DAKOTA data structures (variables, active set, and response attributes of the class). It is more common to add a wrapper function to **DirectFnApplicInterface** which manages the DAKOTA data structures, but allows the simulator subroutine to retain a level of independence from DAKOTA (see Salinas, ModelCenter, and Matlab wrappers as examples).

3. The if-else blocks in the **derived_map_if()**, **derived_map_ac()**, and **derived_map_of()** member functions of the **DirectFnApplicInterface** class must be extended to include the new function names as options. If the new functions are class member functions, then DAKOTA data access may be performed through the existing class member attributes and data objects do not need to be passed through the function parameter list. In this case, the following function prototype is appropriate:

   ```
   int function_name();
   ```

   If, however, the new function names are not members of the **DirectFnApplicInterface** class, then an `extern` declaration may additionally be needed and the function prototype should include passing of the Variables, ActiveSet, and Response data members:

   ```
   int function_name(const Dakota::Variables& vars,
                     const Dakota::ActiveSet& set, Dakota::Response& response);
   ```

4. The DAKOTA system must be recompiled and linked with the new function object files or libraries.

Various header files may have to be included, particularly within the **DirectFnApplicInterface** class, in order to recognize new external functions and compile successfully. Refer to the DAKOTA Developers Manual [31] for additional information on the **DirectFnApplicInterface** class and the DAKOTA data types.

17.2.2 Derivation

As described in "Interfacing with DAKOTA as a Library" in the DAKOTA Developers Manual [31], a derivation approach can be employed to further increase the level of independence between DAKOTA and the host application. In this case, rather than *adding* a new function to the existing **DirectFnApplicInterface** class, a new interface class is derived from **DirectFnApplicInterface** which *redefines* the **derived_map_if()**, **derived_map_ac()**, and **derived_map_of()** virtual functions.

In the approach of Section 17.2.3 below, the class derivation approach avoids the need to recompile the DAKOTA library when the simulation or its direct interface class is modified.

17.2.3 Sandwich

In a "sandwich" implementation, a simulator provides both the "front end" and "back end" with DAKOTA sandwiched in the middle. To accomplish this approach, the simulation code is responsible for interacting with the user (the front end), links DAKOTA in as a library (refer to "Interfacing with DAKOTA as a Library" in the DAKOTA Developers Manual [31]), and plugs in a derived direct interface class to provide a closely-coupled mechanism for performing function evaluations (the back end). This approach makes DAKOTA services available to other codes and frameworks and is currently used by Sandia codes such as Xyce (electrical simulation), Sage (CFD), and SIERRA (multiphysics).

17.3 Existing Direct Interfaces to External Simulators

In addition to built-in polynomial test functions described in Section 13.3.1, DAKOTA includes direct interfaces to Sandia's Salinas code for structural dynamics, Phoenix Integration's ModelCenter framework, and The Mathworks' Matlab scientific computing environment.

17.3.1 Matlab

DAKOTA's direct function interface includes the capability to invoke Matlab for function evaluations, using the Matlab engine API. When using this close-coupling, the Matlab engine is started once when DAKOTA initializes, and then during analysis function evaluations are performed exchanging parameters and results through the Matlab C API. This eliminates the need to use the file system and the expense of initializing the Matlab engine for each function evaluation.

To use the DAKOTA/Matlab interface, DAKOTA must be configured and compiled with the Matlab feature enabled; see the INSTALL file included with the DAKOTA distribution for more information. The DAKOTA/Matlab interface has been built and tested on 32-bit Linux with Matlab 7.0 (R14) and on 64-bit Linux with Matlab 7.1 (R14SP3). Configuration support for Solaris is included, but is untested. Builds on other platforms or with other versions of Matlab may require modifications to the DAKOTA build system as described in INSTALL.

Since the Matlab libraries are linked dynamically, they must be accessible at run time. Before running a DAKOTA binary that includes Matlab support, make sure the path to the Matlab shared object libraries is included in your `LD_LIBRARY_PATH` environment variable. For example to accomplish this in BASH on 32-bit Linux, one might type

```
export LD_LIBRARY_PATH=/usr/local/matlab/bin/glnx86:$LD_LIBRARY_PATH
```

or add such a command to the .bashrc file.

Example files corresponding to the following tutorial are available in `Dakota/examples/linked_interfaces/Matlab`.

DAKOTA/Matlab input file specification

The use of the Matlab direct interface is specified through the combination of the `direct` and `analysis_driver` keywords in an interface specification. The Matlab m-file which performs the analysis is specified through the `analysis_components` keyword. Here is a sample DAKOTA `interface` specification:

```
interface,
  direct
    analysis_driver = 'matlab'
    analysis_components = 'myanalysis.m'
```

Multiple Matlab analysis drivers are supported and may be specified in combination with other direct analysis drivers. Multiple analysis components are supported as for other interfaces as described in Section 13.4. The `.m` extension in the `analysis_component` specification is optional; in fact any characters following the first period in the `analysis_component` specification will be stripped off by the interface, so `myanalysis`, `myanalysis.m`, and `myanalysis.fun.m` will all cause the interface to attempt to execute a Matlab script called `myanalysis.m` for the function evaluation.

Matlab .m file specification

The Matlab analysis file `myanalysis.m` must define a Matlab function that accepts a Matlab structure as its sole argument and returns the same structure in a variable called `Dakota`. A manual execution of the call to the analysis in Matlab should therefore look like:

```
>> Dakota = myanalysis(Dakota)
```

Note that the structure named Dakota will be pushed into the Matlab workspace before the analysis function is called. The structure passed from DAKOTA to the analysis m-function contains essentially the same information that would be passed to a DAKOTA direct function included in `DirectFnApplicInterface.C`, with fields shown in Figure 17.5.

The `directFnASM` field is a boolean matrix version of the active set vector, designed to make it easier for the Matlab user to determine which responses Dakota requires. Its rows 1, 2, and 3 correspond to functions, gradients, and Hessians, respectively, and its columns to each function (response, constraint) from 1–`numFns`. A `directFnASM` matrix entry contains a 1 if the response is needed for that function and a 0 otherwise. For example, for the fictitious

```
directFnASV = [ 1, 2, 3, 4, 5, 6, 7 ],
```

```
Dakota.
  numFns              number of functions (responses, constraints)
  numVars             total number of variables
  numACV              number active continuous variables
  numADV              number active discrete variables
  numDerivVars        number of derivative variables specified in directFnDVV
  xC                  continuous variable values ([1 x numACV])
  xD                  discrete variable values ([1 x numADV])
  xCLabels            continuous variable labels (cell array of numACV strings)
  xDLabels            discrete variable labels (cell array of  numADV strings)
  directFnASV         active set vector ([1 x numFns])
  directFnASM         active set matrix ([3 x numFns]) -- see below
  directFnDVV         derivative variables vector ([1 x numDerivVars])
  directFnDVV_bool    boolean DVV ([1 x numVars]) -- see below
  fnFlag              nonzero if function values requested
  gradFlag            nonzero if gradients requested
  hessFlag            nonzero if hessians requested
```

Figure 17.5: DAKOTA/Matlab parameter data structure.

the corresponding

```
directFnASM = [ 1, 0, 1, 0, 1, 0, 1     (functions)
                0, 1, 1, 0, 0, 1, 1     (gradients)
                0, 0 ,0 ,1, 1, 1, 1 ]   (hessians).
```

The `directFnDVV_bool` field is a similar boolean structure for the `directFnDVV`. It is a `[1 x numVars]` vector with entry 1 if a derivative with respect to that variable is requested and a 0 otherwise. So for an example with 6 functions if

```
directFnDVV = [ 1, 3, 4, 6 ],
```

the corresponding

```
directFnDVV_bool = [ 1, 0, 1, 1, 0, 1 ].
```

The structure `Dakota` returned from the analysis must contain a subset of the fields shown in Figure 17.6. It may contain additional fields and in fact is permitted to be the structure passed in, augmented with any required outputs.

An example Matlab analysis driver `rosenbrock.m` for the Rosenbrock function is shown in Figure 17.7.

```
Dakota.
  fnVals      ([1 x numFns], required if function values requested)
  fnGrads     ([numFns x numDerivVars], required if gradients  requested)
  fnHessians  ([numFns x numDerivVars x numDerivVars],
               required if hessians requested)
  fnLabels    (cell array of numFns strings, optional)
  failure     (optional: zero indicates success, nonzero failure
```

Figure 17.6: DAKOTA/Matlab response data structure.

```
function Dakota = rosenbrock(Dakota)

  Dakota.failure = 0;

  if ( Dakota.numVars ~= 2 | Dakota.numADV | ...
      ( ~isempty( find(Dakota.directFnASM(2,:)) | ...
      find(Dakota.directFnASM(3,:)) ) & Dakota.numDerivVars ~= 2 ) )

    sprintf('Error: Bad number of variables in rosenbrock.m fn.\n');
    Dakota.failure = 1;

  elseif (Dakota.numFns > 2)

    % 1 fn -> opt, 2 fns -> least sq
    sprintf('Error: Bad number of functions in rosenbrock.m fn.\n');
    Dakota.failure = 1;

  else

    if Dakota.numFns > 1
      least_sq_flag = true;
    else
      least_sq_flag = false;
    end

    f0 = Dakota.xC(2)-Dakota.xC(1)*Dakota.xC(1);
    f1 = 1.-Dakota.xC(1);

    % **** f:
    if (least_sq_flag)
      if Dakota.directFnASM(1,1)
        Dakota.fnVals(1) = 10*f0;
      end
      if Dakota.directFnASM(1,2)
        Dakota.fnVals(2) = f1;
      end
    else
      if Dakota.directFnASM(1,1)
        Dakota.fnVals(1) = 100.*f0*f0+f1*f1;
      end
    end

    % **** df/dx:
    if (least_sq_flag)
      if Dakota.directFnASM(2,1)
        Dakota.fnGrads(1,1) = -20.*Dakota.xC(1);
        Dakota.fnGrads(1,2) =  10.;
      end
      if Dakota.directFnASM(2,2)
        Dakota.fnGrads(2,1) = -1.;
        Dakota.fnGrads(2,2) =  0.;
      end

    else

      if Dakota.directFnASM(2,1)
        Dakota.fnGrads(1,1) = -400.*f0*Dakota.xC(1) - 2.*f1;
        Dakota.fnGrads(1,2) =  200.*f0;
      end

    end

    % **** d^2f/dx^2:
    if (least_sq_flag)

      if Dakota.directFnASM(3,1)
        Dakota.fnHessians(1,1,1) = -20.;
        Dakota.fnHessians(1,1,2) = 0.;
        Dakota.fnHessians(1,2,1) = 0.;
        Dakota.fnHessians(1,2,2) = 0.;
      end
      if Dakota.directFnASM(3,2)
        Dakota.fnHessians(2,1:2,1:2) = 0.;
      end

    else

      if Dakota.directFnASM(3,1)
        fx = Dakota.xC(2) - 3.*Dakota.xC(1)*Dakota.xC(1);
        Dakota.fnHessians(1,1,1) = -400.*fx + 2.0;
        Dakota.fnHessians(1,1,2) = -400.*Dakota.xC(1);
        Dakota.fnHessians(1,2,1) = -400.*Dakota.xC(1);
        Dakota.fnHessians(1,2,2) =  200.;
      end

    end

    Dakota.fnLabels = {'f1'};

  end
```

Figure 17.7: Sample Matlab implementation of the Rosenbrock test function for the DAKOTA/Matlab interface.

Chapter 18

Parallel Computing

18.1 Overview

This chapter describes the various parallel computing capabilities provided by DAKOTA. The range of capabilities is extensive and can be daunting at first; therefore, this chapter takes an incremental approach in first describing the simplest single-level parallel computing models (Section 18.2) using asynchronous local, message passing, and hybrid approaches. More advanced uses of DAKOTA can build on this foundation to exploit multiple levels of parallelism, as described in Section 18.3. The chapter concludes with a discussion of using DAKOTA with applications that run as independent MPI processes.

18.1.1 Categorization of parallelism

To understand the parallel computing possibilities, it is instructive to first categorize the opportunities for exploiting parallelism into four main areas [37], consisting of coarse-grained and fine-grained parallelism opportunities within algorithms and their function evaluations:

1. *Algorithmic coarse-grained parallelism*: This parallelism involves the concurrent execution of independent function evaluations, where a "function evaluation" is defined as a data request from an algorithm (which may involve value, gradient, and Hessian data from multiple objective and constraint functions). This concept can also be extended to the concurrent execution of multiple "iterators" within a "strategy." Examples of algorithms containing coarse-grained parallelism include:
   - *Gradient-based algorithms*: finite difference gradient evaluations, speculative optimization, parallel line search.
   - *Nongradient-based algorithms*: genetic algorithms (GAs), pattern search (PS), Monte Carlo sampling.
   - *Approximate methods*: design of computer experiments for building surrogate models.
   - *Concurrent-iterator strategies*: optimization under uncertainty, branch and bound, multi-start local search, Pareto set optimization, island-model GAs.
2. *Algorithmic fine-grained parallelism*: This involves computing the basic computational steps of an optimization algorithm (i.e., the internal linear algebra) in parallel. This is primarily of interest in large-scale optimization problems and simultaneous analysis and design (SAND).

3. *Function evaluation coarse-grained parallelism*: This involves concurrent computation of separable parts of a single function evaluation. This parallelism can be exploited when the evaluation of the response data set requires multiple independent simulations (e.g. multiple loading cases or operational environments) or multiple dependent analyses where the coupling is applied at the optimizer level (e.g., multiple disciplines in the individual discipline feasible formulation [21]).

4. *Function evaluation fine-grained parallelism*: This involves parallelization of the solution steps within a single analysis code. The DOE laboratories have developed parallel analysis codes in the areas of nonlinear mechanics, structural dynamics, heat transfer, computational fluid dynamics, shock physics, and many others.

By definition, coarse-grained parallelism requires very little inter-processor communication and is therefore "embarrassingly parallel," meaning that there is little loss in parallel efficiency due to communication as the number of processors increases. However, it is often the case that there are not enough separable computations on each algorithm cycle to utilize the thousands of processors available on MP machines. For example, a thermal safety application [38] demonstrated this limitation with a pattern search optimization in which the maximum speedup exploiting *only* coarse-grained algorithmic parallelism was shown to be limited by the size of the design problem (coordinate pattern search has at most $2n$ independent evaluations per cycle for n design variables).

Fine-grained parallelism, on the other hand, involves much more communication among processors and care must be taken to avoid the case of inefficient machine utilization in which the communication demands among processors outstrip the amount of actual computational work to be performed. For example, a chemically-reacting flow application [37] illustrated this limitation for a simulation of fixed size in which it was shown that, while simulation run time did monotonically decrease with increasing number of processors, the relative parallel efficiency $\hat{E}$ of the computation for fixed model size decreased rapidly (from $\hat{E} \approx 0.8$ at 64 processors to $\hat{E} \approx 0.4$ at 512 processors). This was due to the fact that the total amount of computation was approximately fixed, whereas the communication demands were increasing rapidly with increasing numbers of processors. Therefore, there is a practical limit on the number of processors that can be employed for fine-grained parallel simulation of a particular model size, and only for extreme model sizes can thousands of processors be efficiently utilized in studies exploiting fine-grained parallelism alone.

These limitations point us to the exploitation of multiple levels of parallelism, in particular the combination of coarse-grained and fine-grained approaches. This will allow us to execute fine-grained parallel simulations on sets of processors where they are most efficient and then replicate this efficiency with many coarse-grained instances. From a software perspective, coarse-grained parallelism by itself (many instances of a single-processor simulation) and fine-grained parallelism by itself (a single instance of a large multiprocessor simulation) can be considered to cover two ends of a spectrum, and we are interested in also supporting anywhere in between (any number of instances of any size simulation). Single-level parallelism approaches are described in Section 18.2, and multilevel parallelism approaches are discussed in Section 18.3.

The available concurrency in function evaluation parallelism is determined by the aspects of a particular systems analysis application, and is therefore highly application-dependent. Algorithmic parallelism, on the other hand, is largely determined by the selection and configuration of a particular algorithm. These selection possibilities within DAKOTA are outlined in the following section.

18.1.2 Parallel DAKOTA algorithms

In DAKOTA Version 4.2, the following parallel algorithms, comprised of iterators and strategies, provide support for coarse-grained algorithmic parallelism. Note that, even if a particular algorithm is serial in terms of its data request concurrency, other concurrency sources (e.g., function evaluation coarse-grained and fine-grained parallelism) may still be available.

Parallel iterators

- Gradient-based optimizers: CONMIN, DOT, NLPQL, NPSOL, and OPT++ can all exploit parallelism through the use of DAKOTA's native finite differencing routine (selected with `method_source dakota` in the responses specification), which will perform concurrent evaluations for each of the parameter offsets. For `n` variables, forward differences result in an $n+1$ concurrency and central differences result in a $2n+1$ concurrency. In addition, CONMIN, DOT, and OPT++ can use speculative gradient techniques [13] to obtain better parallel load balancing. By speculating that the gradient information associated with a given line search point will be used later and computing the gradient information in parallel at the same time as the function values, the concurrency during the gradient evaluation and line search phases can be balanced. NPSOL does not use speculative gradients since this approach is superseded by NPSOL's gradient-based line search in user-supplied derivative mode. NLPQL also supports a distributed line search capability for generating concurrency [117].

- Nongradient-based optimizers: APPSPACK, JEGA methods, and most COLINY methods support parallelism. Serial COLINY methods include Solis-Wets (`coliny_solis_wets`) and certain `exploratory_moves` options (`adaptive_pattern` and `multi_step`) in pattern search (`coliny_pattern_search`). OPT++ PDS (`optpp_pds`) and NCSU DIRECT (`ncsu_direct`) are also currently serial due to incompatibilities in DAKOTA and OPT++/NCSU parallelism models. Finally, `coliny_pattern_search` and `asynch_pattern_search` support dynamic job queues managed with nonblocking synchronization.

- Least squares methods: in an identical manner to the gradient-based optimizers, NL2SOL, NLSSOL, and Gauss-Newton can exploit parallelism through the use of DAKOTA's native finite differencing routine. In addition, NL2SOL and Gauss-Newton can use speculative gradient techniques to obtain better parallel load balancing. NLSSOL does not use speculative gradients since this approach is superseded by NLSSOL's gradient-based line search in user-supplied derivative mode.

- Surrogate-based minimizers: `surrogate_based_local`, `surrogate_based_global`, and `efficient_global` all support parallelism in the initial surrogate construction, but subsequent concurrency varies. In the case of `efficient_global`, only a single point is generated for evaluation for each subsequent cycle and there is no derivatove concurrency for this point. In the case of `surrogate_based_local`, only a single point is generated per subsequent cycle, but derivative concurrency for numerical gradient or Hessian evaluations may be available. And in the case of `surrogate_based_global`, multiple points may be generated on each subsequent cycle, depending on the multipoint return capability of specific minimizers.

- Parameter studies: all parameter study methods (`vector`, `list`, `centered`, and `multidim`) support parallelism. These methods avoid internal synchronization points, so all evaluations are available for concurrent execution.

- Design of experiments: all `dace` (`grid`, `random`, `oas`, `lhs`, `oa_lhs`, `box_behnken`, and `central_composite`), `fsu_quasi_mc` (`halton` and `hammersley`), `fsu_cvt`, and `psuade_moat` methods support parallelism.

- Uncertainty quantification: all nondeterministic methods (`nond_sampling`, `nond_local_reliability`, `nond_global_reliability`, `nond_polynomial_chaos`, `nond_stoch_collocation`, and `nond_evidence`) support parallelism. In the case of `nond_local_reliability`, gradient-based optimization is involved and parallelism can be exploited through the use of DAKOTA's native finite differencing routine. In the case of `nond_global_reliability`, EGRA methods support parallelism in the initial surrogate construction, but subsequently only generate a single point for evaluation per cycle.

Parallel strategies

Certain strategies support concurrency in multiple iterator executions. Currently, the strategies which can exploit this level of parallelism are:

- Hybrid minimization: when the sequential hybrid transfers multiple solution points between methods, single-point minimizers will be executed concurrently using each of the transferred solution points.
- Branch and bound: optimization strategy for mixed-integer nonlinear programming with noncategorical discrete variables.
- Pareto-set optimization: multiobjective optimization strategy for computing sets of points on the Pareto front of nondominated solutions.
- Multi-start iteration: strategy for executing multiple instances of an iterator from different starting points.

In the branch and bound case, the available iterator concurrency grows as the tree develops more branches, so some of the iterator servers may be idle in the initial phases. Similarly, hybrid minimization will display varying levels of iterator concurrency based on differing support of multipoint solution input/output between iterators; however, the use of multiple parallel configurations among the iterator sequence should prevent parallel inefficiencies. Finally, pareto-set and multi-start have a fixed set of jobs to perform and should exhibit good load balancing.

Parallel models

Parallelism support in model classes (see Chapter 11) is an important issue for variable scaling (see Section 7.3.2) and advanced model recursions such as surrogate-based minimization, optimization under uncertainty, and second-order probability (see Chapter 9 and Section 11.6). Support is as follows:

- Single model: parallelism is managed by the singe interface instance.
- Recast model: most parallelism is forwarded on to the sub-model. An exception to this is finite differencing in the presence of variable scaling. Since it is desirable to perform offsets in the scaled space (and avoid minimum step size tolerances), this parallelism is not forwarded to the sub-model, instead being enacted at the recast level.
- Data fit surrogate model: parallelism is supported in the construction of global surrogate models via the concurrent evaluation of points generated by design of experiments methods. Local and multipoint approximations evaluate only a single point at a time, so concurrency is available only from any numerical differencing required for gradient and Hessian data. Since the top-level iterator is interfaced only with the (inexpensive) surrogate, no parallelism is exploited here. Load balancing can be an important issue when performing evaluations for update of existing surrogate models.
- Hierarchical surrogate model: parallelism is supported for both the low and high fidelity models. Since the top-level iterator is interfaced only with the low-fidelity model, and the high-fidelity model is used only for verifications and correction updating, the algorithmic coarse-grained parallelism supported by the iterator is enacted on the low fidelity model and the only parallelism available for high fidelity executions arises from any numerical differencing required for high-fidelity gradient and Hessian data.
- Nested model: Currently, nested models only support concurrency in the sub-iterator execution on the sub-model. This means that the top-level iterator interfaced with the nested model is serialized. In future releases, concurrency will be supported at this top level, allowing techniques such as optimization under uncertainty and second-order probability (see Section 11.5) to support concurrent iterator parallelism.

18.2 Single-level parallelism

DAKOTA's parallel facilities support a broad range of computing hardware, from custom massively parallel supercomputers on the high end, to clusters and networks of workstations (NOWs) in the middle range, to desktop multiprocessors on the low end. Given the reduced scale in the middle to low ranges, it is more common to exploit only one of the levels of parallelism; however, this can still be quite effective in reducing the time to obtain a solution. Three single-level parallelism models will be discussed, and are depicted in Figure 18.1:

- *asynchronous local*: DAKOTA executes on a single processor, but launches multiple jobs concurrently using asynchronous job launching techniques.
- *message passing*: DAKOTA executes in parallel using message passing to communicate between processors. A single job is launched per processor using synchronous job launching techniques.
- *hybrid*: a combination of message passing and asynchronous local. DAKOTA executes in parallel across multiple processors and launches concurrent jobs on each processor.

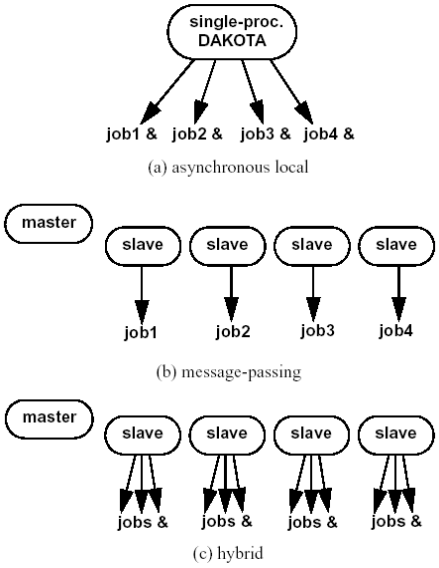


Figure 18.1: External, internal, and hybrid job management.

In each of these cases, jobs are executing concurrently and must be collected in some manner for return to an algorithm. Blocking and nonblocking approaches are provided for this, where the blocking approach is used in most cases:

- *blocking synchronization*: all jobs in the queue are completed before exiting the scheduler and returning the set of results to the algorithm. The job queue fills and then empties completely, which provides a synchronization point for the algorithm.
- *nonblocking synchronization*: the job queue is dynamic, with jobs entering and leaving continuously. There are no defined synchronization points for the algorithm, which requires specialized algorithm logic (only currently supported by `coliny_pattern_search` and `asynch_pattern_search`, which are sometimes referred to as "fully asynchronous" algorithms).

Given these job management capabilities, it is worth noting that the popular term "asynchronous" can be ambiguous when used in isolation. In particular, it can be important to qualify whether one is referring to "asynchronous job launch" (synonymous with any of the three concurrent job launch approaches described above) or "asynchronous job recovery" (synonymous with the latter nonblocking job synchronization approach).

18.2.1 Asynchronous Local Parallelism

This section describes software components which manage simulation invocations local to a processor. These invocations may be either synchronous (i.e., blocking) or asynchronous (i.e., nonblocking). Synchronous evaluations proceed one at a time with the evaluation running to completion before control is returned to DAKOTA.

Asynchronous evaluations are initiated such that control is returned to DAKOTA immediately, prior to evaluation completion, thereby allowing the initiation of additional evaluations which will execute concurrently.

The synchronous local invocation capabilities are used in two contexts: (1) by themselves to provide serial execution on a single processor, and (2) in combination with DAKOTA's message-passing schedulers to provide function evaluations local to each processor. Similarly, the asynchronous local invocation capabilities are used in two contexts: (1) by themselves to launch concurrent jobs from a single processor that rely on external means (e.g., operating system, job queues) for assignment to other processors, and (2) in combination with DAKOTA's message-passing schedulers to provide a hybrid parallelism (see Section 18.2.3). Thus, DAKOTA supports any of the four combinations of synchronous or asynchronous local combined with message passing or without.

Asynchronous local schedulers may be used for managing concurrent function evaluations requested by an iterator or for managing concurrent analyses within each function evaluation. The former iterator/evaluation concurrency supports either blocking (all jobs in the queue must be completed by the scheduler) or nonblocking (dynamic job queue may shrink or expand) synchronization, where blocking synchronization is used by most iterators and nonblocking synchronization is used by fully asynchronous algorithms such as `asynch_pattern_search` and `coliny_pattern_search`. The latter evaluation/analysis concurrency is restricted to blocking synchronization. The "Asynchronous Local" column in Table 18.1 summarizes these capabilities.

DAKOTA supports three local simulation invocation approaches based on the direct function, system call, and fork simulation interfaces. For each of these cases, an input filter, one or more analysis drivers, and an output filter make up the interface, as described in Section 13.4.

Direct function synchronization

The direct function capability may be used synchronously. Synchronous operation of the direct function simulation interface involves a standard procedure call to the input filter, if present, followed by calls to one or more simulations, followed by a call to the output filter, if present (refer to Sections 13.3-13.4 for additional details and examples). Each of these components must be linked as functions within DAKOTA. Control does not return to the calling code until the evaluation is completed and the response object has been populated.

Asynchronous operation will be supported in the future and will involve the use of multithreading (e.g., POSIX threads) to accomplish multiple simultaneous simulations. When spawning a thread (e.g., using `pthread_create`), control returns to the calling code after the simulation is initiated. In this way, multiple threads can be created simultaneously. An array of responses corresponding to the multiple threads of execution would then be recovered in a synchronize operation (e.g., using `pthread_join`).

System call synchronization

The system call capability may be used synchronously or asynchronously. In both cases, the `system` utility from the standard C library is used. Synchronous operation of the system call simulation interface involves spawning the system call (containing the filters and analysis drivers bound together with parentheses and semi-colons) in the foreground. Control does not return to the calling code until the simulation is completed and the response file has been written. In this case, the possibility of a race condition (see below) does not exist and any errors during response recovery will cause an immediate abort of the DAKOTA process (note: detection of the string "fail" is not a response recovery error; see Chapter 21).

Asynchronous operation involves spawning the system call in the background, continuing with other tasks (e.g., spawning other system calls), periodically checking for process completion, and finally retrieving the results. An array of responses corresponding to the multiple system calls is recovered in a synchronize operation.

In this synchronize operation, completion of a function evaluation is detected by testing for the existence of the evaluation's results file using the `stat` utility [83]. Care must be taken when using asynchronous system calls since they are prone to the race condition in which the results file passes the existence test but the recording of the function evaluation results in the file is incomplete. In this case, the read operation performed by DAKOTA will result in an error due to an incomplete data set. In order to address this problem, DAKOTA contains exception handling which allows for a fixed number of response read failures per asynchronous system call evaluation. The number of allowed failures must have a limit, so that an actual response format error (unrelated to the race condition) will eventually abort the system. Therefore, to reduce the possibility of exceeding the limit on allowable read failures, *the user's interface should minimize the amount of time an incomplete results file exists in the directory where its status is being tested*. This can be accomplished through two approaches: (1) delay the creation of the results file until the simulation computations are complete and all of the response data is ready to be written to the results file, or (2) perform the simulation computations in a subdirectory, and as a last step, move the completed results file into the main working directory where its existence is being queried.

If concurrent simulations are executing in a shared disk space, then care must be taken to maintain independence of the simulations. In particular, the parameters and results files used to communicate between DAKOTA and the simulation, as well as any other files used by this simulation, must be protected from other files of the same name used by the other concurrent simulations. With respect to the parameters and results files, these files may be made unique through the use of the `file_tag` option (e.g., `params.in.1`, `results.out.1`, etc.) or the default UNIX temporary file option (e.g., `/var/tmp/aaa0b2Mfv`, etc.). However, if additional simulation files must be protected (e.g., `model.i`, `model.o`, `model.g`, `model.e`, etc.), then an effective approach is to create a tagged working subdirectory for each simulation instance. Section 17.1 provides an example system call interface that demonstrates both the use of tagged working directories and the relocation of completed results files to avoid the race condition.

Fork synchronization

The fork capability is quite similar to the system call; however, it has the advantage that asynchronous fork invocations can avoid the results file race condition that may occur with asynchronous system calls (see Section 13.3.5). The fork interface invokes the filters and analysis drivers using the `fork` and `exec` family of functions, and completion of these processes is detected using the `wait` family of functions. Since `wait` is based on a process id handle rather than a file existence test, an incomplete results file is not an issue.

Depending on the platform, the fork simulation interface executes either a `vfork` or a `fork` call. These calls generate a new child process with its own UNIX process identification number, which functions as a copy of the parent process (dakota). The `execvp` function is then called by the child process, causing it to be replaced by the analysis driver or filter. For synchronous operation, the parent dakota process then awaits completion of the forked child process through a blocking call to `waitpid`. On most platforms, the `fork/exec` procedure is efficient since it operates in a copy-on-write mode, and no copy of the parent is actually created. Instead, the parents address space is borrowed until the `exec` function is called.

The `fork/exec` behavior for asynchronous operation is similar to that for synchronous operation, the only difference being that dakota invokes multiple simulations through the `fork/exec` procedure prior to recovering response results for these jobs using the `wait` function. The combined use of `fork/exec` and `wait` functions in asynchronous mode allows the scheduling of a specified number of concurrent function evaluations and/or concurrent analyses.

Asynchronous Local Example

The test file `Dakota/test/dakota_dace.in` computes 49 orthogonal array samples, which may be evaluated concurrently using parallel computing. When executing DAKOTA with this input file on a single processor, the following execution syntax may be used:

```
dakota -i dakota_dace.in
```

For serial execution (the default), the interface specification within `dakota_dace.in` would appear similar to

```
interface,
        system
          analysis_driver = 'text_book'
```

which results in function evaluation output similar to the following (for `output` set to `quiet` mode):

```
>>>>> Running dace iterator.

------------------------------
Begin Function Evaluation    1
------------------------------
(text_book /tmp/fileG32LEp /tmp/fileP8uYDC)

------------------------------
Begin Function Evaluation    2
------------------------------
(text_book /tmp/fileiqIEEP /tmp/fileBEFlF2)

<snip>

------------------------------
Begin Function Evaluation   49
------------------------------
(text_book /tmp/file4Xyp2p /tmp/filezCohcE)

<<<<< Iterator dace completed.
```

where it is evident that each function evaluation is being performed sequentially.

For parallel execution using asynchronous local approaches, the DAKOTA execution syntax is unchanged as DAKOTA is still launched on a single processor. However, the interface specification is augmented to include the `asynchronous` keyword with optional concurrency limiter to indicate that multiple `analysis_driver` instances will be executed concurrently:

```
interface,
        system asynchronous evaluation_concurrency = 4
          analysis_driver = 'text_book'
```

which results in output excerpts similar to the following:

```
>>>>> Running dace iterator.

------------------------------
```

```
Begin Function Evaluation    1
------------------------------
(Asynchronous job 1 added to queue)

------------------------------
Begin Function Evaluation    2
------------------------------
(Asynchronous job 2 added to queue)

<snip>

------------------------------
Begin Function Evaluation   49
------------------------------
(Asynchronous job 49 added to queue)

Blocking synchronize of 49 asynchronous evaluations
First pass: initiating 4 asynchronous jobs
Initiating function evaluation 1
(text_book /tmp/fileG2uzVX /tmp/fileSqceY8) &
Initiating function evaluation 2
(text_book /tmp/filegFLu5j /tmp/fileeycMcv) &
Initiating function evaluation 3
(text_book /tmp/file8EI3kG /tmp/fileuY2ltR) &
Initiating function evaluation 4
(text_book /tmp/fileEZpDC2 /tmp/fileeMDVLd) &
Second pass: self-scheduling 45 remaining jobs
Waiting on completed jobs
Function evaluation 1 has completed
Initiating function evaluation 5
(text_book /tmp/file8SWrXo /tmp/filem00Y8z) &
Function evaluation 2 has completed
Initiating function evaluation 6
(text_book /tmp/file6PQ5kL /tmp/filegRydxW) &
Function evaluation 3 has completed
Initiating function evaluation 7
(text_book /tmp/filesjB8J7 /tmp/fileUpr4Wi) &
Function evaluation 4 has completed
Initiating function evaluation 8
(text_book /tmp/fileCI6Bbu /tmp/fileWSBaqF) &

<snip>

Function evaluation 49 has completed

<<<<< Iterator dace completed.
```

where it is evident that each of the 49 jobs is first queued and then a blocking synchronization is performed. This synchronization uses a simple scheduler that initiates 4 jobs and then replaces completing jobs with new ones until all 49 are complete.

The default job concurrency for asynchronous local parallelism is all that is available from the algorithm (49 in this case), which could be too many for the computational resources or their usage policies. The concurrency level specification (4 in this case) instructs the scheduler to keep 4 jobs running concurrently, which would be

appropriate for, e.g., a dual-processor dual-core workstation. In this case, it is the operating system's responsibility to assign the concurrent `text_book` jobs to available processors/cores. Specifying greater concurrency than that supported by the hardware will result in additional context switching within a multitasking operating system and will generally degrade performance. Note however that, in this example, there are a total of 5 processes running, one for DAKOTA and four for the concurrent function evaluations. Since the DAKOTA process checks periodically for job completion and sleeps in between checks, it is relatively lightweight and does not require a dedicated processor.

18.2.2 Message Passing Parallelism

DAKOTA uses a "single program-multiple data" (SPMD) parallel programming model. It uses message-passing routines from the Message Passing Interface (MPI) standard [68], [122] to communicate data between processors. The SPMD designation simply denotes that the same DAKOTA executable is loaded on all processors during the parallel invocation. This differs from the MPMD model ("multiple program-multiple data") which would have the DAKOTA executable on one or more processors communicating directly with simulator executables on other processors. The MPMD model has some advantages, but heterogeneous executable loads are not supported by all parallel environments. Moreover, the MPMD model requires simulation code intrusion on the same order as conversion to a subroutine, so subroutine conversion (see Section 17.2) in a direct-linked SPMD model is preferred.

Partitioning

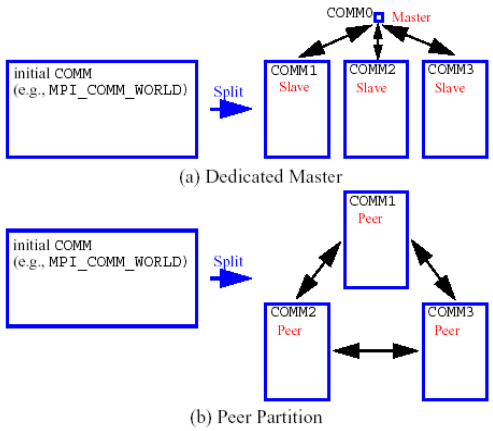


Figure 18.2: Communicator partitioning models.

A level of message passing parallelism can use either of two processor partitioning models:

- *Dedicated master*: a single processor is dedicated to scheduling operations and the remaining processors are split into server partitions.
- *Peer partition*: all processors are allocated to server partitions and the loss of a processor to scheduling is avoided.

These models are depicted in Figure 18.2. The peer partition is desirable since it utilizes all processors for computation; however, it requires either the use of sophisticated mechanisms for distributed scheduling or a problem for which static scheduling of concurrent work performs well (see *Scheduling* below). If neither of these characteristics is present, then use of the dedicated master partition supports a dynamic scheduling which assures that server idleness is minimized.

Scheduling

The following scheduling approaches are available within a level of message passing parallelism:

- *Self-scheduling*: in the dedicated master model, the master processor manages a single processing queue and maintains a prescribed number of jobs (usually one) active on each slave. Once a slave server has

completed a job and returned its results, the master assigns the next job to this slave. Thus, the slaves themselves determine the schedule through their job completion speed. This provides a simple dynamic scheduler in that heterogeneous processor speeds and/or job durations are naturally handled, provided there are sufficient instances scheduled through the servers to balance the variation.

- *Static scheduling*: if scheduling is statically determined at start-up, then no master processor is needed to direct traffic and a peer partitioning approach is applicable. If the static schedule is a good one (ideal conditions), then this approach will have superior performance. However, heterogeneity, when not known *a priori*, can very quickly degrade performance since there is no mechanism to adapt.

In addition, the following scheduling approach is provided by PICO for the scheduling of concurrent optimizations within the branch and bound strategy:

- *Distributed scheduling*: in this approach, a peer partition is used and each peer maintains a separate queue of pending jobs. When one peer's queue is smaller than the other queues, it requests work from its peers (prior to idleness). In this way, it can adapt to heterogeneous conditions, provided there are sufficient instances to balance the variation. Each partition performs communication between computations, and no processors are dedicated to scheduling. Furthermore, it distributes scheduling load beyond a single processor, which can be important for large numbers of concurrent jobs (whose scheduling might overload a single master) or for fault tolerance (avoiding a single point of failure). However, it involves relatively complicated logic and additional communication for queue status and job migration, and its performance is not always superior since a partition can become work-starved if its peers are locked in computation (Note: this logic can be somewhat simplified if a separate thread can be created for communication and migration of jobs).

Message passing schedulers may be used for managing concurrent iterator executions within a strategy, concurrent evaluations within an iterator, or concurrent analyses within an evaluation. In each of these cases, the message passing scheduler is currently restricted to blocking synchronization, in that all jobs in the queue are completed before exiting the scheduler and returning the set of results to the algorithm. Nonblocking message-passing schedulers are under development for the iterator/evaluation concurrency level in support of fully asynchronous algorithms which do not contain synchronization points (e.g., `asynch_pattern_search` and `coliny_pattern_search`). Message passing is also used within a fine-grained parallel analysis code, although this does not involve the use of DAKOTA schedulers (DAKOTA may, at most, pass a communicator partition to the simulation). The "Message Passing" column in Table 18.1 summarizes these capabilities.

Message Passing Example

Revisiting the test file `dakota_dace.in`, DAKOTA will now compute the 49 orthogonal array samples using a message passing approach. In this case, a parallel launch utility is used to execute DAKOTA across multiple processors using syntax similar to the following:

```
mpirun -np 5 -machinefile machines dakota -i dakota_dace.in
```

Since the asynchronous local parallelism will not be used, the interface specification does not include the `asynchronous` keyword and would appear similar to:

```
interface,
        system
          analysis_driver = 'text_book'
```

The relevant excerpts from the DAKOTA output for a dedicated master partition and self-schedule, the default when the maximum concurrency (49) exceeds the available capacity (5), would appear similar to the following:

```
Running MPI executable in parallel on 5 processors.

--------------------------------------------------------------------------
DAKOTA parallel configuration:

Level                   num_servers    procs_per_server    partition/schedule
-----                   -----------    ----------------    ------------------
concurrent iterators         1               5              peer/static
concurrent evaluations       4               1              ded. master/self
concurrent analyses          1               1              peer/static
multiprocessor analysis      1              N/A                N/A

Total parallelism levels =   1
--------------------------------------------------------------------------

>>>>> Running dace iterator.

------------------------------
Begin Function Evaluation    1
------------------------------
(Asynchronous job 1 added to queue)

------------------------------
Begin Function Evaluation    2
------------------------------
(Asynchronous job 2 added to queue)

<snip>

------------------------------
Begin Function Evaluation   49
------------------------------
(Asynchronous job 49 added to queue)

Blocking synchronize of 49 asynchronous evaluations
First pass: assigning 4 jobs among 4 servers
Master assigning function evaluation 1 to server 1
Master assigning function evaluation 2 to server 2
Master assigning function evaluation 3 to server 3
Master assigning function evaluation 4 to server 4
Second pass: self-scheduling 45 remaining jobs
Waiting on completed jobs
job 1 has returned from slave server 1
Master assigning function evaluation 5 to server 1
job 2 has returned from slave server 2
Master assigning function evaluation 6 to server 2
Waiting on completed jobs
job 3 has returned from slave server 3
Master assigning function evaluation 7 to server 3
job 4 has returned from slave server 4
Master assigning function evaluation 8 to server 4
```

```
<snip>

job 49 has returned from slave server 2

<<<<< Iterator dace completed.
```

where it is evident that each of the 49 jobs is first queued and then a blocking synchronization is performed. This synchronization uses a dynamic scheduler that initiates four jobs by sending a message from the master to each of the four servers and then replaces completing jobs with new ones until all 49 are complete. It is important to note that job execution local to each of the four servers is synchronous.

18.2.3 Hybrid Parallelism

The asynchronous local approaches described in Section 18.2.1 can be considered to rely on *external* scheduling mechanisms, since it is generally the operating system or some external queue/load sharing software that allocates jobs to processors. Conversely, the message-passing approaches described in Section 18.2.2 rely on *internal* scheduling mechanisms to distribute work among processors. These two approaches provide building blocks which can be combined in a variety of ways to manage parallelism at multiple levels. At one extreme, DAKOTA can execute on a single processor and rely completely on external means to map all jobs to processors (i.e., using asynchronous local approaches). At the other extreme, DAKOTA can execute on many processors and manage all levels of parallelism, including the parallel simulations, using completely internal approaches (i.e., using message passing at all levels as in Figure 18.4). While all-internal or all-external approaches are common cases, many additional approaches exist between the two extremes in which some parallelism is managed internally and some is managed externally.

These combined approaches are referred to as *hybrid* parallelism, since the internal distribution of work based on message-passing is being combined with external allocation using asynchronous local approaches[1]. Figure 18.1 depicts the asynchronous local, message-passing, and hybrid approaches for a dedicated-master partition. Approaches (b) and (c) both use MPI message-passing to distribute work from the master to the slaves, and approaches (a) and (c) both manage asynchronous jobs local to a processor. The hybrid approach (c) can be seen to be a combination of (a) and (b) since jobs are being internally distributed to slave servers through message-passing and each slave server is managing multiple concurrent jobs using an asynchronous local approach. From a different perspective, one could consider (a) and (b) to be special cases within the range of configurations supported by (c). The hybrid approach is useful for supercomputers that maintain a service/compute node distinction and for supercomputers or networks of workstations that involve clusters of symmetric multiprocessors (SMPs). In the service/compute node case, concurrent multiprocessor simulations are launched into the compute nodes from the service node partition. While an asynchronous local approach from a single service node would be sufficient, spreading the application load by running DAKOTA in parallel across multiple service nodes results in better performance [39]. If the number of concurrent jobs to be managed in the compute partition exceeds the number of available service nodes, then hybrid parallelism is the preferred approach. In the case of a cluster of SMPs (or network of multiprocessor workstations), message-passing can be used to communicate between SMPs, and asynchronous local approaches can be used within an SMP. Hybrid parallelism can again result in improved performance, since the total number of DAKOTA MPI processes is reduced in comparison to a pure message-passing approach over all processors.

Hybrid schedulers may be used for managing concurrent evaluations within an iterator or concurrent analyses

[1] The term "hybrid parallelism" is often used to describe the combination of MPI message passing and OpenMP shared memory parallelism models. This can be considered to be a special case of the meaning here, as OpenMP is based on threads, which is analagous to asynchronous local usage of the direct simulation interface.

within an evaluation. In both of these cases, the scheduler is currently restricted to blocking synchronization, although as for message-passing schedulers described in Section 18.2.2, nonblocking schedulers are under development for the iterator/evaluation concurrency level. The "Hybrid" column in Table 18.1 summarizes these capabilities.

Hybrid Example

Revisiting the test file `dakota_dace.in`, DAKOTA will now compute the 49 orthogonal array samples using a hybrid approach. As for the message passing case, a parallel launch utility is used to execute DAKOTA across multiple processors:

```
mpirun -np 5 -machinefile machines dakota -i dakota_dace.in
```

Since the asynchronous local parallelism will also be used, the interface specification includes the `asynchronous` keyword and appears similar to

```
interface,
        system asynchronous evaluation_concurrency = 2
          analysis_driver = 'text_book'
```

In the hybrid case, the specification of the desired concurrency level must be included, since the default is no longer all available (as it is for asynchronous local parallelism). Rather the default is to employ message passing parallelism, and hybrid parallelism is only available through the specification of asynchronous concurrency greater than one.

The relevant excerpts of the DAKOTA output for a dedicated master partition and self schedule, the default when the maximum concurrency (49) exceeds the maximum available capacity (10), would appear similar to the following:

```
Running MPI executable in parallel on 5 processors.

--------------------------------------------------------------------------
DAKOTA parallel configuration:

Level                   num_servers    procs_per_server    partition/schedule
-----                   -----------    ----------------    ------------------
concurrent iterators         1               5              peer/static
concurrent evaluations       4               1              ded. master/self
concurrent analyses          1               1              peer/static
multiprocessor analysis      1              N/A                N/A

Total parallelism levels =   1
--------------------------------------------------------------------------

>>>>> Running dace iterator.

------------------------------
Begin Function Evaluation    1
------------------------------
(Asynchronous job 1 added to queue)

------------------------------
```

```
Begin Function Evaluation    2
------------------------------
(Asynchronous job 2 added to queue)

<snip>

------------------------------
Begin Function Evaluation   49
------------------------------
(Asynchronous job 49 added to queue)

Blocking synchronize of 49 asynchronous evaluations
First pass: assigning 8 jobs among 4 servers
Master assigning function evaluation 1 to server 1
Master assigning function evaluation 2 to server 2
Master assigning function evaluation 3 to server 3
Master assigning function evaluation 4 to server 4
Master assigning function evaluation 5 to server 1
Master assigning function evaluation 6 to server 2
Master assigning function evaluation 7 to server 3
Master assigning function evaluation 8 to server 4
Second pass: self-scheduling 41 remaining jobs
Waiting on completed jobs

<snip>

job 49 has returned from slave server 4

<<<<< Iterator dace completed.
```

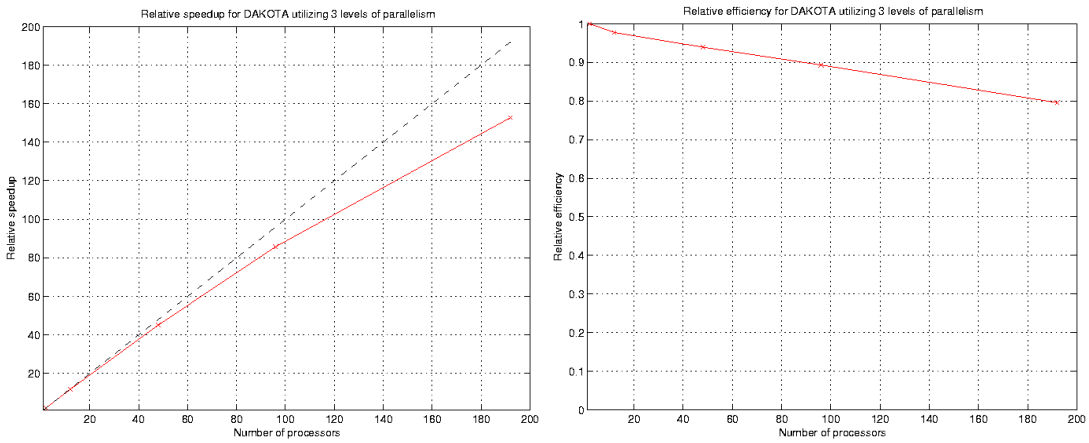
where it is evident that each of the 49 jobs is first queued and then a blocking synchronization is performed. This synchronization uses a dynamic scheduler that initiates eight jobs by sending two messages to each of the four servers and then replaces completing jobs with new ones until all 49 are complete. It is important to note that job execution local to each of the four servers is asynchronous. If the available capacity was increased to meet or exceed the maximum concurrency (e.g., mpirun on 10 processors with `evaluation_concurrency = 5`), then a peer partition with static schedule would be selected by default.

18.3 Multilevel parallelism

Parallel computers within the Department of Energy national laboratories have exceeded a hundred trillion floating point operations per second (100 TeraFLOPS) in Linpack benchmarks and are expected to achieve PetaFLOPS speeds in the near future. This performance is achieved through the use of massively parallel (MP) processing using $O[10^3 - 10^4]$ processors. In order to harness the power of these machines for performing design, uncertainty quantification, and other systems analyses, parallel algorithms are needed which are scalable to thousands of processors.

DAKOTA supports a total of three tiers of scheduling and four levels of parallelism which, in combination, can minimize efficiency losses and achieve near linear scaling on MP computers. The four levels are:

1. concurrent iterators within a strategy (scheduling performed by DAKOTA)
2. concurrent function evaluations within each iterator (scheduling performed by DAKOTA)



(a) Relative speedup.

(b) Relative efficiency.

Figure 18.3: Fixed-size scaling results for three levels of parallelism.

3. concurrent analyses within each function evaluation (scheduling performed by DAKOTA)

4. multiprocessor analyses (work distributed by a parallel analysis code)

for which the first two are classified as algorithmic coarse-grained parallelism, the third is function evaluation coarse-grained parallelism, and the fourth is function evaluation fine-grained parallelism (see Section 18.1.1). Algorithmic fine-grained parallelism is not currently supported, although the development of large-scale parallel SAND techniques is a current research direction [8].

A particular application may support one or more of these parallelism types, and DAKOTA provides for convenient selection and combination of each of the supported levels. If multiple types of parallelism can be exploited, then the question may arise as to how the amount of parallelism at each level should be selected so as to maximize the overall parallel efficiency of the study. For performance analysis of multilevel parallelism formulations and detailed discussion of these issues, refer to [39]. In many cases, *the user may simply employ DAKOTA's automatic parallelism configuration facilities*, which implement the recommendations from the aforementioned paper.

Figure 18.3 shows typical fixed-size scaling performance using a modified version of the extended `text_book` problem (see Section 22.1). Three levels of parallelism (concurrent evaluations within an iterator, concurrent analyses within each evaluation, and multiprocessor analyses) are exercised. Despite the use of a fixed problem size and the presence of some idleness within the scheduling at multiple levels, the efficiency is still reasonably high[2]. Greater efficiencies are obtainable for scaled speedup studies (or for larger problems in fixed-size studies) and for problems optimized for minimal scheduler idleness (by, e.g., managing all concurrency in as few scheduling levels as possible). Note that speedup and efficiency are measured relative to the case of a single instance of a multiprocessor analysis, since it was desired to investigate the effectiveness of the DAKOTA schedulers independent from the efficiency of the parallel analysis.

[2]Note that overhead is reduced in these scaling studies by deactivating the evaluation cache and restart file logging.

18.3.1 Asynchronous Local Parallelism

In most cases, the use of asynchronous local parallelism is the termination point for multilevel parallelism, in that any level of parallelism lower than an asynchronous local level will be serialized. The exception to this rule is reforking of forked processes for concurrent analyses within forked evaluations. In this case, a new process is created using fork for one of several concurrent evaluations; however, the new process is not replaced immediately using exec. Rather, the new process is reforked to create additional child processes for executing concurrent analyses within each concurrent evaluation process. This capability is not supported by system calls and provides one of the key advantages to using fork over system (see Section 13.3.5).

18.3.2 Message Passing Parallelism

Partitioning of levels

DAKOTA uses MPI communicators to identify groups of processors. The global `MPI_COMM_WORLD` communicator provides the total set of processors allocated to the DAKOTA run. `MPI_COMM_WORLD` can be partitioned into new intra-communicators which each define a set of processors to be used for a multiprocessor server. Each of these servers may be further partitioned to nest one level of parallelism within the next. At the lowest parallelism level, these intra-communicators can be passed into a simulation for use as the simulation's computational context, provided that the simulation has been designed, or can be modified, to be modular on a communicator (i.e., it does not assume ownership of `MPI_COMM_WORLD`). New intra-communicators are created with the `MPI_Comm_split` routine, and in order to send messages between these intra-communicators, new inter-communicators are created with calls to `MPI_Intercomm_create`. To minimize overhead, DAKOTA creates new intra- and inter-communicators only when the parent communicator provides insufficient context for the scheduling at a particular level. In addition, multiple parallel configurations (containing a set of communicator partitions) can be allocated for use in studies with multiple iterators and models (e.g., 16 servers of 64 processors each could be used for iteration on a lower fidelity model, followed by two servers of 512 processors each for subsequent iteration on a higher fidelity model). Each of the parallel configurations are allocated at object construction time and are reported at the beginning of the DAKOTA output.

Each tier within DAKOTA's nested parallelism hierarchy can use the dedicated master and peer partition approaches described in Section 18.2.2. To recursively partition the subcommunicators of Figure 18.2, `COMM1/2/3` in the dedicated master or peer partition case would be further subdivided using the appropriate partitioning model for the next lower level of parallelism.

Scheduling within levels

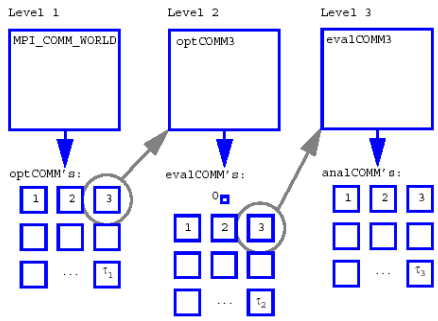


Figure 18.4: Recursive partitioning for nested parallelism.

DAKOTA is designed to allow the freedom to configure each parallelism level with either the dedicated master partition/self-scheduling combination or the peer partition/static scheduling combination. In addition, certain external libraries may provide additional options (e.g., PICO supports distributed scheduling in peer partitions). As an example, Figure 18.4 shows a case in which a branch and bound strategy employs peer partition/distributed scheduling at level 1, each optimizer partition employs concurrent function evaluations in a dedicated master partition/self-scheduling model at level 2, and each function evaluation partition employs concurrent multiprocessor analyses in a peer partition/static scheduling

model at level 3. In this case, `MPI_COMM_WORLD` is subdivided into `optCOMM1/2/3/.../`τ_1, each `optCOMM` is further subdivided into `evalCOMM0` (master) and `evalCOMM1/2/3/.../`τ_2 (slaves), and each slave `evalCOMM` is further subdivided into `analCOMM1/2/3/.../`τ_3. Logic for selection of τ_i is discussed in [39].

18.3.3 Hybrid Parallelism

Hybrid parallelism approaches can take several forms when used in the multilevel parallel context. A conceptual boundary can be considered to exist for which all parallelism above the boundary is managed internally using message-passing and all parallelism below the boundary is managed externally using asynchronous local approaches. Hybrid parallelism approaches can then be categorized based on whether this boundary between internal and external management occurs within a parallelism level (*intra-level*) or between two parallelism levels (*inter-level*). In the intra-level case, the jobs for the parallelism level containing the boundary are scheduled using a hybrid scheduler, in which a capacity multiplier is used for the number of jobs to assign to each server. Each server is then responsible for concurrently executing its capacity of jobs using an asynchronous local approach. In the inter-level case, one level of parallelism manages its parallelism internally using a message-passing approach and the next lower level of parallelism manages its parallelism externally using an asynchronous local approach. That is, the jobs for the higher level of parallelism are scheduled using a standard message-passing scheduler, in which a single job is assigned to each server. However, each of these jobs has multiple components, as managed by the next lower level of parallelism, and each server is responsible for executing these sub-components concurrently using an asynchronous local approach.

For example, consider a multiprocessor DAKOTA run which involves an iterator scheduling a set of concurrent function evaluations across a cluster of SMPs. A hybrid parallelism approach will be applied in which message-passing parallelism is used between SMPs and asynchronous local parallelism is used within each SMP. In the hybrid intra-level case, multiple function evaluations would be scheduled to each SMP, as dictated by the capacity of the SMPs, and each SMP would manage its own set of concurrent function evaluations using an asynchronous local approach. Any lower levels of parallelism would be serialized. In the hybrid inter-level case, the function evaluations would be scheduled one per SMP, and the analysis components within each of these evaluations would be executed concurrently using asynchronous local approaches within the SMP. Thus, the distinction can be viewed as whether the concurrent jobs on each server in Figure 18.1c reflect the same level of parallelism as that being scheduled by the master (intra-level) or one level of parallelism below that being scheduled by the master (inter-level).

18.4 Capability Summary

Table 18.1 shows a matrix of the supported job management approaches for each of the parallelism levels, with supported simulation interfaces and synchronization approaches shown in parentheses. The concurrent iterator and multiprocessor analysis parallelism levels can only be managed with message-passing approaches. In the former case, this is due to the fact that a separate process or thread for an iterator is not currently supported. The latter case reflects a finer point on the definition of external parallelism management. While a multiprocessor analysis can most certainly be launched (e.g., using `mpirun`/`yod`) from one of DAKOTA's analysis drivers, resulting in a parallel analysis external to DAKOTA (which is consistent with asynchronous local and hybrid approaches), this parallelism is not visible to DAKOTA and therefore does not qualify as parallelism that DAKOTA manages (and therefore is not included in Table 18.1). The concurrent evaluation and analysis levels can be managed either with message-passing, asynchronous local, or hybrid techniques, with the exceptions that the direct interface does not

Table 18.1: Support of job management approaches within parallelism levels. Shown in parentheses are supported simulation interfaces and supported synchronization approaches.

Parallelism Level	Asynchronous Local	Message Passing	Hybrid
concurrent iterators within a strategy		**X** (blocking only)	
concurrent function evaluations within an iterator	**X** (system, fork) (blocking, nonblocking)	**X** (system, fork, direct) (blocking only)	**X** (system, fork) (blocking only)
concurrent analyses within a function evaluation	**X** (fork only) (blocking only)	**X** (system, fork, direct) (blocking only)	**X** (fork only) (blocking only)
fine-grained parallel analysis		**X**	

support asynchronous operations (asynchronous local or hybrid) at either of these levels and the system call interface does not support asynchronous operations (asynchronous local or hybrid) at the concurrent analysis level. The direct interface restrictions are present since multithreading in not yet supported and the system call interface restrictions result from the inability to manage concurrent analyses within a nonblocking function evaluation system call. Finally, nonblocking synchronization is only currently supported for asynchronous local parallelism at the concurrent function evaluation level. In time, message passing and hybrid parallelism approaches will also support nonblocking synchronization at this level.

18.5 Running a Parallel DAKOTA Job

Section 18.2 provides a few examples of serial and parallel execution of DAKOTA using asynchronous local, message passing, and hybrid approaches to single-level parallelism. The following sections provides a more complete discussion of the parallel execution syntax and available specification controls.

18.5.1 Single-processor execution

The command for running DAKOTA on a single-processor and exploiting asynchronous local parallelism is the same as for running DAKOTA on a single-processor for a serial study, e.g.:

```
dakota -i dakota.in > dakota.out
```

See Section 2.1.5 for additional information on single-processor command syntax.

18.5.2 Multiprocessor execution

Running a DAKOTA job on multiple processors requires the use of an executable loading facility such as `mpirun`, `mpiexec`, `poe`, or `yod`. On a network of workstations, the `mpirun` script is commonly used to initiate a parallel DAKOTA job, e.g.:

```
mpirun -np 12 dakota -i dakota.in > dakota.out
mpirun -machinefile machines -np 12 dakota -i dakota.in > dakota.out
```

where both examples specify the use of 12 processors, the former selecting them from a default system resources file and the latter specifying particular machines in a machine file (see [67] for details).

On a massively parallel computer such as ASCI Red, similar facilities are available from the Cougar operating system via the `yod` executable loading facility:

```
yod -sz 512 dakota -i dakota.in > dakota.out
```

In each of these cases, MPI command line arguments are used by MPI (extracted first in the call to `MPI_Init`) and DAKOTA command line arguments are used by DAKOTA (extracted second by DAKOTA's command line handler). An issue that can arise with these command line arguments is that the mpirun script distributed with MPICH has been observed to have problems with certain file path specifications (e.g., a relative path such as "`../some_file`"). These path problems are most easily resolved by using local linkage (all referenced files or soft links to these files appear in the same directory).

Finally, when running on computer resources that employ NQS/PBS batch schedulers, the single-processor `dakota` command syntax or the multiprocessor `mpirun` command syntax might be contained within an executable script file which is submitted to the batch queue. For example, on Cplant, the command

```
qsub -l size=512 run_dakota
```

could be submitted to the PBS queue for execution. On ASCI Red, the NQS syntax is similar:

```
qsub -q snl -lP 512 -lT 6:00:00 run_dakota
```

These commands allocate 512 compute nodes for the study, and execute the `run_dakota` script on a service node. If this script contains a single-processor `dakota` command, then DAKOTA will execute on a single service node from which it can launch parallel simulations into the compute nodes using analysis drivers that contain `yod` commands (any `yod` executions occurring at any level underneath the `run_dakota` script are mapped to the 512 compute node allocation). If the script submitted to `qsub` contains a multiprocessor `mpirun` command, then DAKOTA will execute across multiple service nodes so that it can spread the application load in either a message-passing or hybrid parallelism approach. Again, analysis drivers containing `yod` commands would be responsible for utilizing the 512 compute nodes. And, finally, if the script submitted to `qsub` contains a `yod` of the `dakota` executable, then DAKOTA will execute directly on the compute nodes and manage all of the parallelism internally (note that a `yod` of this type without a `qsub` would be mapped to the interactive partition, rather than to the batch partition).

Not all supercomputers employ the same model for service/compute partitions or provide the same support for tiling of concurrent multiprocessor simulations within a single NQS/PBS allocation. For this reason, templates for parallel job configuration are being catalogued within `Dakota/examples/script_interfaces` and `Dakota/examples/parallelism` (in the software distributions) that are intended to provide guidance for individual machine idiosyncrasies.

18.6 Specifying Parallelism

Given an allotment of processors, DAKOTA contains logic based on the theoretical work in [39] to automatically determine an efficient parallel configuration, consisting of partitioning and scheduling selections for each of the parallelism levels. This logic accounts for problem size, the concurrency supported by particular iterative algorithms, and any user inputs or overrides. The following points are important components of the automatic configuration logic which can be helpful in estimating the total number of processors to allocate and in selecting configuration overrides:

- If the capacity of the servers in a peer configuration is sufficient to schedule all jobs in one pass, then a peer partition and static schedule will be selected. If this capacity is not sufficient, then a dedicated-master partition and dynamic schedule will be used. These selections can be overridden with self/static scheduling request specifications for the concurrent iterator, evaluation, and analysis parallelism levels. For example, if it is known that processor speeds and job durations have little variability, then overriding the automatic configuration with a static schedule request could eliminate the unnecessary loss of a processor to scheduling.
- With the exception of the concurrent-iterator parallelism level (iterator executions tend to have high variability in duration), concurrency is pushed up. That is, available processors will be assigned to concurrency at the higher parallelism levels first. If more processors are available than needed for concurrency at a level, then the server size is increased to support concurrency in the next lower level of parallelism. This process is continued until all available processors have been assigned. These assignments can be overridden with a servers specification for the concurrent iterator, evaluation, and analysis parallelism levels and with a processors per analysis specification for the multiprocessor analysis parallelism level. For example, if it is desired to parallelize concurrent analyses within each function evaluation, then an `evaluation_servers = 1` override would serialize the concurrent function evaluations level and assure processor availability for concurrent analyses.

In the following sections, the user inputs and overrides are described, followed by specification examples for single and multi-processor DAKOTA executions.

18.6.1 The interface specification

Specifying parallelism within an interface can involve the use of the `asynchronous`, `evaluation_concurrency`, and `analysis_concurrency` keywords to specify concurrency local to a processor (i.e., asynchronous local parallelism). This `asynchronous` specification has dual uses:

- When running DAKOTA on a single-processor, the `asynchronous` keyword specifies the use of asynchronous invocations local to the processor (these jobs then rely on external means to be allocated to other processors). The default behavior is to simultaneously launch all function evaluations available from the iterator as well as all available analyses within each function evaluation. In some cases, the default behavior can overload a machine or violate a usage policy, resulting in the need to limit the number of concurrent jobs using the `evaluation_concurrency` and `analysis_concurrency` specifications.
- When executing DAKOTA across multiple processors and managing jobs with a message-passing scheduler, the `asynchronous` keyword specifies the use of asynchronous invocations local to each server processor, resulting in a hybrid parallelism approach (see Section 18.2.3). In this case, the default behavior is one job per server, which must be overridden with an `evaluation_concurrency` specification and/or an `analysis_concurrency` specification. When a hybrid parallelism approach is specified, the capacity of the servers (used in the automatic configuration logic) is defined as the number of servers times the number of asynchronous jobs per server.

In addition, `evaluation_servers`, `evaluation_self_scheduling`, and `evaluation_static_scheduling` keywords can be used to override the automatic parallelism configuration for concurrent function evaluations; `analysis_servers`, `analysis_self_scheduling`, and `analysis_static_scheduling` keywords can be used to override the automatic parallelism configuration for concurrent analyses; and the `processors_per_analysis` keyword can be used to override the automatic parallelism configuration for the size of multiprocessor analyses used in a direct function simulation interface. Each of these keywords appears as part of the interface commands specification in the DAKOTA Reference Manual [30].

18.6.2 The strategy specification

To specify concurrency in iterator executions, the `iterator_servers`, `iterator_self_scheduling`, and `iterator_static_scheduling` keywords are used to override the automatic parallelism configuration. See the strategy commands specification in the DAKOTA Reference Manual [30] for additional information.

18.6.3 Single-processor DAKOTA specification

Specifying a single-processor DAKOTA job that exploits parallelism through asynchronous local approaches (see Figure 18.1a) requires inclusion of the `asynchronous` keyword in the interface specification. Once the input file is defined, single-processor DAKOTA jobs are executed using the command syntax described previously in Section 18.5.1.

Example 1

For example, the following specification runs an NPSOL optimization which will perform asynchronous finite differencing:

```
method,
        npsol_sqp

variables,
        continuous_design = 5
          initial_point  0.2  0.05 0.08 0.2  0.2
          lower_bounds   0.15 0.02 0.05 0.1  0.1
          upper_bounds   2.0  2.0  2.0  2.0  2.0

interface,
        system,
          asynchronous
          analysis_drivers = 'text_book'

responses,
        num_objective_functions = 1
        num_nonlinear_inequality_constraints = 2
        numerical_gradients
          interval_type central
          method_source dakota
          fd_gradient_step_size = 1.e-4
        no_hessians
```

Note that `method_source dakota` selects DAKOTA's internal finite differencing routine so that the concurrency in finite difference offsets can be exploited. In this case, central differencing has been selected and 11 function evaluations (one at the current point plus two offsets in each of five variables) can be performed simultaneously for each NPSOL response request. These 11 evaluations will be launched with system calls in the background and presumably assigned to additional processors through the operating system of a multiprocessor compute server or other comparable method. The concurrency specification may be included if it is necessary to limit the maximum number of simultaneous evaluations. For example, if a maximum of six compute processors were available, the command

```
evaluation_concurrency = 6
```

could be added to the `asynchronous` specification within the `interface` keyword from the preceding example.

Example 2

If, in addition, multiple analyses can be executed concurrently within a function evaluation (e.g., from multiple load cases or disciplinary analyses that must be evaluated to compute the response data set), then an input specification similar to the following could be used:

```
method,
        npsol_sqp

variables,
        continuous_design = 5
          initial_point  0.2  0.05 0.08 0.2  0.2
          lower_bounds   0.15 0.02 0.05 0.1  0.1
          upper_bounds   2.0  2.0  2.0  2.0  2.0

interface,
        fork
          asynchronous
            evaluation_concurrency = 6
            analysis_concurrency = 3
          analysis_drivers = 'text_book1' 'text_book2' 'text_book3'

responses,
        num_objective_functions = 1
        num_nonlinear_inequality_constraints = 2
        numerical_gradients
          method_source dakota
          interval_type central
          fd_gradient_step_size = 1.e-4
        no_hessians
```

In this case, the default concurrency with just an `asynchronous` specification would be all 11 function evaluations and all 3 analyses, which can be limited by the `evaluation_concurrency` and `analysis_concurrency` specifications. The input file above limits the function evaluation concurrency, but not the analysis concurrency (a specification of 3 is the default in this case and could be omitted). Changing the input to `evaluation_concurrency = 1` would serialize the function evaluations, and changing the input to `analysis_concurrency = 1` would serialize the analyses.

18.6.4 Multiprocessor DAKOTA specification

In multiprocessor executions, server evaluations are synchronous (Figure 18.1b) by default and the `asynchronous` keyword is only used if a hybrid parallelism approach (Figure 18.1c) is desired. Multiprocessor DAKOTA jobs are executed using the command syntax described previously in Section 18.5.2.

Example 3

To run Example 1 using a message-passing approach, the `asynchronous` keyword would be removed (since the servers will execute their evaluations synchronously), resulting in the following interface specification:

```
interface,
        system,
          analysis_drivers = 'text_book'
```

Running DAKOTA on 4 processors (syntax: `mpirun -np 4 dakota -i dakota.in`) would result in the following parallel configuration report from the DAKOTA output:

```
-----------------------------------------------------------------------------
DAKOTA parallel configuration:

Level                   num_servers    procs_per_server    partition/schedule
-----                   -----------    ----------------    ------------------
concurrent iterators         1               4              peer/static
concurrent evaluations       3               1              ded. master/self
concurrent analyses          1               1              peer/static
multiprocessor analysis      1              N/A                N/A

Total parallelism levels =   1
-----------------------------------------------------------------------------
```

The dedicated master partition and self-scheduling algorithm are automatically selected for the concurrent evaluations parallelism level since the number of function evaluations (11) is greater than the maximum capacity of the servers (4). Since one of the processors is dedicated to being the master, only 3 processors are available for computation and the 11 evaluations can be completed in approximately 4 passes through the servers. If it is known that there is little variability in evaluation duration, then this logic could be overridden to use a static schedule through use of the `evaluation_static_scheduling` specification:

```
interface,
        system,
          evaluation_static_scheduling
          analysis_drivers = 'text_book'
```

Running DAKOTA again on 4 processors (syntax: `mpirun -np 4 dakota -i dakota.in`) would now result in this parallel configuration report:

```
-----------------------------------------------------------------------------
DAKOTA parallel configuration:

Level                   num_servers    procs_per_server    partition/schedule
-----                   -----------    ----------------    ------------------
concurrent iterators         1               4              peer/static
concurrent evaluations       4               1              peer/static
concurrent analyses          1               1              peer/static
multiprocessor analysis      1              N/A                N/A

Total parallelism levels =   1
-----------------------------------------------------------------------------
```

Now the 11 jobs will be statically distributed among 4 peer servers, since the processor previously dedicated to scheduling has been converted to a compute server. This could be more efficient if the evaluation durations are sufficiently similar, but there is no mechanism to adapt to heterogeneity in processor speeds or simulation expense.

As a related example, consider the case where each of the workstations used in the parallel execution has multiple processors. In this case, a hybrid parallelism approach which combines message-passing parallelism with asynchronous local parallelism (see Figure 18.1c) would be a good choice. To specify hybrid parallelism, one uses the same `asynchronous` specification as was used for the single-processor examples, e.g.:

```
interface,
         system
           asynchronous evaluation_concurrency = 3
           analysis_drivers = 'text_book'
```

With 3 function evaluations concurrent on each server, the capacity of a 4 processor DAKOTA execution (syntax: `mpirun -np 4 dakota -i dakota.in`) has increased to 12 evaluations. Since all 11 jobs can now be scheduled in a single pass, a static schedule is automatically selected (without any override request):

```
-----------------------------------------------------------------------------
DAKOTA parallel configuration:

Level                   num_servers    procs_per_server    partition/schedule
-----                   -----------    ----------------    ------------------
concurrent iterators         1               4              peer/static
concurrent evaluations       4               1              peer/static
concurrent analyses          1               1              peer/static
multiprocessor analysis      1              N/A                N/A

Total parallelism levels =   1
-----------------------------------------------------------------------------
```

Example 4

To run Example 2 using a message-passing approach, the `asynchronous` specification is again removed:

```
interface,
         fork
           analysis_drivers = 'text_book1' 'text_book2' 'text_book3'
```

Running this example on 6 processors (syntax: `mpirun -np 6 dakota -i dakota.in`) would result in the following parallel configuration report:

```
-----------------------------------------------------------------------------
DAKOTA parallel configuration:

Level                   num_servers    procs_per_server    partition/schedule
-----                   -----------    ----------------    ------------------
concurrent iterators         1               6              peer/static
concurrent evaluations       5               1              ded. master/self
concurrent analyses          1               1              peer/static
multiprocessor analysis      1              N/A                N/A
```

```
Total parallelism levels =   1
-------------------------------------------------------------------------------
```

in which all of the processors have been assigned to support evaluation concurrency due to the "push up" automatic configuration logic. Note that the default configuration could be a poor choice in this case, since 11 jobs scheduled through 5 servers will likely have significant idleness towards the end of the scheduling. To assign some of the available processors to the concurrent analysis level, the following input could be used:

```
interface,
         fork
           analysis_drivers = 'text_book1' 'text_book2' 'text_book3'
           evaluation_static_scheduling
           evaluation_servers = 2
```

which results in the following 2-level parallel configuration:

```
-------------------------------------------------------------------------------
DAKOTA parallel configuration:

Level                   num_servers    procs_per_server    partition/schedule
-----                   -----------    ----------------    ------------------
concurrent iterators         1               6              peer/static
concurrent evaluations       2               3              peer/static
concurrent analyses          3               1              peer/static
multiprocessor analysis      1              N/A                N/A

Total parallelism levels =   2
-------------------------------------------------------------------------------
```

The six processors available have been split into two evaluation servers of three processors each, where the three processors in each evaluation server manage the three analyses, one per processor.

Next, consider the following 3-level parallel case, in which `text_book1`, `text_book2`, and `text_book3` from the previous examples now execute on two processors each. In this case, the `processors_per_analysis` keyword is added and the `fork` interface is changed to a `direct` interface since the fine-grained parallelism of the three simulations is managed internally:

```
interface,
         direct
           analysis_drivers = 'text_book1' 'text_book2' 'text_book3'
           evaluation_static_scheduling
           evaluation_servers = 2
           processors_per_analysis = 2
```

This results in the following parallel configuration for a 12 processor DAKOTA run (syntax: `mpirun -np 12 dakota -i dakota.in`):

```
-------------------------------------------------------------------------------
DAKOTA parallel configuration:

Level                   num_servers    procs_per_server    partition/schedule
-----                   -----------    ----------------    ------------------
concurrent iterators         1              12              peer/static
```

```
concurrent evaluations       2                6              peer/static
concurrent analyses          3                2              peer/static
multiprocessor analysis      2               N/A                N/A

Total parallelism levels =   3
-------------------------------------------------------------------------------
```

An important point to recognize is that, since each of the parallel configuration inputs has been tied to the interface specification up to this point, these parallel configurations can be reallocated for each interface in a multi-iterator/multi-model strategy. For example, a DAKOTA execution on 40 processors might involve the following two interface specifications:

```
interface,
        direct,
          id_interface = 'COARSE'
          analysis_driver = 'sim1'
          processors_per_analysis = 5

interface,
        direct,
          id_interface = 'FINE'
          analysis_driver = 'sim2'
          processors_per_analysis = 10
```

for which the coarse model would employ 8 servers of 5 processors each and the fine model would employ 4 servers of 10 processors each.

Next, consider the following 4-level parallel case that employs the Pareto set optimization strategy. In this case, `iterator_servers` and `iterator_static_scheduling` requests are included in the strategy specification:

```
strategy,
         pareto_set
           iterator_servers = 2
           iterator_static_scheduling
           opt_method_pointer = 'NLP'
           random_weight_sets = 4
```

Adding this strategy specification to the input file from the previous 12 processor example results in the following parallel configuration for a 24 processor DAKOTA run
(syntax: `mpirun -np 24 dakota -i dakota.in`):

```
-------------------------------------------------------------------------------
DAKOTA parallel configuration:

Level                   num_servers    procs_per_server    partition/schedule
-----                   -----------    ----------------    ------------------
concurrent iterators         2               12              peer/static
concurrent evaluations       2                6              peer/static
concurrent analyses          3                2              peer/static
multiprocessor analysis      2               N/A                N/A

Total parallelism levels =   4
-------------------------------------------------------------------------------
```

Example 5

As a final example, consider a multi-start optimization conducted on 384 processors of ASCI Red. A job of this size must be submitted to the batch queue, using syntax similar to:

```
qsub -q snl -lP 384 -lT 6:00:00 run_dakota
```

where the `run_dakota` script appears as

```
#!/bin/sh
cd /scratch/<some_workdir>
yod -sz 384 dakota -i dakota.in > dakota.out
```

and the strategy and interface specifications from the `dakota.in` input file appear as

```
strategy,
        multi_start
          method_pointer = 'CPS'
          iterator_servers = 8
          random_starts = 8

interface,
        direct,
          analysis_drivers = 'text_book1' 'text_book2' 'text_book3'
          evaluation_servers = 8
          evaluation_static_scheduling
          processors_per_analysis = 2
```

The resulting parallel configuration is reported as

```
------------------------------------------------------------------------------
DAKOTA parallel configuration:

Level                   num_servers    procs_per_server    partition/schedule
-----                   -----------    ----------------    ------------------
concurrent iterators         8               48               peer/static
concurrent evaluations       8                6               peer/static
concurrent analyses          3                2               peer/static
multiprocessor analysis      2               N/A                 N/A

Total parallelism levels =   4
------------------------------------------------------------------------------
```

Since the concurrency at each of the nested levels has a multiplicative effect on the number of processors that can be utilized, it is easy to see how large numbers of processors can be put to effective use in reducing the time to reach a solution, even when, as in this example, the concurrency per level is relatively low.

18.7 Application Parallelism Use Cases

This section describes several common use cases for running DAKOTA on parallel computing clusters with various combinations of DAKOTA and application parallelism. In three of the four cases addressed, the application

Table 18.2: Cases for DAKOTA and application-level parallelism with M available processors and each application job requiring N processors. Cases 1–3 assume that DAKOTA and any application runs will execute wholly within a single scheduled job, whereas Case 4 is relevant when analysis jobs must be individually submitted to a scheduler.

Case	DAKOTA	Application	Notes
1	parallel	serial	$M-1$ (or M) simultaneous application instances each $N=1$ processor
2	serial	parallel	1 simultaneous application instance on N processors
3	serial	parallel	$\approx (M-1)/N$ or $\approx M/N$ simultaneous N processor jobs
4	serial	parallel	submit *expensive* N processor application jobs to a scheduler (e.g., qsub)

Table 18.3: Files found in Dakota/examples/parallelism/Case* to illustrate DAKOTA+application parallelism. Not all files appear in each directory.

pbs_submission	script file describing the job to be submitted to the scheduler (e.g., qsub, Case 1–3 only)
dakota_pstudy.in	DAKOTA input file describing the study to execute
text_book_par_driver	script to create working directories, prepare inputs, and run the analysis code
text_book_par	the binary "analysis code" to run (works in serial or parallel)
post_process	collect the results of a DAKOTA study performed by submitting analysis jobs to a scheduler, e.g., qsub (Case4 only)

launched by DAKOTA is assumed MPI-enabled and run as an independent parallel process. For demonstration purposes, the following characteristics are shared among the usage examples:

- DAKOTA performs a vector parameter study requiring 20 model evaluations (application runs).
- For each evaluation, DAKOTA uses a fork simulation interface to call a shell script `text_book_par_driver` to launch the application. This script is a stand-in for a typical DAKOTA-application black box interface (as described in Chapter 17.1), and includes mock application input preparation, execution, and postprocessing to return necessary metrics to DAKOTA.
- The application executed is a modified version of the text book example driver, `text_book_par`, capable of serial or parallel execution.

The combinations of DAKOTA and application parallelism are summarized in Table 18.2. In each case, M denotes the total number of processors allocated and N denotes the number of processors used by a single application analysis. For most scenarios, Cases 1–3, where DAKOTA and the application jobs run within a single cluster processor allocation (queued job), are preferred. However for particularly long-running or large jobs, or platforms that not supporting the first scheduling modes, Case 4 may be most appropriate. Relevant example files are included in directories `Dakota/examples/parallelism/Case*` with the DAKOTA distribution and described in Table 18.3.

Case 1: Multiple serial analysis jobs

In this case, DAKOTA will launch multiple simultaneous single processor application runs (massively serial analysis code execution, an embarrassingly parallel model). DAKOTA is run in parallel, making this example an elaboration of the message-passing single-level parallel mode described in Section 18.2. Specifically in this example, DAKOTA is run in parallel with $M=6$ processors (`pbs_submission`):

```
mpiexec -n 6 dakota dakota_pstudy.in
```

and its default master-slave schedule will launch $M - 1$ simultaneous analysis jobs, and as each job completes, another will be launched, until all jobs are complete. Several options are possible in this case:

- If the possible DAKOTA application concurrency equals M, DAKOTA will use a peer-to-peer scheduler, and run the M jobs concurrently. When the possible concurrency is greater than M, DAKOTA will by default launch $M - 1$ jobs with a master-slave model. Specifying `static_schedule` in the DAKOTA input, will override the default master-slave scheduler and DAKOTA will launch M jobs, but jobs will be launched blocking, so all M will complete, then another M will be scheduled.
- If the analysis is extremely inexpensive, performance may be improved by launching multiple evaluation jobs local to each DAKOTA MPI process, specifying

  ```
  asynchronous evaluation_concurrency = [2 or more]
  ```

- It is also possible to launch only one DAKOTA process per node, and then use either asynchronous local as above, or launch the application in parallel using only the local processors (shared-memory MPI parallelism):

  ```
  mpiexec -pernode -n 3 dakota dakota_pstudy.in
  ```

Caveat: This example assumes the application is capable of serial execution, which on some platforms or MPI implementations is not equivalent to `mpiexec -n 1`. Some MPI/scheduler combinations will not permit another MPI process to run on a resource assigned to the DAKOTA processes.

Case 2: One simultaneous parallel analysis job

This case is relevant for multi-processor analysis jobs, typically where the analysis is expensive (i.e., is long-running or sufficient processors are not available to run more than one simultaneous analysis). Note that for extremely long-running parallel jobs, Case 4 below may be more appropriate.

In this case, DAKOTA runs in serial

```
dakota dakota_pstudy.in
```

and the driver script launches the application with `mpiexec -n K`, where $K \leq M$, to launch the application code within the processor allocation:

```
mpiexec -n 6 text_book_par application.in application.out
```

Case 3: Multiple simultaneous parallel analysis jobs

In this "job tiling" case, a single scheduled processor allocation is partitioned to run $\approx (M-1)/N$ or $\approx M/N$ parallel application jobs, each requiring N processors. We describe two current ways to accomplish this (though other solutions exist): use option (a) if the application will work correctly in an MPICH/MVAPICH environment and option (b) otherwise.

Case 3a: Mpiexec server mode

Mpiexec (http://www.osc.edu/ pw/mpiexec/) works in concert with MPICH implementations, extending mpirun to run jobs in a PBS environment with additional features. It offers a background server option which can be used to tile multiple MPI jobs within a single parallel resource allocation. (Note that with MPICH, there is a difference between `mpirun` and `mpiexec`, unlike with OpenMPI, where both are typically aliases for `orterun`.)

In this case, an `mpiexec` server process is started and backgrounded to service application requests for processors; DAKOTA runs in serial (`pbs_submission`):

```
mpiexec -server &

dakota dakota_pstudy.in
```

and asynchronously launches $M/N = 3$ evaluations (`dakota_pstudy.in`):

```
interface, application fork, asynchronous evaluation_concurrency = 3
  analysis_driver = 'text_book_par_driver'
```

The simulator script calls `mpiexec -n 2` to run the analysis in parallel and the mpiexec server assigns a subset of the available processors to the particular MPI task (`text_book_par`):

```
mpiexec -n 2 text_book_par application.in application.out
```

An error will result if more application tasks are launched than the processor allocation permits. An error may result if the application does not exit cleanly. At present similar capability is not supported by OpenMPI, although a daemon mode similar to Mpiexec has been proposed.

Case 3b: Script-based partitioning

This Case 3 variant applies when the application must be compiled with OpenMPI or another MPI implementation that does not support a server mode for job tiling, but does support the use of machine files specifying the resources on which to run the application job. A set of scripts are used to manage the partitioning of the M processor allocation among N analysis jobs, each with a machines file consisting of a unique subset of the assigned resources. Note that this will not work with early OpenMPI versions with some resource managers (e.g., OpenMPI 1.2 with Torque), where machinefiles, even if a proper subset of `$PBS_NODEFILE`, are ignored. This will however work with OpenMPI 1.3 and newer.

In this case the `pbs_submission` script includes variables specifying how to create a separate node file for each job:

```
# number of simultaneous application jobs to run
export NUM_APPLIC_JOBS=2
# number of processors neded per application job
export NUM_APPLIC_PROCS=3

# divide the assigned resources into a nodefile for DAKOTA and
# NUM_APPLIC_JOBS nodefiles for the application runs
setup_nodes.sh $PBS_NODEFILE
```

Similarly to Case 3a, DAKOTA runs in serial and uses asynchronous evaluation concurrency to launch the jobs. The `text_book_par_driver` now contains logic to lock a node file for the application run and return it when complete. As each job completes, the next is scheduled.

Case 4: Parallel analysis jobs submitted to a queue

This case describes running DAKOTA to submit parallel jobs to a batch queue. This option is likely only useful when the cost of an individual analysis evaluation is high (such that the job requires far too many processors or hours to run all the evaluations) and there is no feedback to DAKOTA required to generate subsequent evaluation points. So this scenario is likely more relevant for sensitivity analysis and uncertainty quantification than optimization.

In the first pass, DAKOTA runs (likely interactively) in serial on a login node or other node capable of job submission:

```
dakota dakota_pstudy.in
```

For each evaluation, the simulator script (`text_book_par_driver`) will generate a `pbs_submission` script and submit it to the scheduler. Dummy results are returned to DAKOTA which will exit when all jobs have been scheduled.

In the second pass, when analysis is complete, the analysis driver is changed to `post_process` and DAKOTA is executed on a login node to collect the results of the study.

Chapter 19

DAKOTA Usage Guidelines

19.1 Problem Exploration

The first objective in an analysis is to characterize the problem so that appropriate algorithms can be chosen. In the case of optimization, typical questions that should be addressed include: Are the design variables continuous, discrete, or mixed? Is the problem constrained or unconstrained? How expensive are the response functions to evaluate? Will the response functions behave smoothly as the design variables change or will there be nonsmoothness and/or discontinuities? Are the response functions likely to be multimodal, such that global optimization may be warranted? Is analytic gradient data available, and if not, can I calculate gradients accurately and cheaply? Additional questions that are pertinent for characterization of uncertainty quantification problems include: Can I accurately model the probabilistic distributions of my uncertain variables? Are the response functions relatively linear? Am I interested in a full random process characterization of the response functions, or just statistical results?

If there is not sufficient information from the problem description to answer these questions, then additional problem characterization activities may be warranted. One particularly useful characterization activity that DAKOTA enables is parameter space exploration through the use of parameter studies and design of experiments methods. The parameter space can be systematically interrogated to create sufficient information to evaluate the trends in the response functions and to determine if these trends are noisy or smooth, unimodal or multimodal, relatively linear or highly nonlinear, etc. In addition, the parameter studies may reveal that one or more of the parameters do not significantly affect the results and can be removed from the problem formulation. This can yield a potentially large savings in computational expense for the subsequent studies. Refer to Chapters 4 and 5 for additional information on parameter studies and design of experiments methods.

19.2 Optimization Method Selection

In selecting an optimization method, important considerations include the type of variables in the problem (continuous, discrete, mixed), whether a global search is needed or a local search is sufficient, and the required constraint support (unconstrained, bound constrained, or generally constrained). Less obvious, but equally important, considerations include the efficiency of convergence to an optimum (i.e., convergence rate) and the robustness of the method in the presence of challenging design space features (e.g., nonsmoothness).

Gradient-based optimization methods are highly efficient, with the best convergence rates of all of the optimization

methods. If analytic gradient and Hessian information can be provided by an application code, a full Newton method will provide quadratic convergence rates near the solution. More commonly, only gradient information is available and a quasi-Newton method is chosen in which the Hessian information is approximated from an accumulation of gradient data. In this case, superlinear convergence rates can be obtained. These characteristics make gradient-based optimization methods the methods of choice when the problem is smooth, unimodal, and well-behaved. However, when the problem exhibits nonsmooth, discontinuous, or multimodal behavior, these methods can also be the least robust since inaccurate gradients will lead to bad search directions, failed line searches, and early termination, and the presence of multiple minima will be missed.

Thus, for gradient-based optimization, a critical factor is the gradient accuracy. Analytic gradients are ideal, but are often unavailable. For many engineering applications, a finite difference method will be used by the optimization algorithm to estimate gradient values. DAKOTA allows the user to select the step size for these calculations, as well as choose between forward-difference and central-difference algorithms. The finite difference step size should be selected as small as possible, to allow for local accuracy and convergence, but not so small that the steps are "in the noise." This requires an assessment of the local smoothness of the response functions using, for example, a parameter study method. Central differencing, in general, will produce more reliable gradients than forward differencing, but at roughly twice the expense.

Nongradient-based methods exhibit much slower convergence rates for finding an optimum, and as a result, tend to be much more computationally demanding than gradient-based methods. Nongradient local optimization methods, such as pattern search algorithms, often require from several hundred to a thousand or more function evaluations, depending on the number of variables, and nongradient global optimization methods such as genetic algorithms may require from thousands to tens-of-thousands of function evaluations. Clearly, for nongradient optimization studies, the computational cost of the function evaluation must be relatively small in order to obtain an optimal solution in a reasonable amount of time. In addition, nonlinear constraint support in nongradient methods is an open area of research and, while supported by many nongradient methods in DAKOTA, is not as refined as constraint support in gradient-based methods. However, nongradient methods can be more robust and more inherently parallel than gradient-based approaches. They can be applied in situations were gradient calculations are too expensive or unreliable. In addition, some nongradient-based methods can be used for global optimization which gradient-based techniques, by themselves, cannot. For these reasons, nongradient-based methods deserve consideration when the problem may be nonsmooth, multimodal, or poorly behaved.

Approaches that seek to improve the effectiveness or efficiency of optimizers and least squares methods through the use of surrogate models include the surrogate-based local, surrogate-based global, and efficient global methods. Chapter 9 provides further information on these approaches. The surrogate-based local approach (see Section 9.2) brings the efficiency of gradient-based optimization/least squares methods to nonsmooth or poorly behaved problems by smoothing noisy or discontinuous response results with a data fit surrogate model (e.g., a quadratic polynomial) and then minimizing on the smooth surrogate using efficient gradient-based techniques. The surrogate-based global approach (see Section 9.3) similarly employs optimizers/least squares methods with surrogate models, but rather than localizing through the use of trust regions, seeks global solutions using global methods. And the efficient global approach (see Section 9.4) uses the specific combination of Gaussian process surrogate models in combination with the DIRECT global optimizer. Similar to these surrogate-based approaches, the hybrid and multistart optimization strategies seek to bring the efficiency of gradient-based optimization methods to global optimization problems. In the former case, a global optimization method can be used for a few cycles to locate promising regions and then local gradient-based optimization is used to efficiently converge on one or more optima. In the latter case, a stratification technique is used to disperse a series of local gradient-based optimization runs through parameter space. Without surrogate data smoothing, however, these strategies are best for smooth multimodal problems. Section 10.2 and Section 10.3 provide more information on these approaches.

Table 19.1 provides a convenient reference for choosing an optimization method or strategy to match the characteristics of the user's problem, where blank fields inherit the value from above. With respect to constraint support,

it should be understood that the methods with more advanced constraint support are also applicable to the lower constraint support levels; they are listed only at their highest level of constraint support for brevity.

19.3 UQ Method Selection

The need for computationally efficient methods is further amplified in the case of the quantification of uncertainty in computational simulations. Sampling-based methods are the most robust uncertainty techniques available, are applicable to almost all simulations, and possess rigorous error bounds; consequently, they should be used whenever the function is relatively inexpensive to compute. However, in the case of terascale computational simulations, the number of function evaluations required by traditional techniques such as Monte Carlo and Latin hypercube sampling (LHS) quickly becomes prohibitive, especially if tail statistics are needed. Additional sampling options include quasi-Monte Carlo (QMC) sampling and importance sampling (IS), and incremental sampling may also be used to incrementally add samples to an existing sample set.

Alternatively, one can apply the traditional sampling techniques to a surrogate function approximating the expensive computational simulation (see Section 11.6.3). However, if this approach is selected, the user should be aware that it is very difficult to assess the accuracy of the results obtained. Unlike in the case of surrogate-based local minimization (see Section 9.2), there is no simple pointwise calculation to verify the accuracy of the approximate results. This is due to the functional nature of uncertainty quantification, i.e. the accuracy of the surrogate over the entire parameter space needs to be considered, not just around a candidate optimum as in the case of surrogate-based local. This issue especially manifests itself when trying to estimate low probability events such as the catastrophic failure of a system.

Another class of UQ methods known as local reliability methods (e.g., MV, AMV/AMV^2, AMV+/AMV^2+, TANA, and FORM/SORM) are more computationally efficient in general than the sampling methods and are effective when applied to reasonably well-behaved response functions; i.e., functions that are smooth, unimodal, and linear or mildly nonlinear. They can be used to provide qualitative sensitivity information concerning which uncertain variables are important (with relatively few function evaluations), or compute full cumulative or complementary cumulative response functions (with additional computational effort). Since they rely on gradient calculations to compute local optima (most probable points of failure), they scale well for increasing numbers of random variables, but issues with nonsmooth, discontinuous, and multimodal response functions are relevant concerns. In addition, even if there is a single MPP and it is calculated accurately, first-order and second-order integrations may fail to accurately capture the shape of the failure domain. In these cases, adaptive importance sampling around the MPP can be helpful. Overall, local reliability methods should be used with some care and their accuracy should be verified whenever possible.

An effective alternative to local reliability analysis when confronted with nonsmooth, multimodal, and/or highly nonlinear response functions is efficient global reliability analysis (EGRA). This technique employs Gaussian process global surrogate models to accurately resolve the failure domain and then employs multimodal adaptive importance sampling to resolve the probabilities. For relatively low dimensional problems (i.e, on the order of 10 variables), this method displays the efficiency of local reliability analysis with the accuracy of exhaustive sampling. While extremely promising, this method is still relatively new and is the subject of ongoing refinements as we deploy it to additional applications.

The next class of UQ methods available in DAKOTA is comprised of stochastic expansion methods (polynomial chaos and stochastic collocation), which are general purpose techniques provided that the response functions possess finite second order moments. Further, these methods approximate the full random process/field and capture the underlying functional relationship between a key response metric and its random variables, rather than just approximating statistics such as mean and standard deviation. This class of methods parallels traditional variational methods in mechanics; in that vein, efforts are underway to compute rigorous error bounds of the approximations

Table 19.1: Guidelines for optimization and nonlinear least squares method selection.

Variable Type	Function Surface	Solution Type	Constraints	Applicable Methods
continuous	smooth	local opt	none	optpp_cg
			bounds	dot_bfgs, dot_frcg, conmin_frcg
			general	npsol_sqp, nlpql_sqp, dot_mmfd, dot_slp, dot_sqp, conmin_mfd, optpp_newton, optpp_q_newton, optpp_fd_newton
		local least sq	bounds	nl2sol
			general	nlssol_sqp, optpp_g_newton
		local multiobjective	general	weighted sums (one soln), pareto_set strategy (multiple solns)
		global opt	general	hybrid strategy, multi_start strategy
		global least sq	general	hybrid strategy, multi_start strategy
	nonsmooth	local opt	bounds	optpp_pds
			general	asynch_pattern_search, coliny_cobyla, coliny_pattern_search, coliny_solis_wets
		local/global opt	general	surrogate_based_local
		local/global least sq	general	surrogate_based_local
		global opt	bounds	ncsu_direct
			general	coliny_ea, coliny_direct, efficient_global, soga, surrogate_based_global
		global least sq	general	efficient_global, surrogate_based_global
		global multiobjective	general	moga (multiple solns)
discrete categorical	n/a	global opt	general	soga, coliny_ea
		global multiobjective	general	moga (multiple solns)
discrete noncategorical	n/a	local opt	general	branch_and_bound strategy
mixed categorical	nonsmooth	global opt	general	soga, coliny_ea
		global multiobjective	general	moga (multiple solns)
mixed noncategorical	smooth	local opt	general	branch_and_bound strategy

Table 19.2: Guidelines for UQ method selection.

Method Classification	Desired Problem Characteristics	Applicable Methods
Sampling	nonsmooth, multimodal response functions; response evaluations are relatively inexpensive	nond_sampling (Monte Carlo or LHS)
Local reliability	smooth, unimodal response functions; larger sets of random variables	nond_local_reliability (MV, AMV/AMV^2, AMV+/AMV^2+, TANA, FORM/SORM)
Global reliability	nonsmooth, multimodal response functions; low dimensional	nond_global_reliability
Stochastic expansions	nonsmooth, multimodal response functions; low dimensional; capture of functional form useful for subsequent analyses	nond_polynomial_chaos, nond_stoch_collocation
Epistemic	some uncertainties are poorly characterized	nond_evidence, second-order probability using nested models

produced by the methods. Another strength of these methods is their potential use in a multiphysics environment as a means to propagate the uncertainty through a series of simulations while retaining as much information as possible at each stage of the analysis. The current challenge in the development of these methods, as for other global surrogate-based methods, is effective scaling for large numbers of random variables.

The final class of UQ methods available in DAKOTA are focused on epistemic uncertainties, or uncertainties resulting from a lack of knowledge. In these problems, the assignment of input probability distributions when data is sparse can be somewhat suspect. One approach to handling epistemic uncertainties is Dempster-Shafer theory of evidence (DAKOTA method `nond_evidence`). Another method, which supports a mixture of epistemic and aleatoric uncertainties, is second-order probability used nested models (see Section 11.5). In this method, an outer epistemic level selects realizations of epistemic parameters (augmented variables) and/or realizations of random variable distribution parameters (inserted variables) from intervals. These realizations define the probability distributions for an inner aleatoric level performing probabilistic analyses. In combination, the study generates a family of CDF/CCDF respresentations which can be represented as a “horse tail” plot.

The recommendations for UQ methods are summarized in Table 19.2.

19.4 Parameter Study/DOE/DACE/Sampling Method Selection

Parameter studies, classical design of experiments (DOE), design/analysis of computer experiments (DACE), and sampling methods share the purpose of exploring the parameter space. If directed studies with a defined structure are desired, then parameter study methods (see Chapter 4) are recommended. For example, a quick assessment of the smoothness of a response function is best addressed with a vector or centered parameter study. Also, performing local sensitivity analysis is best addressed with these methods. If, however, a global space-filling set of samples is desired, then the DOE, DACE, and sampling methods are recommended (see Chapter 5). These techniques are useful for scatter plot and variance analysis as well as surrogate model construction. The distinction between DOE and DACE methods is that the former are intended for physical experiments containing an element of nonrepeatability (and therefore tend to place samples at the extreme parameter vertices), whereas the latter are intended for repeatable computer experiments and are more space-filling in nature. The distinction between DOE/DACE and sampling is drawn based on the distributions of the parameters. DOE/DACE methods typically assume uniform distributions, whereas the sampling approaches in DAKOTA support a broad range of probability distributions. To use `nond_sampling` in a design of experiments mode (as opposed to an uncertainty

Table 19.3: Guidelines for selection of parameter study, DOE, DACE, and sampling methods.

Method Classification	Applications	Applicable Methods
parameter study	sensitivity analysis, directed parameter space investigations	centered_parameter_study, list_parameter_study, multidim_parameter_study, vector_parameter_study
classical design of experiments	physical experiments (parameters are uniformly distributed)	dace (box_behnken, central_composite)
design of computer experiments	variance analysis, space filling designs (parameters are uniformly distributed)	dace (grid, random, oas, lhs, oa_lhs), fsu_quasi_mc (halton, hammersley), fsu_cvt, psuade_moat
sampling	space filling designs (parameters have general probability distributions)	nond_sampling (Monte Carlo or LHS) with all_variables flag

quantification mode), the `all_variables` flag should be included in the method specification of the DAKOTA input file.

These method selection recommendations are summarized in Table 19.3.

Chapter 20

Restart Capabilities and Utilities

20.1 Restart Management

DAKOTA was developed for solving problems that require multiple calls to computationally expensive simulation codes. In some cases you may want to conduct the same optimization, but to a tighter final convergence tolerance. This would be costly if the entire optimization analysis had to be repeated. Interruptions imposed by computer usage policies, power outages, and system failures could also result in costly delays. However, DAKOTA automatically records the variable and response data from all function evaluations so that new executions of DAKOTA can pick up where previous executions left off.

The DAKOTA restart file (e.g., `dakota.rst`) is written in a portable binary format. The portability derives from use of the XDR standard. As shown in Section 2.1.5, the primary restart commands for DAKOTA are `-read_restart`, `-write_restart`, and `-stop_restart`.

To write a restart file using a particular name, the `-write_restart` command line input (may be abbreviated as `-w`) is used:

```
dakota -i dakota.in -write_restart my_restart_file
```

If no `-write_restart` specification is used, then DAKOTA will still write a restart file, but using the default name `dakota.rst` instead of a user-specified name. To turn restart recording off, the user may select `deactivate restart_file` in the `interface` specification (refer to the Interface Commands chapter in the DAKOTA Reference Manual [30] for additional information). This can increase execution speed and reduce disk storage requirements, but at the expense of a loss in the ability to recover and continue a run that terminates prematurely. Obviously, this option is not recommended when function evaluations are costly or prone to failure.

To restart DAKOTA from a restart file, the `-read_restart` command line input (may be abbreviated as `-r`) is used:

```
dakota -i dakota.in -read_restart my_restart_file
```

If no `-read_restart` specification is used, then DAKOTA will not read restart information from any file (i.e., the default is no restart processing).

If the `-write_restart` and `-read_restart` specifications identify the same file (including the case where `-write_restart` is not specified and `-read_restart` identifies `dakota.rst`), then new evaluations will

be appended to the existing restart file. If the `-write_restart` and `-read_restart` specifications identify different files, then the evaluations read from the file identified by `-read_restart` are first written to the `-write_restart` file. Any new evaluations are then appended to the `-write_restart` file. In this way, restart operations can be chained together indefinitely with the assurance that all of the relevant evaluations are present in the latest restart file.

To read in only a portion of a restart file, the `-stop_restart` control (may be abbreviated as `-s`) is used to specify the number of entries to be read from the database. Note that this integer value corresponds to the restart record processing counter (as can be seen when using the `print` utility; see Section 20.2.1 below), which may differ from the evaluation numbers used in the previous run if, for example, any duplicates were detected (since these duplicates are not recorded in the restart file). In the case of a `-stop_restart` specification, it is usually desirable to specify a new restart file using `-write_restart` so as to remove the records of erroneous or corrupted function evaluations. For example, to read in the first 50 evaluations from `dakota.rst`:

```
dakota -i dakota.in -r dakota.rst -s 50 -w dakota_new.rst
```

The `dakota_new.rst` file will contain the 50 processed evaluations from `dakota.rst` as well as any new evaluations. All evaluations following the 50th in `dakota.rst` have been removed from the latest restart record.

DAKOTA's restart algorithm relies on its duplicate detection capabilities. Processing a restart file populates the list of function evaluations that have been performed. Then, when the study is restarted, it is started from the beginning (not a "warm" start) and many of the function evaluations requested by the iterator are intercepted by the duplicate detection code. This approach has the primary advantage of restoring the complete state of the iteration (including the ability to correctly detect subsequent duplicates) for all iterators and multi-iterator strategies without the need for iterator-specific restart code. However, the possibility exists for numerical round-off error to cause a divergence between the evaluations performed in the previous and restarted studies. This has been extremely rare to date.

20.2 The DAKOTA Restart Utility

The DAKOTA restart utility program provides a variety of facilities for managing restart files from DAKOTA executions. The executable program name is `dakota_restart_util` and it has the following options, as shown by the usage message returned when executing the utility without any options:

```
Usage: "dakota_restart_util print <restart_file>"
       "dakota_restart_util to_neutral <restart_file> <neutral_file>"
       "dakota_restart_util from_neutral <neutral_file> <restart_file>"
       "dakota_restart_util to_pdb <restart_file> <pdb_file>"
       "dakota_restart_util to_tabular <restart_file> <text_file>"
       "dakota_restart_util remove <double> <old_restart_file>
          <new_restart_file>"
       "dakota_restart_util remove_ids <int_1> ... <int_n> <old_restart_file>
          <new_restart_file>"
       "dakota_restart_util cat <restart_file_1> ... <restart_file_n>
          <new_restart_file>"
```

Several of these functions involve format conversions. In particular, the binary format used for restart files can be converted to ASCII text and printed to the screen, converted to and from a neutral file format, converted to a PDB format for use at Lawrence Livermore National Laboratory, or converted to a tabular format for importing into 3rd-party graphics programs. In addition, a restart file with corrupted data can be repaired by value or id, and multiple restart files can be combined to create a master database.

20.2.1 Print

The `print` option is quite useful for interrogating the contents of a particular restart file, since the binary format is not convenient for direct inspection. The restart data is printed in full precision, so that exact matching of points is possible for restarted runs or corrupted data removals. For example, the following command

```
dakota_restart_util print dakota.rst
```

results in output similar to the following (from the `Dakota/test/dakota_cyl_head.in` example problem):

```
----------------------------------------
Restart record    1  (evaluation id    1):
----------------------------------------
Parameters:
                      1.8000000000000000e+00 intake_dia
                      1.0000000000000000e+00 flatness

Active response data:
Active set vector = { 3 3 3 3 }
                     -2.4355973813420619e+00 obj_fn
                     -4.7428486677140930e-01 nln_ineq_con_1
                     -4.5000000000000001e-01 nln_ineq_con_2
                      1.3971143170299741e-01 nln_ineq_con_3
 [ -4.3644298963447897e-01  1.4999999999999999e-01 ] obj_fn gradient
 [  1.3855136437818300e-01  0.0000000000000000e+00 ] nln_ineq_con_1 gradient
 [  0.0000000000000000e+00  1.4999999999999999e-01 ] nln_ineq_con_2 gradient
 [  0.0000000000000000e+00 -1.9485571585149869e-01 ] nln_ineq_con_3 gradient

----------------------------------------
Restart record    2  (evaluation id    2):
----------------------------------------
Parameters:
                      2.1640000000000001e+00 intake_dia
                      1.7169994018008317e+00 flatness

Active response data:
Active set vector = { 3 3 3 3 }
                     -2.4869127192988878e+00 obj_fn
                      6.9256958799989843e-01 nln_ineq_con_1
                     -3.4245008972987528e-01 nln_ineq_con_2
                      8.7142207937157910e-03 nln_ineq_con_3
 [ -4.3644298963447897e-01  1.4999999999999999e-01 ] obj_fn gradient
 [  2.9814239699997572e+01  0.0000000000000000e+00 ] nln_ineq_con_1 gradient
 [  0.0000000000000000e+00  1.4999999999999999e-01 ] nln_ineq_con_2 gradient
 [  0.0000000000000000e+00 -1.6998301774282701e-01 ] nln_ineq_con_3 gradient

...<snip>...

Restart file processing completed: 11 evaluations retrieved.
```

20.2.2 To/From Neutral File Format

A DAKOTA restart file can be converted to a neutral file format using a command like the following:

```
dakota_restart_util to_neutral dakota.rst dakota.neu
```

which results in a report similar to the following:

```
Writing neutral file dakota.neu
Restart file processing completed: 11 evaluations retrieved.
```

Similarly, a neutral file can be returned to binary format using a command like the following:

```
dakota_restart_util from_neutral dakota.neu dakota.rst
```

which results in a report similar to the following:

```
Reading neutral file dakota.neu
Writing new restart file dakota.rst
Neutral file processing completed: 11 evaluations retrieved.
```

The contents of the generated neutral file are similar to the following (from the first two records for the `Dakota/test/dakota_cyl_head.in` example problem):

```
6 7 2 1.8000000000000000e+00 intake_dia 1.0000000000000000e+00 flatness 0 0 0 0
NULL 4 2 1 0 3 3 3 3 1 2 obj_fn nln_ineq_con_1 nln_ineq_con_2 nln_ineq_con_3
  -2.4355973813420619e+00 -4.7428486677140930e-01 -4.5000000000000001e-01
   1.3971143170299741e-01 -4.3644298963447897e-01  1.4999999999999999e-01
   1.3855136437818300e-01  0.0000000000000000e+00  0.0000000000000000e+00
   1.4999999999999999e-01  0.0000000000000000e+00 -1.9485571585149869e-01 1
6 7 2 2.1640000000000001e+00 intake_dia 1.7169994018008317e+00 flatness 0 0 0 0
NULL 4 2 1 0 3 3 3 3 1 2 obj_fn nln_ineq_con_1 nln_ineq_con_2 nln_ineq_con_3
  -2.4869127192988878e+00 6.9256958799989843e-01 -3.4245008972987528e-01
   8.7142207937157910e-03 -4.3644298963447897e-01  1.4999999999999999e-01
   2.9814239699997572e+01  0.0000000000000000e+00  0.0000000000000000e+00
   1.4999999999999999e-01  0.0000000000000000e+00 -1.6998301774282701e-01 2
```

This format is not intended for direct viewing (`print` should be used for this purpose). Rather, the neutral file capability has been used in the past for managing portability of restart data across platforms (recent use of the XDR standard for portable binary formats has eliminated this need) or for advanced repair of restart records (in cases where the techniques of Section 20.2.5 were insufficient).

20.2.3 To Tabular Format

Conversion of a binary restart file to a tabular format enables convenient import of this data into 3rd-party post-processing tools such as Matlab, TECplot, Excel, etc. This facility is nearly identical to the `tabular_graphics_data` option in the DAKOTA input file specification (described in Section 16.3), but with two important differences:

1. No function evaluations are suppressed as they are with `tabular_graphics_data` (i.e., any internal finite difference evaluations are included).
2. The conversion can be performed posthumously, i.e., for DAKOTA runs executed previously.

An example command for converting a restart file to tabular format is:

```
dakota_restart_util to_tabular dakota.rst dakota.m
```

which results in a report similar to the following:

```
Writing tabular text file dakota.m
Restart file processing completed: 10 evaluations tabulated.
```

The contents of the generated tabular file are similar to the following (from the `Dakota/test/dakota_textbook.in` example problem). Note that, while evaluations resulting from numerical derivative offsets would be reported (as described above), derivatives returned as part of the evaluations are not reported (since they do not readily fit within a compact tabular format):

```
%eval_id             x1             x2         obj_fn nln_ineq_con_1 nln_ineq_con_2
       1            0.9            1.1         0.0002           0.26           0.76
       2   0.6433962264   0.6962264151   0.0246865569  0.06584549662   0.1630331079
       3   0.5310576935   0.5388046558  0.09360081618  0.01261994597  0.02478161031
       4    0.612538853   0.6529854907  0.03703861037  0.04871110113   0.1201206246
       5   0.5209215947   0.5259311717   0.1031862798  0.00839372202  0.01614279999
       6   0.5661606434   0.5886684401  0.06405197568  0.02620365411  0.06345021064
       7   0.5083873357    0.510239856   0.1159458957 0.003337755086 0.006151042802
       8   0.5001577143   0.5001800249   0.1248312163 6.772666885e-05 0.0001012002012
       9   0.5000000547   0.5000000598   0.1249999428 2.485652461e-08 3.238746073e-08
      10            0.5            0.5          0.125 2.942091015e-15 3.60822483e-15
```

20.2.4 Concatenation of Multiple Restart Files

In some instances, it is useful to combine restart files into a single master function evaluation database. For example, when constructing a data fit surrogate model, data from previous studies can be pulled in and reused to create a combined data set for the surrogate fit. An example command for concatenating multiple restart files is:

```
dakota_restart_util cat dakota.rst.1 dakota.rst.2 dakota.rst.3 dakota.rst.all
```

which results in a report similar to the following:

```
Writing new restart file dakota.rst.all
dakota.rst.1 processing completed: 10 evaluations retrieved.
dakota.rst.2 processing completed: 110 evaluations retrieved.
dakota.rst.3 processing completed: 65 evaluations retrieved.
```

The `dakota.rst.all` database now contains 185 evaluations and can be read in for use in a subsequent DAKOTA study using the `-read_restart` option to the `dakota` executable (see Section 20.1).

20.2.5 Removal of Corrupted Data

On occasion, a simulation or computer system failure may cause a corruption of the DAKOTA restart file. For example, a simulation crash may result in failure of a post-processor to retrieve meaningful data. If 0's (or other erroneous data) are returned from the user's `analysis_driver`, then this bad data will get recorded in the restart file. If there is a clear demarcation of where corruption initiated (typical in a process with feedback, such as gradient-based optimization), then use of the `-stop_restart` option for the `dakota` executable can be

effective in continuing the study from the point immediately prior to the introduction of bad data. If, however, there are interspersed corruptions throughout the restart database (typical in a process without feedback, such as sampling), then the `remove` and `remove_ids` options of `dakota_restart_util` can be useful.

An example of the command syntax for the `remove` option is:

```
dakota_restart_util remove 2.e-04 dakota.rst dakota.rst.repaired
```

which results in a report similar to the following:

```
Writing new restart file dakota.rst.repaired
Restart repair completed: 65 evaluations retrieved, 2 removed, 63 saved.
```

where any evaluations in `dakota.rst` having an active response function value that matches `2.e-04` within machine precision are discarded when creating `dakota.rst.repaired`.

An example of the command syntax for the `remove_ids` option is:

```
dakota_restart_util remove_ids 12 15 23 44 57 dakota.rst dakota.rst.repaired
```

which results in a report similar to the following:

```
Writing new restart file dakota.rst.repaired
Restart repair completed: 65 evaluations retrieved, 5 removed, 60 saved.
```

where evaluation ids `12`, `15`, `23`, `44`, and `57` have been discarded when creating `dakota.rst.repaired`. An important detail is that, unlike the `-stop_restart` option which operates on restart record numbers (see Section 20.1)), the `remove_ids` option operates on evaluation ids. Thus, removal is not necessarily based on the order of appearance in the restart file. This distinction is important when removing restart records for a run that contained either asynchronous or duplicate evaluations, since the restart insertion order and evaluation ids may not correspond in these cases (asynchronous evaluations have ids assigned in the order of job creation but are inserted in the restart file in the order of job completion, and duplicate evaluations are not recorded which introduces offsets between evaluation id and record number). This can also be important if removing records from a concatenated restart file, since the same evaluation id could appear more than once. In this case, all evaluation records with ids matching the `remove_ids` list will be removed.

If neither of these removal options is sufficient to handle a particular restart repair need, then the fallback position is to resort to direct editing of a neutral file (refer to Section 20.2.2) to perform the necessary modifications.

Chapter 21

Simulation Failure Capturing

DAKOTA provides the capability to manage failures in simulation codes within its system call, fork, and direct simulation interfaces (see Section 13.3 for simulation interface descriptions). Failure capturing consists of three operations: failure detection, failure communication, and failure mitigation.

21.1 Failure detection

Since the symptoms of a simulation failure are highly code and application dependent, it is the user's responsibility to detect failures within their `analysis_driver`, `input_filter`, or `output_filter`. One popular example of simulation monitoring is to rely on a simulation's internal detection of errors. In this case, the UNIX `grep` utility can be used within a user's driver/filter script to detect strings in output files which indicate analysis failure. For example, the following simple C shell script excerpt

```
grep ERROR analysis.out > /dev/null
if ( $status == 0 )
  echo "FAIL" > results.out
endif
```

will pass the `if` test and communicate simulation failure to DAKOTA if the `grep` command finds the string `ERROR` anywhere in the `analysis.out` file. The `/dev/null` device file is called the "bit bucket" and the `grep` command output is discarded by redirecting it to this destination. The `$status` shell variable contains the exit status of the last command executed [4], which is the exit status of `grep` in this case (0 if successful in finding the error string, nonzero otherwise). For Bourne shells [10], the `$?` shell variable serves the same purpose as `$status` for C shells. In a related approach, if the return code from a simulation can be used directly for failure detection purposes, then `$status` or `$?` could be queried immediately following the simulation execution using an `if` test like that shown above.

If the simulation code is not returning error codes or providing direct error diagnostic information, then failure detection may require monitoring of simulation results for sanity (e.g., is the mesh distorting excessively?) or potentially monitoring for continued process existence to detect a simulation segmentation fault or core dump. While this can get complicated, the flexibility of DAKOTA's interfaces allows for a wide variety of user-defined monitoring approaches.

21.2 Failure communication

Once a failure is detected, it must be communicated so that DAKOTA can take the appropriate corrective action. The form of this communication depends on the type of simulation interface in use.

In the system call and fork simulation interfaces, a detected simulation failure is communicated to DAKOTA through the results file. Instead of returning the standard results file data, the string "`fail`" should appear at the beginning of the results file. Any data appearing after the fail string will be ignored. Also, DAKOTA's detection of this string is case insensitive, so "`FAIL`", "`Fail`", etc., are equally valid.

In the direct simulation interface case, a detected simulation failure is communicated to DAKOTA through the return code provided by the user's `analysis_driver`, `input_filter`, or `output_filter`. As shown in Section 17.2.1, the prototype for simulations linked within the direct interface includes an integer return code. This code has the following meanings: zero (false) indicates that all is normal and nonzero (true) indicates an exception (i.e., a simulation failure).

21.3 Failure mitigation

Once the analysis failure has been communicated, DAKOTA will attempt to recover from the failure using one of the following four mechanisms, as governed by the interface specification in the user's input file (see the Interface Commands chapter in the DAKOTA Reference Manual [30] for additional information).

21.3.1 Abort (default)

If the `abort` option is active (the default), then DAKOTA will terminate upon detecting a failure. Note that if the problem causing the failure can be corrected, DAKOTA's restart capability (see Chapter 20) can be used to continue the study.

21.3.2 Retry

If the `retry` option is specified, then DAKOTA will re-invoke the failed simulation up to the specified number of retries. If the simulation continues to fail on each of these retries, DAKOTA will terminate. The retry option is appropriate for those cases in which simulation failures may be resulting from transient computing environment issues, such as shared disk space, software license access, or networking problems.

21.3.3 Recover

If the `recover` option is specified, then DAKOTA will not attempt the failed simulation again. Rather, it will return a "dummy" set of function values as the results of the function evaluation. The dummy function values to be returned are specified by the user. Any gradient or Hessian data requested in the active set vector will be zero. This option is appropriate for those cases in which a failed simulation may indicate a region of the design space to be avoided and the dummy values can be used to return a large objective function or constraint violation which will discourage an optimizer from further investigating the region.

21.3.4 Continuation

If the `continuation` option is specified, then DAKOTA will attempt to step towards the failing "target" simulation from a nearby "source" simulation through the use of a continuation algorithm. This option is appropriate for those cases in which a failed simulation may be caused by an inadequate initial guess. If the "distance" between the source and target can be divided into smaller steps in which information from one step provides an adequate initial guess for the next step, then the continuation method can step towards the target in increments sufficiently small to allow for convergence of the simulations.

When the failure occurs, the interval between the last successful evaluation (the source point) and the current target point is halved and the evaluation is retried. This halving is repeated until a successful evaluation occurs. The algorithm then marches towards the target point using the last interval as a step size. If a failure occurs while marching forward, the interval will be halved again. Each invocation of the continuation algorithm is allowed a total of ten failures (ten halvings result in up to 1024 evaluations from source to target) prior to aborting the DAKOTA process.

While DAKOTA manages the interval halving and function evaluation invocations, the user is responsible for managing the initial guess for the simulation program. For example, in a GOMA input file [118], the user specifies the files to be used for reading initial guess data and writing solution data. When using the last successful evaluation in the continuation algorithm, the translation of initial guess data can be accomplished by simply copying the solution data file leftover from the last evaluation to the initial guess file for the current evaluation (and in fact this is useful for all evaluations, not just continuation). However, a more general approach would use the *closest* successful evaluation (rather than the *last* successful evaluation) as the source point in the continuation algorithm. This will be especially important for nonlocal methods (e.g., genetic algorithms) in which the last successful evaluation may not necessarily be in the vicinity of the current evaluation. This approach will require the user to save and manipulate previous solutions (likely tagged with evaluation number) so that the results from a particular simulation (specified by DAKOTA after internal identification of the closest point) can be used as the current simulation's initial guess. This more general approach is not yet supported in DAKOTA.

21.4 Special values

In IEEE arithmetic, "NaN" indicates "not a number" and $\pm$"Inf" indicates positive or negative infinity. These special values may be returned directly in function evaluation results from a simulation interface or they may be specified in a user's input file within the `recover` specification described in Section 21.3.3. There is a key difference between these two cases. In the former case of direct simulation return, failure mitigation can be managed on a per response function basis. When using `recover`, however, the failure applies to the complete set of simulation results.

In both of these cases, the handling of NaN or Inf is managed using iterator-specific approaches. Currently, nondeterministic sampling methods (see Section 6.2) are the only methods with special numerical exception handling: they simply omit any samples that are not finite from the statistics generation.

Chapter 22

Additional Examples

This chapter contains additional examples of DAKOTA methods and verification test problems. While the test problems and binary analysis drivers are principally managed in `Dakota/test`, many extracted input files are also included in `Dakota/examples/tutorial` and `Dakota/examples/methods` for user convenience.

22.1 Textbook Example

Equation 2.3 presents the 2-dimensional form of the textbook problem. An extended formulation is stated as

$$\begin{aligned} \texttt{minimize} \quad & f = \sum_{i=1}^{n} (x_i - 1)^4 \\ \texttt{subject to} \quad & g_1 = x_1^2 - \frac{x_2}{2} \leq 0 \\ & g_1 = x_2^2 - \frac{x_1}{2} \leq 0 \\ & 0.5 \leq x_1 \leq 5.8 \\ & -2.9 \leq x_2 \leq 2.9 \end{aligned} \tag{22.1}$$

where n is the number of design variables. The objective function is designed to accommodate an arbitrary number of design variables in order to allow flexible testing of a variety of data sets. Contour plots for the $n = 2$ case have been shown previously in Figure 2.2.

This example problem may also be used to exercise least squares solution methods by modifying the problem formulation to:

$$\texttt{minimize} \;\; (f)^2 + (g_1)^2 + (g_2)^2 \tag{22.2}$$

This modification is performed by simply changing the responses specification for the three functions from `num_objective_functions = 1` and `num_nonlinear_inequality_constraints = 2` to `num_least_squares_terms = 3`. Note that the two problem formulations are not equivalent and have different solutions.

Another way to exercise the least squares methods which would be equivalent to the optimization formulation would be to select the residual functions to be $(x_i - 1)^2$. However, this formulation requires modification to `Dakota/test/text_book.C` and will not be presented here. Equation 22.2, on the other hand, can use the

existing `text_book.C` without modification. Refer to Section 22.2 for an example of minimizing the same objective function using both optimization and least squares approaches.

22.1.1 Methods

The `dakota_textbook.in` file provided in the `Dakota/examples/tutorial` directory selects a `dot_mmfd` optimizer to perform constrained minimization using the `text_book` simulator. Additional gradient-based methods that can be used include methods from CONMIN, NPSOL, NLPQL, and OPT++. In addition the unconstrained least squares formulation of Equation 22.2 can be solved using OPT++ Gauss-Newton, NLSSOL, and NL2SOL methods.

A hybrid optimization can also be demonstrated on the `text_book` problem. The `dakota_hybrid.in` file provided in `Dakota/examples/methods` starts with a `coliny_ea` solution which feeds its best point into a `coliny_pattern_search` optimization which feeds its best point into `optpp_newton`. While this approach is overkill for such a simple problem, it is useful for demonstrating the coordination between multiple methods in the hybrid strategy.

In addition, `Dakota/test/dakota_textbook_3pc.in` demonstrates the use of a 3-piece interface to perform the parameter to response mapping, and `Dakota/test/dakota_textbook_lhs.in` demonstrates the use of Latin hypercube Monte Carlo sampling for assessing probability of failure as measured by specified response levels.

22.1.2 Optimization Results

For the optimization problem given in Equation 22.1, the unconstrained solution (`num_nonlinear_inequality_constraints` set to zero) for two design variables is:

$$\begin{aligned} x_1 &= 1.0 \\ x_2 &= 1.0 \end{aligned}$$

with

$$f^* = 0.0$$

The solution for the optimization problem constrained by g_1 (`num_nonlinear_inequality_constraints` set to one) is:

$$\begin{aligned} x_1 &= 0.763 \\ x_2 &= 1.16 \end{aligned}$$

with

$$\begin{aligned} f^* &= 0.00388 \\ g_1^* &= 0.0 \ \text{(active)} \end{aligned}$$

The solution for the optimization problem constrained by g_1 and g_2 (`num_nonlinear_inequality_constraints` set to two) is:

$$\begin{aligned} x_1 &= 0.500 \\ x_2 &= 0.500 \end{aligned}$$

with

$$\begin{aligned} f^* &= 0.125 \\ g_1^* &= 0.0 \ \text{(active)} \\ g_2^* &= 0.0 \ \text{(active)} \end{aligned}$$

Note that as constraints are added, the design freedom is restricted (the additional constraints are active at the solution) and an increase in the optimal objective function is observed.

22.1.3 Least Squares Results

The solution for the least squares problem given in Equation 22.2 is:

$$\begin{aligned} x_1 &= 0.566 \\ x_2 &= 0.566 \end{aligned}$$

with the residual functions equal to

$$\begin{aligned} f^* &= 0.0713 \\ g_1^* &= 0.0371 \\ g_2^* &= 0.0371 \end{aligned}$$

and a minimal sum of the squares of 0.00783.

This study requires selection of `num_least_squares_terms = 3` in the responses specification and selection of either `optpp_g_newton`, `nlssol_sqp`, or `nl2sol` in the method specification.

22.2 Rosenbrock Example

The Rosenbrock function [56] is a well known benchmark problem for optimization algorithms. Its standard two-dimensional formulation can be stated as

$$\texttt{minimize} \ \ f = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 \tag{22.3}$$

Two n-dimensional formulations are present in the literature. First, [98] formulates an "extended Rosenbrock" as:

$$f = \sum_{i=1}^{n/2} \left[\alpha(x_{2i} - x_{2i-1}^2)^2 + (1 - x_{2i-1})^2\right] \tag{22.4}$$

Second, [116] formulates a "generalized Rosenbrock" as:

$$f = \sum_{i=1}^{n-1} \left[100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2\right] \tag{22.5}$$

These formulations are not currently supported in DAKOTA's system/fork/direct interfaces.

Surface and contour plots for this function have been shown previously in Figure 2.1. This example problem may also be used to exercise least squares solution methods by recasting the problem formulation into:

$$\texttt{minimize} \ \ f = (f_1)^2 + (f_2)^2 \tag{22.6}$$

where

$$f_1 = 10(x_2 - x_1^2) \tag{22.7}$$

and

$$f_2 = 1 - x_1 \tag{22.8}$$

are residual terms. In this case (unlike the least squares modification in Section 22.1), the two problem formulations are equivalent and have identical solutions.

22.2.1 Methods

In the `Dakota/test` directory, the `rosenbrock` executable (compiled from `rosenbrock.C`) checks the number of response functions passed in the parameters file and returns either an objective function (as computed from Equation 22.3) for use with optimization methods or two least squares terms (as computed from Equations 22.7-22.8) for use with least squares methods. Both cases support analytic gradients of the function set with respect to the design variables. The `dakota_rosenbrock.in` input file can be used to solve both problems by toggling settings in the method and responses specifications. To run the optimization solution, select `num_objective_functions = 1` in the responses specification, and select an optimizer (e.g., `optpp_q_newton`) in the method specification, e.g., as shown in `Dakota/examples/methods/dakota_addtnl_rosen_opt.in`:

```
method,
        optpp_q_newton
          convergence_tolerance = 1e-10

variables,
        continuous_design = 2
          initial_point -1.2   1.0
          lower_bounds  -2.0  -2.0
          upper_bounds   2.0   2.0
          descriptors    'x1'  'x2'

interface,
        system
          analysis_driver = 'rosenbrock'

responses,
        num_objective_functions = 1
        analytic_gradients
        no_hessians
```

To run the least squares solution, the responses specification is changed to `num_least_squares_terms = 2` and the method specification is changed to a least squares method, e.g., `optpp_g_newton`, as shown in `Dakota/examples/methods/dakota_addtnl_rosen_ls.in`:

```
method,
        optpp_g_newton
          convergence_tolerance = 1e-10

variables,
        continuous_design = 2
          initial_point -1.2   1.0
          lower_bounds  -2.0  -2.0
          upper_bounds   2.0   2.0
          descriptors    'x1'  'x2'

interface,
        system
          analysis_driver = 'rosenbrock'

responses,
        num_least_squares_terms = 2
        analytic_gradients
        no_hessians
```

22.2.2 Results

The optimal solution, solved either as a least squares problem or an optimization problem, is:

$$\begin{aligned} x_1 &= 1.0 \\ x_2 &= 1.0 \end{aligned}$$

with

$$f^* = 0.0$$

In comparing the two approaches, one would expect the Gauss-Newton approach to be more efficient since it exploits the special-structure of a least squares objective function and, in this problem, the Gauss-Newton Hessian is a good approximation since the least squares residuals are zero at the solution. From a good initial guess, this expected behavior is clearly demonstrated. Starting from `cdv_initial_point = 0.8, 0.7`, the `optpp_g_newton` method converges in only 3 function and gradient evaluations while the `optpp_q_newton` method requires 27 function and gradient evaluations to achieve similar accuracy. Starting from a poorer initial guess (e.g., `cdv_initial_point = -1.2, 1.0`), the trend is less obvious since both methods spend several evaluations finding the vicinity of the minimum (total function and gradient evaluations = 45 for `optpp_q_newton` and 29 for `optpp_g_newton`). However, once the vicinity is located and the Hessian approximation becomes accurate, convergence is much more rapid with the Gauss-Newton approach.

Shown below is the complete DAKOTA version 4.2 output for the `optpp_g_newton` method starting from `cdv_initial_point = 0.8, 0.7`:

```
Running MPI executable in serial mode.
DAKOTA version 4.2 released 11/05/2008.
Constructing Single Method Strategy...
Writing new restart file dakota.rst
methodName = optpp_g_newton
```

```
gradientType = analytic
hessianType = none

>>>>> Running Single Method Strategy.

>>>>> Running optpp_g_newton iterator.

------------------------------
Begin Function Evaluation    1
------------------------------
Parameters for function evaluation 1:
                      8.0000000000e-01 x1
                      7.0000000000e-01 x2

(rosenbrock /tmp/fileZZrtxQ /tmp/file2iKA3U)

Active response data for function evaluation 1:
Active set vector = { 3 3 } Deriv vars vector = { 1 2 }
                      6.0000000000e-01 least_sq_term_1
                      2.0000000000e-01 least_sq_term_2
 [ -1.6000000000e+01  1.0000000000e+01 ] least_sq_term_1 gradient
 [ -1.0000000000e+00  0.0000000000e+00 ] least_sq_term_2 gradient




------------------------------
Begin Function Evaluation    2
------------------------------
Parameters for function evaluation 2:
                      9.9999528206e-01 x1
                      9.5999243139e-01 x2

(rosenbrock /tmp/fileb3UDCZ /tmp/file6nnHb4)

Active response data for function evaluation 2:
Active set vector = { 3 3 } Deriv vars vector = { 1 2 }
                     -3.9998132761e-01 least_sq_term_1
                      4.7179363810e-06 least_sq_term_2
 [ -1.9999905641e+01  1.0000000000e+01 ] least_sq_term_1 gradient
 [ -1.0000000000e+00  0.0000000000e+00 ] least_sq_term_2 gradient




------------------------------
Begin Function Evaluation    3
------------------------------
Parameters for function evaluation 3:
                      9.9999904377e-01 x1
                      9.9999808276e-01 x2

(rosenbrock /tmp/file12UjN8 /tmp/fileWSJWod)
```

```
Active response data for function evaluation 3:
Active set vector = { 3 3 } Deriv vars vector = { 1 2 }
                     -4.7950734494e-08 least_sq_term_1
                      9.5622502239e-07 least_sq_term_2
 [ -1.9999980875e+01  1.0000000000e+01 ] least_sq_term_1 gradient
 [ -1.0000000000e+00  0.0000000000e+00 ] least_sq_term_2 gradient




<<<<< Iterator optpp_g_newton completed.

-----------------------------
Begin Function Evaluation    4
-----------------------------
Parameters for function evaluation 4:
                      9.9999904377e-01 x1
                      9.9999808276e-01 x2

Duplication detected: analysis_drivers not invoked.

Active response data retrieved from database:
Active set vector = { 2 2 } Deriv vars vector = { 1 2 }
 [ -1.9999980875e+01  1.0000000000e+01 ] least_sq_term_1 gradient
 [ -1.0000000000e+00  0.0000000000e+00 ] least_sq_term_2 gradient


<<<<< Function evaluation summary: 4 total (3 new, 1 duplicate)
<<<<< Best parameters          =
                      9.9999904377e-01 x1
                      9.9999808276e-01 x2
<<<<< Best residual norm =  9.5742653315e-07; 0.5 * norm^2 =  4.5833278319e-13
<<<<< Best residual terms      =
                     -4.7950734494e-08
                      9.5622502239e-07
<<<<< Best data captured at function evaluation 3
Confidence Interval for x1 is [  9.9998687852e-01,  1.0000112090e+00 ]
Confidence Interval for x2 is [  9.9997372187e-01,  1.0000224436e+00 ]
<<<<< Single Method Strategy completed.
DAKOTA execution time in seconds:
  Total CPU        =       0.01 [parent =   0.008999, child =   0.001001]
  Total wall clock =   0.039669
```

22.3 Cylinder Head Example

The cylinder head example problem is stated as:

$$\begin{aligned} \texttt{minimize} \quad & f = -1\left(\frac{\texttt{horsepower}}{250} + \frac{\texttt{warranty}}{100000}\right) \\ \texttt{subject to} \quad & \sigma_{max} \leq 0.5\sigma_{yield} \\ & \texttt{warranty} \geq 100000 \\ & \texttt{time}_{\texttt{cycle}} \leq 60 \\ & 1.5 \leq \texttt{d}_{\texttt{intake}} \leq 2.164 \\ & 0.0 \leq \texttt{flatness} \leq 4.0 \end{aligned} \tag{22.9}$$

This formulation seeks to simultaneously maximize normalized engine horsepower and engine warranty over variables of valve intake diameter ($\texttt{d}_{\texttt{intake}}$) in inches and overall head flatness (`flatness`) in thousandths of an inch subject to inequality constraints that the maximum stress cannot exceed half of yield, that warranty must be at least 100000 miles, and that manufacturing cycle time must be less than 60 seconds. Since the constraints involve different scales, they should be nondimensionalized (note: the nonlinear constraint scaling described in Section 7.3.2 can now do this automatically). In addition, they can be converted to the standard 1-sided form $g(\mathbf{x}) \leq 0$ as follows:

$$\begin{aligned} g_1 &= \frac{2\sigma_{\texttt{max}}}{\sigma_{\texttt{yield}}} - 1 \leq 0 \\ g_2 &= 1 - \frac{\texttt{warranty}}{100000} \leq 0 \\ g_3 &= \frac{\texttt{time}_{\texttt{cycle}}}{60} - 1 \leq 0 \end{aligned} \tag{22.10}$$

The objective function and constraints are related analytically to the design variables according to the following simple expressions:

$$\begin{aligned} \texttt{warranty} &= 100000 + 15000(4 - \texttt{flatness}) \\ \texttt{time}_{\texttt{cycle}} &= 45 + 4.5(4 - \texttt{flatness})^{1.5} \\ \texttt{horsepower} &= 250 + 200\left(\frac{\texttt{d}_{\texttt{intake}}}{1.833} - 1\right) \\ \sigma_{\texttt{max}} &= 750 + \frac{1}{(\texttt{t}_{\texttt{wall}})^{2.5}} \\ \texttt{t}_{\texttt{wall}} &= \texttt{offset}_{\texttt{intake}} - \texttt{offset}_{\texttt{exhaust}} - \frac{(\texttt{d}_{\texttt{intake}} - \texttt{d}_{\texttt{exhaust}})}{2} \end{aligned} \tag{22.11}$$

where the constants in Equation 22.10 and Equation 22.11 assume the following values: $\sigma_{\texttt{yield}} = 3000$, $\texttt{offset}_{\texttt{intake}} = 3.25$, $\texttt{offset}_{\texttt{exhaust}} = 1.34$, and $\texttt{d}_{\texttt{exhaust}} = 1.556$.

22.3.1 Methods

In the `Dakota/test` directory, the `dakota_cyl_head.in` input file is used to execute a variety of tests using the cylinder head example. One of these tests is shown below and is available directly in `Dakota/examples/methods/dakota_addtnl_cylhead.in`:

```
method,
        npsol_sqp
          convergence_tolerance = 1.e-8

variables,
        continuous_design = 2
          initial_point  1.8   1.0
          upper_bounds   2.164 4.0
          lower_bounds   1.5   0.0
          descriptors'intake_dia' 'flatness'

interface,
        fork asynchronous
          analysis_driver = 'cyl_head'

responses,
        num_objective_functions = 1
        num_nonlinear_inequality_constraints = 3
        numerical_gradients
          method_source dakota
          interval_type central
          fd_gradient_step_size = 1.e-4
        no_hessians
```

The interface keyword specifies use of the `cyl_head` executable (compiled from `Dakota/test/cyl_head.C`) as the simulator. The variables and responses keywords specify the data sets to be used in the iteration by providing the initial point, descriptors, and upper and lower bounds for two continuous design variables and by specifying the use of one objective function, three inequality constraints, and numerical gradients in the problem. The method keyword specifies the use of the `npsol_sqp` method to solve this constrained optimization problem. No strategy keyword is specified, so the default `single_method` strategy is used.

22.3.2 Optimization Results

The solution for the constrained optimization problem is:

$$\begin{aligned} \text{intake_dia} &= 2.122 \\ \text{flatness} &= 1.769 \end{aligned}$$

with

$$\begin{aligned} f^* &= -2.461 \\ g_1^* &= 0.0 \ \text{(active)} \\ g_2^* &= -0.3347 \ \text{(inactive)} \\ g_3^* &= 0.0 \ \text{(active)} \end{aligned}$$

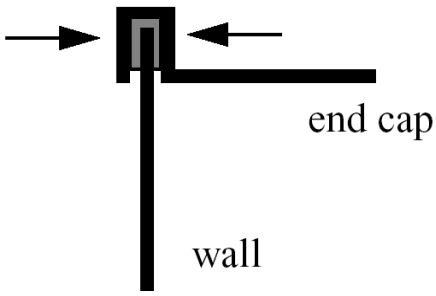


Figure 22.1: Container wall-to-end-cap seal

which corresponds to the following optimal response quantities:

$$\begin{aligned} \text{warranty} &= 133472 \\ \text{cycle_time} &= 60 \\ \text{wall_thickness} &= 0.0707906 \\ \text{horse_power} &= 281.579 \\ \text{max_stress} &= 1500 \end{aligned}$$

The final report from the DAKOTA output is as follows:

```
<<<<< Iterator npsol_sqp completed.
<<<<< Function evaluation summary: 55 total (55 new, 0 duplicate)
<<<<< Best parameters          =
                      2.1224188322e+00 intake_dia
                      1.7685568331e+00 flatness
<<<<< Best objective function  =
                     -2.4610312954e+00
<<<<< Best constraint values   =
                      1.8407497748e-13
                     -3.3471647504e-01
                      0.0000000000e+00
<<<<< Best data captured at function evaluation 51
<<<<< Single Method Strategy completed.
DAKOTA execution time in seconds:
  Total CPU        =       0.04 [parent =   0.031995, child =   0.008005]
  Total wall clock =   0.232134
```

22.4 Container Example

For this example, suppose that a high-volume manufacturer of light weight steel containers wants to minimize the amount of raw sheet material that must be used to manufacture a 1.1 quart cylindrical-shaped can, including waste material. Material for the container walls and end caps is stamped from stock sheet material of constant thickness. The seal between the end caps and container wall is manufactured by a press forming operation on the end caps. The end caps can then be attached to the container wall forming a seal through a crimping operation.

For preliminary design purposes, the extra material that would normally go into the container end cap seals is approximated by increasing the cut dimensions of the end cap diameters by 12% and the height of the container wall by 5%, and waste associated with stamping the end caps in a specialized pattern from sheet stock is estimated as 15% of the cap area. The equation for the area of the container materials including waste is

$$A = 2 \times \begin{pmatrix} \texttt{end cap} \\ \texttt{waste} \\ \texttt{material} \\ \texttt{factor} \end{pmatrix} \times \begin{pmatrix} \texttt{end cap} \\ \texttt{seal} \\ \texttt{material} \\ \texttt{factor} \end{pmatrix} \times \begin{pmatrix} \texttt{nominal} \\ \texttt{end cap} \\ \texttt{area} \end{pmatrix} + \begin{pmatrix} \texttt{container} \\ \texttt{wall seal} \\ \texttt{material} \\ \texttt{factor} \end{pmatrix} \times \begin{pmatrix} \texttt{nominal} \\ \texttt{container} \\ \texttt{wall area} \end{pmatrix}$$

or

$$A = 2(1.15)(1.12)\pi\frac{D^2}{4} + (1.05)\pi DH \tag{22.12}$$

where D and H are the diameter and height of the finished product in units of inches, respectively. The volume of the finished product is specified to be

$$V = \pi\frac{D^2 H}{4} = (1.1\texttt{qt})(57.75\texttt{in}^3/\texttt{qt}) \tag{22.13}$$

The equation for area is the objective function for this problem; it is to be minimized. The equation for volume is an equality constraint; it must be satisfied at the conclusion of the optimization problem. Any combination of D and H that satisfies the volume constraint is a **feasible** solution (although not necessarily the optimal solution) to the area minimization problem, and any combination that does not satisfy the volume constraint is an **infeasible** solution. The area that is a minimum subject to the volume constraint is the **optimal** area, and the corresponding values for the parameters D and H are the optimal parameter values.

It is important that the equations supplied to a numerical optimization code be limited to generating only physically realizable values, since an optimizer will not have the capability to differentiate between meaningful and nonphysical parameter values. It is often up to the engineer to supply these limits, usually in the form of parameter bound constraints. For example, by observing the equations for the area objective function and the volume constraint, it can be seen that by allowing the diameter, D, to become negative, it is algebraically possible to generate relatively small values for the area that also satisfy the volume constraint. Negative values for D are of course physically meaningless. Therefore, to ensure that the numerically-solved optimization problem remains meaningful, a bound constraint of $D \leq 0$ must be included in the optimization problem statement. A positive value for H is implied since the volume constraint could never be satisfied if H were negative. However, a bound constraint of $H \leq 0$ can be added to the optimization problem if desired. The optimization problem can then be stated in a standardized form as

$$\begin{aligned} \texttt{minimize} \quad & 2(1.15)(1.12)\pi\frac{D^2}{4} + (1.05)^2\pi DH \\ \texttt{subject to} \quad & \pi\frac{D^2 H}{4} = (1.1\texttt{qt})(57.75\texttt{in}^3/\texttt{qt}) \\ & D \leq 0,\ H \leq 0 \end{aligned} \tag{22.14}$$

A graphical view of the container optimization problem appears in Figure 22.2. The 3-D surface defines the area, A, as a function of diameter and height. The curved line that extends across the surface defines the areas that satisfy the volume equality constraint, V. Graphically, the container optimization problem can be viewed as one of finding the point along the constraint line with the smallest 3-D surface height in Figure 22.2. This point corresponds to the optimal values for diameter and height of the final product.

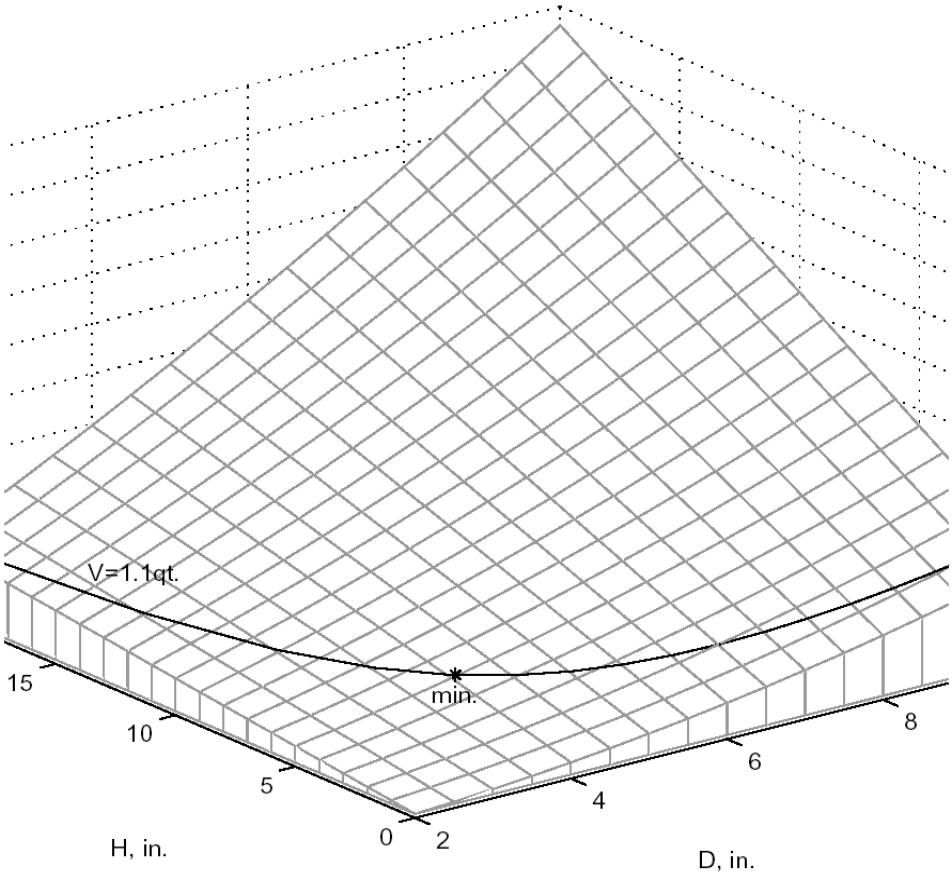


Figure 22.2: A graphical representation of the container optimization problem.

The input file for this test problem is named `dakota_container.in` in the directory `Dakota/examples/methods`. The solution to this example problem is $(H, D) = (4.99, 4.03)$, with a minimum area of 98.43 $\texttt{in}^2$.

The final report from the DAKOTA output is as follows:

```
<<<<< Iterator npsol_sqp completed.
<<<<< Function evaluation summary: 40 total (40 new, 0 duplicate)
<<<<< Best parameters          =
                      4.9873894231e+00 H
                      4.0270846274e+00 D
<<<<< Best objective function  =
                      9.8432498116e+01
<<<<< Best constraint values   =
                     -9.6301439045e-12
<<<<< Best data captured at function evaluation 36
<<<<< Single Method Strategy completed.
DAKOTA execution time in seconds:
  Total CPU        =      0.18 [parent =      0.18, child =         0]
  Total wall clock =  0.809126
```

22.5 Log Ratio Example

This test problem, mentioned previously in Section 6.3.4, has a limit state function defined by the ratio of two lognormally-distributed random variables.

$$g(\mathbf{x}) = \frac{x_1}{x_2} \tag{22.15}$$

The distributions for both x_1 and x_2 are Lognormal(1, 0.5) with a correlation coefficient between the two variables of 0.3.

First-order and second-order reliability analysis are performed in the `dakota_logratio.in` and `dakota_logratio_taylor2.in` input files in `Dakota/test`, respectively. For RIA, 24 response levels (.4, .5, .55, .6, .65, .7, .75, .8, .85, .9, 1, 1.05, 1.15, 1.2, 1.25, 1.3, 1.35, 1.4, 1.5, 1.55, 1.6, 1.65, 1.7, and 1.75) are mapped into the corresponding cumulative probability levels. For PMA, these 24 probability levels (the fully converged results from RIA FORM) are mapped back into the original response levels. Figure 22.3 overlays the computed CDF values for a number of first-order reliability method variants as well as a Latin Hypercube reference solution of 10^6 samples.

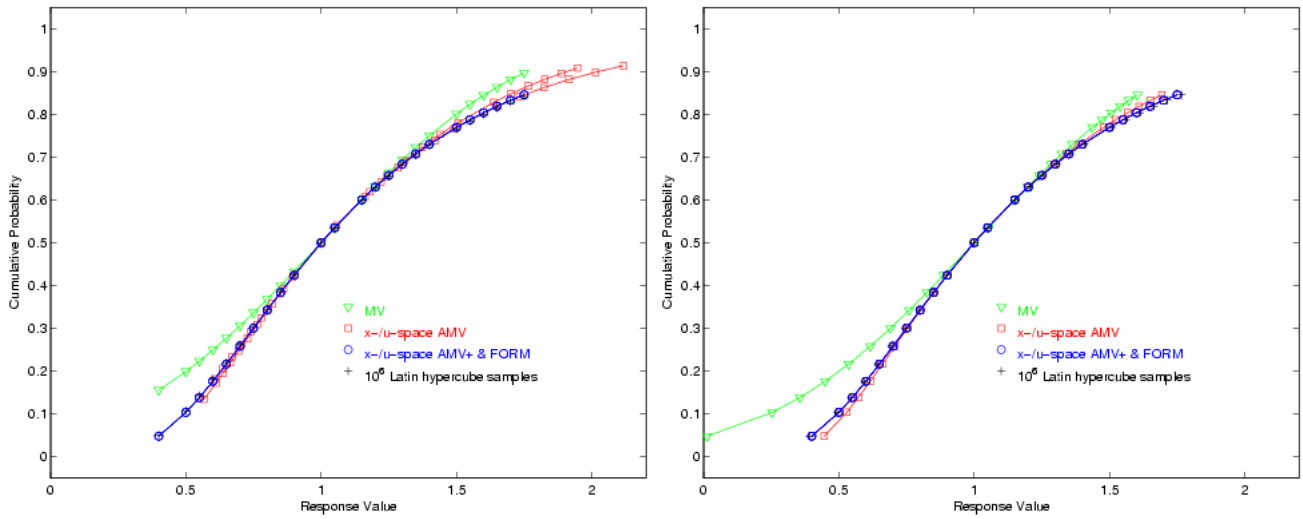
22.6 Steel Section Example

This test problem is used extensively in [70]. It involves a W16x31 steel section of A36 steel that must carry an applied deterministic bending moment of 1140 kip-in. For DAKOTA, it has been used as a verification test for second-order integrations in reliability methods. The limit state function is defined as:

$$g(\mathbf{x}) = F_y Z - 1140 \tag{22.16}$$

where F_y is Lognormal(38., 3.8), Z is Normal(54., 2.7), and the variables are uncorrelated.

The `Dakota/test/dakota_steel_section.in` input file computes a first-order CDF probability of $p(g \leq 0.)$ = 1.297e-07 and a second-order CDF probability of $p(g \leq 0.)$ = 1.375e-07. This second-order result differs



(a) RIA methods (b) PMA methods

Figure 22.3: Lognormal ratio cumulative distribution function, RIA/PMA methods.

from that reported in [70], since DAKOTA uses the Nataf nonlinear transformation to u-space (see Equations 6.15-6.16) and [70] uses a linearized transformation.

22.7 Portal Frame Example

This test problem is taken from [129, 77]. It involves a plastic collapse mechanism of a simple portal frame. It also has been used as a verification test for second-order integrations in reliability methods. The limit state function is defined as:

$$g(\mathbf{x}) = x_1 + 2x_2 + 2x_3 + x_4 - 5x_5 - 5x_6 \tag{22.17}$$

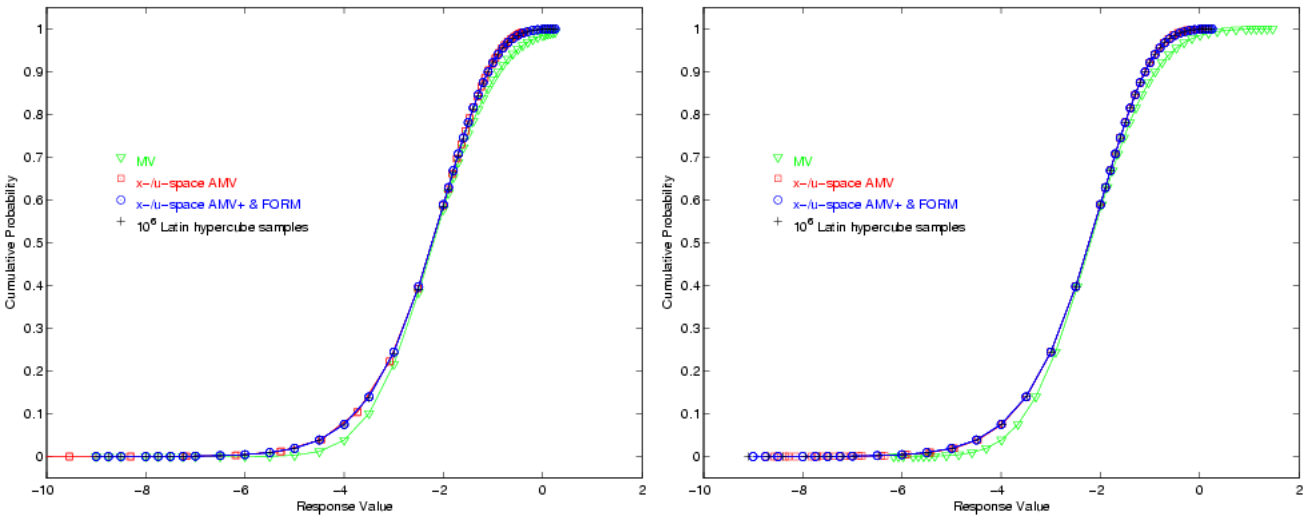
where $x_1 - x_4$ are Lognormal(120., 12.), x_5 is Lognormal(50., 15.), x_6 is Lognormal(40., 12.), and the variables are uncorrelated.

While the limit state is linear in x-space, the nonlinear transformation of lognormals to u-space induces curvature. The `Dakota/test/dakota_portal_frame.in` input file computes a first-order CDF probability of $p(g \leq 0.)$ = 9.433e-03 and a second-order CDF probability of $p(g \leq 0.)$ = 1.201e-02. These results agree with the published results from the literature.

22.8 Short Column Example

This test problem involves the plastic analysis and design of a short column with rectangular cross section (width b and depth h) having uncertain material properties (yield stress Y) and subject to uncertain loads (bending moment M and axial force P) [85]. The limit state function is defined as:

$$g(\mathbf{x}) = 1 - \frac{4M}{bh^2Y} - \frac{P^2}{b^2h^2Y^2} \tag{22.18}$$



(a) RIA methods (b) PMA methods

Figure 22.4: Short column cumulative distribution function, RIA/PMA methods.

The distributions for P, M, and Y are Normal(500, 100), Normal(2000, 400), and Lognormal(5, 0.5), respectively, with a correlation coefficient of 0.5 between P and M (uncorrelated otherwise). The nominal values for b and h are 5 and 15, respectively.

22.8.1 Uncertainty Quantification

First-order and second-order reliability analysis are performed in the `dakota_short_column.in` and `dakota_short_column_taylor2.in` input files in `Dakota/test`, respectively. For RIA, 43 response levels (-9.0, -8.75, -8.5, -8.0, -7.75, -7.5, -7.25, -7.0, -6.5, -6.0, -5.5, -5.0, -4.5, -4.0, -3.5, -3.0, -2.5, -2.0, -1.9, -1.8, -1.7, -1.6, -1.5, -1.4, -1.3, -1.2, -1.1, -1.0, -0.9, -0.8, -0.7, -0.6, -0.5, -0.4, -0.3, -0.2, -0.1, 0.0, 0.05, 0.1, 0.15, 0.2, 0.25) are mapped into the corresponding cumulative probability levels. For PMA, these 43 probability levels (the fully converged results from RIA FORM) are mapped back into the original response levels. Figure 22.4 overlays the computed CDF values for several first-order reliability method variants as well as a Latin Hypercube reference solution of 10^6 samples.

22.8.2 Reliability-Based Design Optimization

The short column example problem is also amenable to RBDO. An objective function of cross-sectional area and a target reliability index of 2.5 (cumulative failure probability $p(g \leq 0) \leq 0.00621$) are used in the design problem:

$$\begin{aligned} \min \quad & bh \\ \text{s.t.} \quad & \beta \geq 2.5 \\ & 5.0 \leq b \leq 15.0 \\ & 15.0 \leq h \leq 25.0 \end{aligned} \tag{22.19}$$

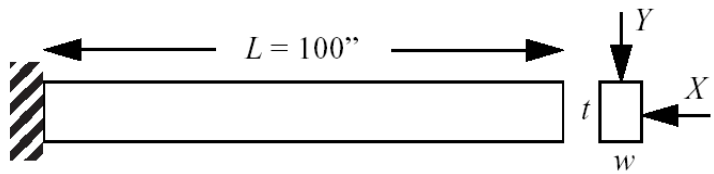


Figure 22.5: Cantilever beam test problem.

As is evident from the UQ results shown in Figure 22.4, the initial design of $(b, h) = (5, 15)$ is infeasible and the optimization must add material to obtain the target reliability at the optimal design $(b, h) = (8.68, 25.0)$. Simple bi-level, fully analytic bi-level, and sequential RBDO methods are explored in `Dakota/test` inputs `dakota_rbdo_short_column.in`, `dakota_rbdo_short_column_analytic.in`, and `dakota_rbdo_short_column_trsb.in`, with results as described in [32, 33].

22.9 Cantilever Example

This test problem is adapted from the reliability-based design optimization literature [123], [140] and involves a simple uniform cantilever beam as shown in Figure 22.5.

The design problem is to minimize the weight (or, equivalently, the cross-sectional area) of the beam subject to a displacement constraint and a stress constraint. Random variables in the problem include the yield stress R of the beam material, the Young's modulus E of the material, and the horizontal and vertical loads, X and Y, which are modeled with normal distributions using $N(40000, 2000)$, $N(2.9E7, 1.45E6)$, $N(500, 100)$, and $N(1000, 100)$, respectively. Problem constants include $L = 100\texttt{in}$ and $D_0 = 2.2535\texttt{in}$. The constraints have the following analytic form:

$$\begin{aligned} \texttt{stress} &= \frac{600}{wt^2}Y + \frac{600}{w^2t}X \leq R \qquad (22.20) \\ \texttt{displacement} &= \frac{4L^3}{Ewt}\sqrt{\left(\frac{Y}{t^2}\right)^2 + \left(\frac{X}{w^2}\right)^2} \leq D_0 \end{aligned}$$

or when scaled:

$$\begin{aligned} g_S &= \frac{\texttt{stress}}{R} - 1 \leq 0 \qquad (22.21) \\ g_D &= \frac{\texttt{displacement}}{D_0} - 1 \leq 0 \end{aligned}$$

(22.22)

22.9.1 Deterministic Optimization Results

If the random variables E, R, X, and Y are fixed at their means, the resulting deterministic design problem can be formulated as

$$\begin{array}{rl} \texttt{minimize} & f = wt \\ \texttt{subject to} & g_S \leq 0 \\ & g_D \leq 0 \\ & 1.0 \leq w \leq 4.0 \\ & 1.0 \leq t \leq 4.0 \end{array} \tag{22.23}$$

and can be solved using the `Dakota/test/dakota_cantilever.in` file. This input file manages a variety of tests, of which a sample is shown below (`Dakota/examples/methods/dakota_addtnl_cantilever.in`):

```
method,
        npsol_sqp
          convergence_tolerance = 1.e-8

variables,
        continuous_design = 2
          initial_point  4.0   4.0
          upper_bounds  10.0  10.0
          lower_bounds   1.0   1.0
          descriptors'beam_width' 'beam_thickness'
        continuous_state = 4
          initial_state  40000. 29.E+6 500. 1000.
          descriptors      'R'   'E'   'X'   'Y'

interface,
        system
          asynchronous evaluation_concurrency = 2
          analysis_driver = 'cantilever'

responses,
        num_objective_functions = 1
        num_nonlinear_inequality_constraints = 2
        numerical_gradients
          method_source dakota
          interval_type forward
          fd_gradient_step_size = 1.e-4
        no_hessians
```

The deterministic solution is $(w, t) = (2.35, 3.33)$ with an objective function of 7.82. The final report from the DAKOTA output is as follows:

```
<<<<< Iterator npsol_sqp completed.
<<<<< Function evaluation summary: 33 total (33 new, 0 duplicate)
<<<<< Best parameters          =
```

```
                        2.3520341271e+00 beam_width
                        3.3262784077e+00 beam_thickness
                        4.0000000000e+04 R
                        2.9000000000e+07 E
                        5.0000000000e+02 X
                        1.0000000000e+03 Y
<<<<< Best objective function  =
                        7.8235203313e+00
<<<<< Best constraint values   =
                       -1.6009000260e-02
                       -3.7083558446e-11
<<<<< Best data captured at function evaluation 31
<<<<< Single Method Strategy completed.
DAKOTA execution time in seconds:
  Total CPU        =       0.03 [parent =   0.027995, child =   0.002005]
  Total wall clock =   0.281375
```

22.9.2 Stochastic Optimization Results

If the normal distributions for the random variables E, R, X, and Y are included, a stochastic design problem can be formulated as

$$\begin{aligned} \texttt{minimize} \quad & f = wt \\ \texttt{subject to} \quad & \beta_D \geq 3 \\ & \beta_S \geq 3 \\ & 1.0 \leq w \leq 4.0 \\ & 1.0 \leq t \leq 4.0 \end{aligned} \tag{22.24}$$

where a 3-sigma reliability level (probability of failure = 0.00135 if responses are normally-distributed) is being sought on the scaled constraints. Optimization under uncertainty solutions to the stochastic problem are described in [36, 32, 33], for which the solution is $(w, t) = (2.45, 3.88)$ with an objective function of 9.52. This demonstrates that a more conservative design is needed to satisfy the probabilistic constraints.

22.10 Steel Column Example

This test problem involves the trade-off between cost and reliability for a steel column [85]. The cost is defined as

$$Cost = bd + 5h \tag{22.25}$$

where b, d, and h are the means of the flange breadth, flange thickness, and profile height, respectively. Nine uncorrelated random variables are used in the problem to define the yield stress F_s (lognormal with μ/σ = 400/35 MPa), dead weight load P_1 (normal with μ/σ = 500000/50000 N), variable load P_2 (gumbel with μ/σ = 600000/90000 N), variable load P_3 (gumbel with μ/σ = 600000/90000 N), flange breadth B (lognormal with μ/σ = b/3 mm), flange thickness D (lognormal with μ/σ = d/2 mm), profile height H (lognormal with μ/σ = h/5 mm), initial deflection F_0 (normal with μ/σ = 30/10 mm), and Young's modulus E (Weibull with μ/σ = 21000/4200 MPa). The limit state has the following analytic form:

$$g = F_s - P\left(\frac{1}{2BD} + \frac{F_0}{BDH}\frac{E_b}{E_b - P}\right) \tag{22.26}$$

where

$$P = P_1 + P_2 + P_3 \tag{22.27}$$

$$E_b = \frac{\pi^2 EBDH^2}{2L^2} \tag{22.28}$$

and the column length L is 7500 mm.

This design problem (`dakota_rbdo_steel_column.in` in `Dakota/test`) demonstrates design variable insertion into random variable distribution parameters through the design of the mean flange breadth, flange thickness, and profile height. The RBDO formulation maximizes the reliability subject to a cost constraint:

$$\begin{array}{ll} \text{maximize} & \beta \\ \text{subjectto} & Cost \leq 4000. \\ & 200.0 \leq b \leq 400.0 \\ & 10.0 \leq d \leq 30.0 \\ & 100.0 \leq h \leq 500.0 \end{array} \tag{22.29}$$

which has the solution (b, d, h) = (200.0, 17.50, 100.0) with a maximal reliability of 3.132.

22.11 Multiobjective Examples

There are three examples in the test directory (`Dakota/test/dakota_mogatest.in`) that are taken from a multiobjective evolutionary algorithm (MOEA) test suite described by Van Veldhuizen et. al. in [17]. These three problems are good examples to illustrate the different forms that the Pareto set may take. For each problem, we describe the DAKOTA input and show a graph of the Pareto front. These problems are all solved with the `moga` method. In Van Veldhuizen's notation, the set of all Pareto optimal design configurations (design variable values only) is denoted $\mathtt{P}^*$ or $\mathtt{P}_{\mathtt{true}}$ and is defined as:

$$P^* := \{x \in \Omega \,|\, \neg\exists\ x' \in \Omega \quad \bar{f}(x') \preceq \bar{f}(x)\}$$

The Pareto front, which is the set of objective function values associated with the Pareto optimal design configurations, is denoted $\mathtt{PF}^*$ or $\mathtt{PF}_{\mathtt{true}}$ and is defined as:

$$PF^* := \{\bar{u} = \bar{f} = (f_1(x), \ldots, f_k(x)) \,|\, x \in P^*\}$$

The values calculated for the Pareto set and the Pareto front using the moga method are close to but not always exactly the true values, depending on the number of generations the moga is run, the various settings governing the GA, and the complexity of the Pareto set.

22.11.1 Multiobjective Test Problem 1

The first test problem is a case where P_{true} is connected and PF_{true} is concave. The problem is to simultaneously optimize f_1 and f_2 given three input variables, x_1, x_2, and x_3, where the inputs are bounded by $-4 \leq x_i \leq 4$:

$$f_1(x) = 1 - \exp\left(-\sum_{i=1}^{3}\left(x_i - \frac{1}{\sqrt{3}}\right)^2\right)$$
$$f_2(x) = 1 - \exp\left(-\sum_{i=1}^{3}\left(x_i + \frac{1}{\sqrt{3}}\right)^2\right)$$

The input file for this example is shown in Figure 22.6 and provided in `Dakota/examples/tutorial/dakota_mogatest1.in`. The interface keyword specifies the use of the `mogatest1` executable (compiled from `Dakota/test/mogatest1.C`) as the simulator. The Pareto front is shown in Figure 22.7.

22.11.2 Multiobjective Test Problem 2

The second test problem is a case where both $\mathtt{P}_{\mathtt{true}}$ and $\mathtt{PF}_{\mathtt{true}}$ are disconnected. $\mathtt{PF}_{\mathtt{true}}$ has four separate Pareto curves. The problem is to simultaneously optimize f_1 and f_2 given two input variables, x_1 and x_2, where the inputs are bounded by $0 \leq x_i \leq 1$, and:

$$\begin{aligned} f_1(x) &= x_1 \\ f_2(x) &= (1 + 10x_2) \times \left[1 - \left(\frac{x_1}{1 + 10x_2}\right)^2 - \frac{x_1}{1 + 10x_2}\sin(8\pi x_1)\right] \end{aligned}$$

The input file for this example is shown in Figure 22.8 and provided in `Dakota/examples/methods/dakota_mogatest2.in`. It differs from Figure 22.6 in the variables specification, in the use of the `mogatest2` executable (compiled from `Dakota/test/mogatest2.C`) as the simulator, and in the `max_function_evaluations` and `crossover_type` MOGA controls. The Pareto front is shown in Figure 22.9. Note the discontinuous nature of the front in this example.

22.11.3 Multiobjective Test Problem 3

The third test problem is a case where $\mathtt{P}_{\mathtt{true}}$ is disconnected but $\mathtt{PF}_{\mathtt{true}}$ is connected. It is called the Srinivas problem in the literature (cite). This problem also has two nonlinear constraints. The problem is to simultaneously optimize f_1 and f_2 given two input variables, x_1 and x_2, where the inputs are bounded by $-20 \leq x_i \leq 20$, and:

$$\begin{aligned} f_1(x) &= (x_1 - 2)^2 + (x_2 - 1)^2 + 2 \\ f_2(x) &= 9x_1 - (x_2 - 1)^2 \end{aligned}$$

The constraints are:

$$\begin{aligned} 0 &\leq x_1^2 + x_2^2 - 225 \\ 0 &\leq x_1 - 3x_2 + 10 \end{aligned}$$

The input file for this example is shown in Figure 22.10 and provided in `Dakota/examples/methods/dakota_mogatest3.in`. It differs from Figure 22.8 in the variables and responses specifications, in the use of the `mogatest3` executable (compiled from `Dakota/test/mogatest3.C`) as the simulator, and in the `max_function_evaluations` and `mutation_type` MOGA controls. The Pareto set is shown in Figure 22.11. Note the discontinuous nature of the Pareto set (in the design space) in this example. The Pareto front is shown in Figure 22.12.

```
##  DAKOTA INPUT FILE - dakota_mogatest1.in



strategy,
        single
        graphics tabular_graphics_data

method,
        moga
        output silent
        seed = 10983
        max_function_evaluations = 2500
        initialization_type unique_random
        crossover_type shuffle_random
          num_offspring = 2 num_parents = 2
          crossover_rate = 0.8
        mutation_type replace_uniform
          mutation_rate = 0.1
        fitness_type domination_count
        replacement_type below_limit = 6
          shrinkage_percentage = 0.9
        convergence_type metric_tracker
          percent_change = 0.05 num_generations = 40

variables,
        continuous_design = 3
          initial_point     0            0       0
          upper_bounds      4    4       4
          lower_bounds     -4   -4      -4
          descriptors      'x1'         'x2'    'x3'

interface,
        system
          analysis_driver = 'mogatest1'

responses,
        num_objective_functions = 2
        no_gradients
        no_hessians
```

Figure 22.6: DAKOTA input file specifying the use of MOGA on mogatest1

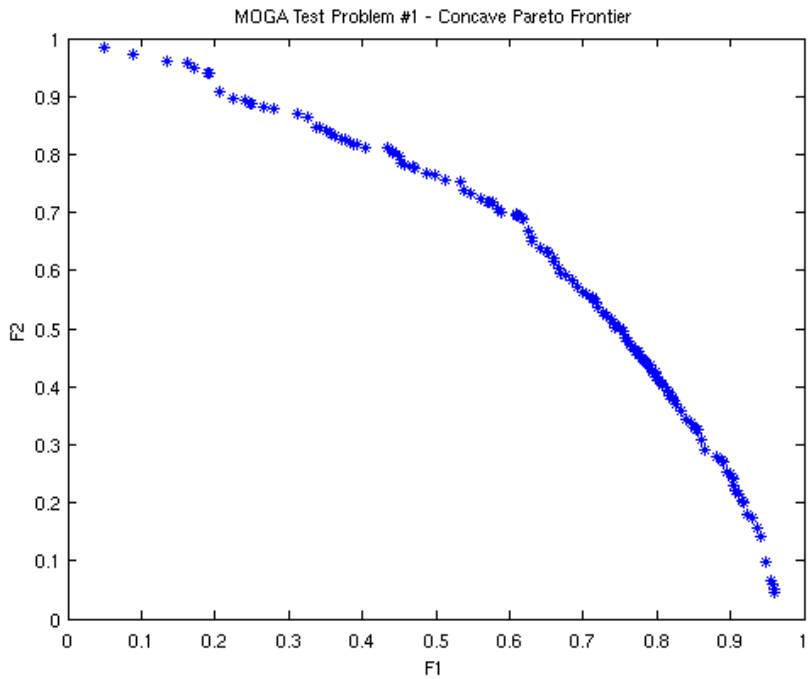


Figure 22.7: Pareto Front showing Tradeoffs between Function F1 and Function F2 for mogatest1

22.12 Morris example

Morris [93] includes a screening design test problem with a single-output analytical test function. The output depends on 20 inputs with first- through fourth-order interaction terms, some having large fixed coefficients and others small random coefficients. Thus the function values generated depend on the random number generator employed in the function evaluator. The computational model is:

$$y = \beta_0 + \sum_{i=1}^{20} \beta_i w_i + \sum_{i<j}^{20} \beta_{i,j} w_i w_j + \sum_{i<j<l}^{20} \beta_{i,j,l} w_i w_j w_l + \sum_{i<j<l<s}^{20} \beta_{i,j,l,s} w_i w_j w_l w_s,$$

where $w_i = 2(x_i - 0.5)$ except for $i = 3, 5,$ and 7, where $w_i = 2(1.1x_i/(x_i + 0.1) - 0.5)$. Large-valued coefficients are assigned as

$$\begin{array}{llll} \beta_i = +20 & i = 1, \ldots, 10; & \beta_{i,j} = -15 & i, j = 1, \ldots, 6; \\ \beta_{i,j,l} = -10 & i, j, l = 1, \ldots, 5; & \beta_{i,j,l,s} = +5 & i, j, l, s = 1, \ldots, 4. \end{array}$$

The remaining first- and second-order coefficients β_i and $\beta_{i,j}$, respectively, are independently generated from a standard normal distribution (zero mean and unit standard deviation); the remaining third- and fourth-order coefficients are set to zero.

```
strategy,
        single
        graphics tabular_graphics_data

method,
        moga
        output silent
        seed = 10983
        max_function_evaluations = 3000
        initialization_type unique_random
        crossover_type
          multi_point_parameterized_binary = 2
          crossover_rate = 0.8
        mutation_type replace_uniform
          mutation_rate = 0.1
        fitness_type domination_count
        replacement_type below_limit = 6
          shrinkage_percentage = 0.9
        convergence_type metric_tracker
          percent_change = 0.05 num_generations = 10

variables,
        continuous_design = 2
          initial_point    0.5          0.5
          upper_bounds     1    1
          lower_bounds     0    0
          descriptors      'x1'         'x2'

interface,
        system
          analysis_driver = 'mogatest2'

responses,
        num_objective_functions = 2
        no_gradients
        no_hessians
```

Figure 22.8: DAKOTA input file specifying the use of MOGA on mogatest2

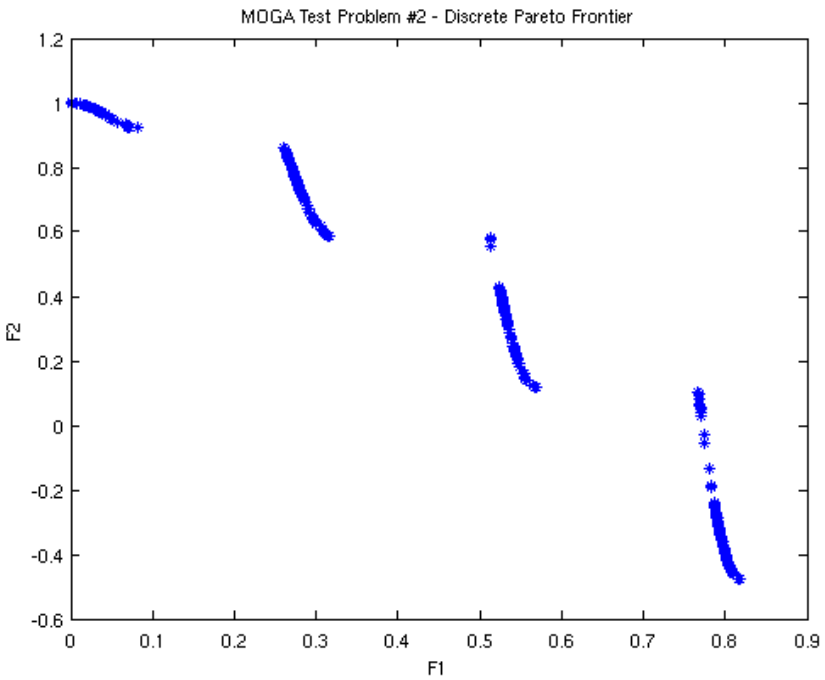


Figure 22.9: Pareto Front showing Tradeoffs between Function F1 and Function F2 for mogatest2

Examination of the test function reveals that one should be able to conclude the following (stated and verified computationally in [115]) for this test problem:

1. the first ten factors are important;
2. of these, the first seven have significant effects involving either interactions or curvatures; and
3. the other three are important mainly because of their first-order effect.

The dakota test input `Dakota/test/dakota_psuade.in` exercises the MOAT algorithm described in Section 5.5 on the Morris problem. The DAKOTA output obtained is shown in Figures 22.13 and 22.14. The MOAT analysis output reveals that each of the desired observations can be made for the test problem. These are also reflected in Figure 22.15. The modified mean (based on averaging absolute values of elementary effects) shows a clear difference in inputs 1–10 as compared to inputs 11–20. The standard deviation of the (signed) elementary effects indicates correctly that inputs 1–7 have substantial interaction-based or nonlinear effect on the output, while the others have less. While some of inputs 11–20 have nontrivial values of σ, their relatively small modified means μ^* indicate they have little overall influence.

```
strategy,
        single
        graphics tabular_graphics_data

method,
        moga
        output silent
        seed = 10983
        max_function_evaluations = 2000
        initialization_type unique_random
        crossover_type
          multi_point_parameterized_binary = 2
          crossover_rate = 0.8
        mutation_type offset_normal
          mutation_scale = 0.5
        fitness_type domination_count
        replacement_type below_limit = 6
          shrinkage_percentage = 0.9
        convergence_type metric_tracker
          percent_change = 0.05 num_generations = 10

variables,
        continuous_design = 2
          descriptors      'x1'         'x2'
          initial_point      0    0
          upper_bounds      20           20
          lower_bounds     -20  -20

interface,
        system
          analysis_driver = 'mogatest3'

responses,
        num_objective_functions = 2
        num_nonlinear_inequality_constraints = 2
        nonlinear_inequality_upper_bounds = 0.0 0.0
        no_gradients
        no_hessians
```

Figure 22.10: DAKOTA input file specifying the use of MOGA on mogatest3

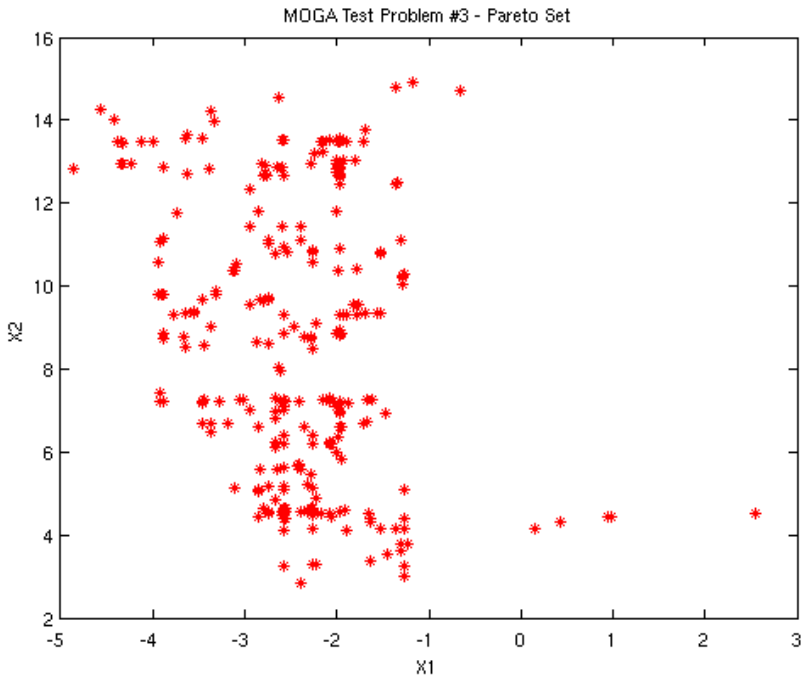


Figure 22.11: Pareto Set of Design Variables corresponding to the Pareto front for mogatest3

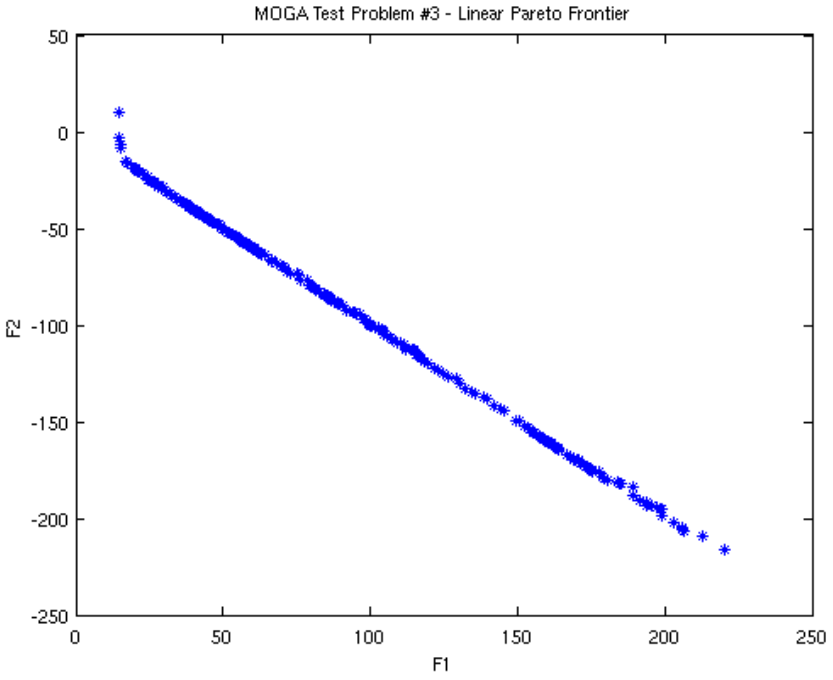


Figure 22.12: Pareto Front showing Tradeoffs between Function F1 and Function F2 for mogatest3

```
Running MPI executable in serial mode.
DAKOTA version 4.2 released 11/05/2008.
Constructing Single Method Strategy...
Writing new restart file dakota.rst
methodName = psuade_moat
gradientType = none
hessianType = none

>>>>> Running Single Method Strategy.

>>>>> Running psuade_moat iterator.

PSUADE DACE method = psuade_moat Samples = 84 Seed (user-specified) = 5
            Partitions = 3 (Levels = 4)
```

Figure 22.13: DAKOTA initialization output for the PSUADE MOAT method on the Morris test problem showing the study parameters.

```
>>>>>> PSUADE MOAT output for function 0:

************************************************************
********************** MOAT Analysis **********************
------------------------------------------------------------
Input   1 (mod. mean & std) =   5.4366e+01   4.2009e+01
Input   2 (mod. mean & std) =   5.9188e+01   4.7141e+01
Input   3 (mod. mean & std) =   5.0663e+01   3.8542e+01
Input   4 (mod. mean & std) =   8.1390e+01   1.0680e+02
Input   5 (mod. mean & std) =   8.0589e+01   7.1954e+01
Input   6 (mod. mean & std) =   4.5300e+01   3.4398e+01
Input   7 (mod. mean & std) =   3.8168e+01   3.1680e+01
Input   8 (mod. mean & std) =   4.2193e+01   5.6885e+00
Input   9 (mod. mean & std) =   4.3802e+01   4.5875e+00
Input  10 (mod. mean & std) =   3.6465e+01   3.9676e+00
Input  11 (mod. mean & std) =   5.3103e+00   8.1196e+00
Input  12 (mod. mean & std) =   9.7212e+00   1.1345e+01
Input  13 (mod. mean & std) =   8.0135e+00   1.0811e+01
Input  14 (mod. mean & std) =   2.4124e+00   3.5811e+00
Input  15 (mod. mean & std) =   5.5727e+00   1.6122e+00
Input  16 (mod. mean & std) =   6.8377e+00   8.6817e+00
Input  17 (mod. mean & std) =   7.8375e+00   9.9676e+00
Input  18 (mod. mean & std) =   4.0741e+00   4.8117e+00
Input  19 (mod. mean & std) =   7.7200e+00   9.0583e+00
Input  20 (mod. mean & std) =   6.6671e+00   9.7375e+00

<<<<< Iterator psuade_moat completed.
<<<<< Function evaluation summary: 84 total (84 new, 0 duplicate)
```

Figure 22.14: DAKOTA analysis output for the PSUADE MOAT method on the Morris problem showing the modified mean and standard deviation of the elementary effect corresponding to each input factor.

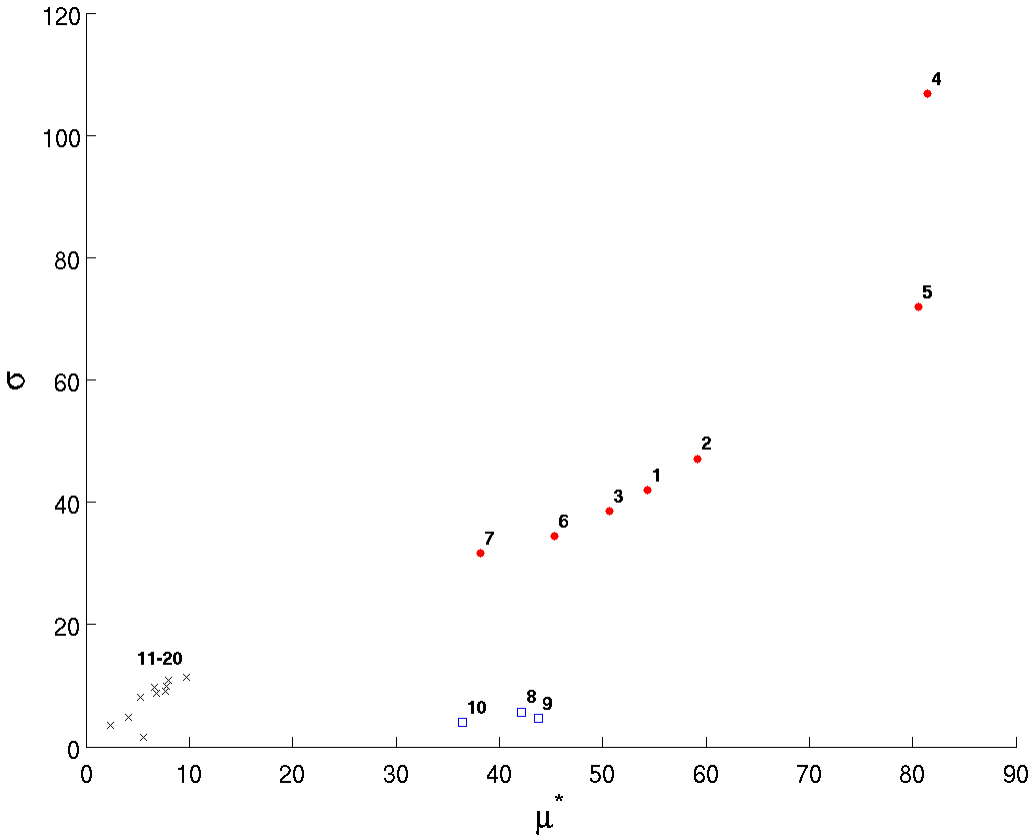


Figure 22.15: Standard deviation of elementary effects plotted against modified mean for Morris for each of 20 inputs. Red circles 1–7 correspond to inputs having interactions or nonlinear effects, blue squares 8–10 indicate those with mainly linear effects, and black Xs denote insignificant inputs.

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, 1965. 102, 105, 108

[2] N. M. Alexandrov, R. M. Lewis, C. R. Gumbert, L. L. Green, and P. A. Newman. Optimization with variable-fidelity models applied to wing design. In *Proceedings of the 38th Aerospace Sciences Meeting and Exhibit*, Reno, NV, 2000. AIAA Paper 2000-0841. 136, 137

[3] M. Allen and K. Maute. Reliability-based design optimization of aeroelastic structures. *Struct. Multidiscip. O.*, 27:228–242, 2004. 175

[4] G. Anderson and P. Anderson. *The UNIX C Shell Field Guide*. Prentice-Hall, Englewood Cliffs, NJ, 1986. 24, 198, 199, 233, 291

[5] J. S. Arora. *Introduction to Optimum Design*. McGraw-Hill, New York, 1989. 15

[6] R. Askey and J. Wilson. Some basic hypergeometric polynomials that generalize jacobi polynomials. In *Mem. Amer. Math. Soc. 319*, Providence, RI, 1985. AMS. 101

[7] V. Barthelmann, E. Novak, and K. Ritter. High dimensional polynomial interpolation on sparse grids. *Adv. Comput. Math.*, 12(4):273–288, 2000. Multivariate polynomial interpolation. 106

[8] R. A. Bartlett and L. T. Biegler. rSQP++: An object-oriented framework for successive quadratic programming. Abstract for First Sandia Workshop on Large-scale PDE-constrained Optimization, Santa Fe, NM, April 4, 2001; to appear in Springer-Verlag Lecture Notes in Computational Science and Engineering. 262

[9] B. J. Bichon, M. S. Eldred, L. P. Swiler, S. Mahadevan, and J. M. McFarland. Efficient global reliability analysis for nonlinear implicit performance functions. *AIAA Journal*, 46(10):2459–2468, 2008. 95, 96

[10] B. Blinn. *Portable Shell Programming: An Extensive Collection of Bourne Shell Examples*. Prentice Hall PTR, New Jersey, 1996. 233, 291

[11] G. E. P. Box and D. R. Cox. An analysis of transformations. *J. Royal Stat. Soc., Series B*, 26:211–252, 1964. 91

[12] K. Breitung. Asymptotic approximation for multinormal integrals. *J. Eng. Mech., ASCE*, 110(3):357–366, 1984. 94

[13] R. H. Byrd, R. B. Schnabel, and G. A. Schultz. Parallel quasi-newton methods for unconstrained optimization. *Mathematical Programming*, 42:273–306, 1988. 249

[14] M. R. Celis, J. .E. Dennis, and R. .A Tapia. A trust region strategy for nonlinear equality constrained optimization. In P. .T. Boggs, R. H. Byrd, and R. B. Schnabel, editors, *Numerical Optimization 1984*, pages 71–82. SIAM, Philadelphia, USA, 1985. 137

[15] K. J. Chang, R. T. Haftka, G. L. Giles, and P.-J. Kao. Sensitivity-based scaling for approximating structural response. *J. Aircraft*, 30:283–288, 1993. 160

[16] X. Chen and N.C. Lind. Fast probability integration by three-parameter normal tail approximation. *Struct. Saf.*, 1:269–276, 1983. 91

[17] C. A. Coello, D. A. Van Veldhuizen, and G. B. Lamont. *Evolutionary Algorithms for Solving Multi-Objective Problems*. Kluwer Academic/Plenum Publishers, New York, 2002. 43, 313

[18] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust-Region Methods*. MPS-SIAM Series on Optimization, SIAM-MPS, Philadelphia, 2000. 139

[19] N. Cressie. *Statistics of Spatial Data*. John Wiley and Sons, New York, 1991. 164

[20] J. E. Dennis, D. M. Gay, and R. E. Welsch. ALGORITHM 573: NL2SOL–an adaptive nonlinear least-squares algorithm. *ACM Trans. Math. Software*, 7:369–383, 1981. 36, 56, 130

[21] J. E. Dennis and R. M. Lewis. Problem formulations and other optimization issues in multidisciplinary optimization. In *Proc. AIAA Symposium on Fluid Dynamics*, number AIAA-94-2196, Colorado Springs, Colordao, June 1994. 248

[22] J. E. Dennis and V. J. Torczon. Derivative-free pattern search methods for multidisciplinary design problems. In *Proc. 5th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, number AIAA-94-4349, pages 922–932, Panama City, FL, September 7–9 1994. 120

[23] A. Der Kiureghian and P. L. Liu. Structural reliability under incomplete information. *J. Eng. Mech., ASCE*, 112(EM-1):85–104, 1986. 89, 91, 92

[24] A. Dey and S. Mahadevan. Ductile structural system reliability analysis using adaptive importance sampling. *Structural Safety*, 20:137–154, 1998. 97

[25] Q. Du, V. Faber, and M. Gunzburger. Centroidal voronoi tessellations: Applications and algorithms. *SIAM Review*, 41:637–676, 1999. 77

[26] X. Du and W. Chen. Sequential optimization and reliability assessment method for efficient probabilistic design. *J. Mech. Design*, 126:225–233, 2004. 176, 179

[27] J. Eckstein, W. E. Hart, and C. A. Phillips. Resource management in a parallel mixed integer programming package. In *Proc. 1997 Intel Supercomputer Users Group Conference*, Albuquerque, NM, June 11–13 1997. `http://www.cs.sandia.gov/ISUG97/program.html`. 155

[28] J. Eckstein, W. E. Hart, and C. A. Phillips. PICO: An object-oriented framework for parallel branch and bound. In D. Butnariu, Y. Censor, and S. Reich, editors, *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*. Elsevier Science Publishers, Amsterdam, Netherlands, 2001. 55, 155

[29] M. S. Eldred. Optimization strategies for complex engineering applications. Technical Report SAND98-0340, Sandia National Laboratories, Albuquerque, NM, 1998. 13, 195

[30] M. S. Eldred, B. M. Adams, K. Haskell, W. J. Bohnhoff, J. P. Eddy, D. M. Gay, W. E. Hart, P. D. Hough, T. G. Kolda, L. P. Swiler, and J.-P. Watson. DAKOTA, a multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis: Version 4.2 reference manual. Technical Report SAND2006-4055, Sandia National Laboratories, Albuquerque, NM, Updated November 2008. Available online from `http://www.cs.sandia.gov/dakota/documentation.html`. 26, 37, 40, 43, 49, 53, 82, 116, 117, 118, 119, 120, 121, 130, 131, 137, 149, 150, 153, 160, 168, 171, 181, 183, 185, 190, 191, 193, 209, 213, 216, 219, 230, 267, 268, 285, 292

[31] M. S. Eldred, B. M. Adams, K. Haskell, W. J. Bohnhoff, J. P. Eddy, D. M. Gay, W. E. Hart, P. D. Hough, T. G. Kolda, L. P. Swiler, and J.-P. Watson. DAKOTA, a multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis: Version 4.2 developers manual. Technical Report SAND2006-4056, Sandia National Laboratories, Albuquerque, NM, Updated November 2008. Available online from `http://www.cs.sandia.gov/dakota/documentation.html`. 60, 159, 160, 162, 194, 241, 242

[32] M. S. Eldred, H. Agarwal, V. M. Perez, Jr. Wojtkiewicz, S. F., and J. E. Renaud. Investigation of reliability method formulations in dakota/uq. *Structure & Infrastructure Engineering: Maintenance, Management, Life-Cycle Design & Performance*, 3(3):199–213, 2007. 81, 92, 169, 176, 179, 310, 312

[33] M. S. Eldred and B. J. Bichon. Second-order reliability formulations in DAKOTA/UQ. In *Proceedings of the 47th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, number AIAA-2006-1828, Newport, RI, May 1–4 2006. 92, 169, 176, 179, 310, 312

[34] M. S. Eldred and D. M. Dunlavy. Formulations for surrogate-based optimization with data fit, multifidelity, and reduced-order models. In *Proceedings of the 11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, number AIAA-2006-7117, Portsmouth, VA, September 6–8 2006. 179

[35] M. S. Eldred, A. A. Giunta, and S. S. Collis. Second-order corrections for surrogate-based optimization with model hierarchies. In *Proceedings of the 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Albany, NY,, Aug. 30–Sept. 1, 2004. AIAA Paper 2004-4457. 144, 160, 161, 180

[36] M. S. Eldred, A. A. Giunta, S. F. Wojtkiewicz Jr., and T. G. Trucano. Formulations for surrogate-based optimization under uncertainty. In *Proc. 9th AIAA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, number AIAA-2002-5585, Atlanta, GA, September 4–6, 2002. 169, 174, 312

[37] M. S. Eldred and W. E. Hart. Design and implementation of multilevel parallel optimization on the Intel TeraFLOPS. In *Proc. 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, number AIAA-98-4707, pages 44–54, St. Louis, MO, September 2–4 1998. 247, 248

[38] M. S. Eldred, W. E. Hart, W. J. Bohnhoff, V. J. Romero, S. A. Hutchinson, and A. G. Salinger. Utilizing object-oriented design to build advanced optimization strategies with generic implementation. In *Proc. 6th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, number AIAA-96-4164, pages 1568–1582, Bellevue, WA, September 4–6 1996. 195, 248

[39] M. S. Eldred, W. E. Hart, B. D. Schimel, and B. G. van Bloemen Waanders. Multilevel parallelism for optimization on MP computers: Theory and experiment. In *Proc. 8th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, number AIAA-2000-4818, Long Beach, CA, 2000. 259, 262, 264, 266

[40] M. S. Eldred, D. E. Outka, W. J. Bohnhoff, W. R. Witkowski, V. J. Romero, E. R. Ponslet, and K. S. Chen. Optimization of complex mechanics simulations with object-oriented software design. *Computer Modeling and Simulation in Engineering*, 1(3), August 1996. 195

[41] M. S. Eldred and B. D. Schimel. Extended parallelism models for optimization on massively parallel computers. In *Proc. 3rd World Congress of Structural and Multidisciplinary Optimization (WCSMO-3)*, number 16-POM-2, Amherst, NY, May 17–21 1999. 155

[42] M. S. Eldred, C. G. Webster, and P. Constantine. Evaluation of non-intrusive approaches for wiener-askey generalized polynomial chaos. In *abstract submitted for the 10th AIAA Non-Deterministic Approaches Conference*, Chicago, IL, April 2008. 169

[43] G. M. Fadel, M. F. Riley, and J.-F. M. Barthelemy. Two point exponential approximation method for structural optimization. *Structural Optimization*, 2(2):117–124, 1990. 58, 93, 162

[44] D. Flaggs. JPrePost user's manual. In preparation. 236

[45] R. Fletcher, S. Leyffer, and P. L. Toint. On the global convergence of a filter-SQP algorithm. *SIAM J. Optim.*, 13(1):44–59, 2002. 138

[46] R. Fourer, D. M. Gay, and B. W. Kernighan. *AMPL: A Modeling Language for Mathematical Programming, 2nd ed.* Duxbury Press/Brooks/Cole Publishing Co., Pacific Grove, CA, 2003. 191

[47] P. Frauenfelder, C. Schwab, and R. A. Todor. Finite elements for elliptic problems with stochastic coefficients. *Comput. Methods Appl. Mech. Engrg.*, 194(2-5):205–228, 2005. 106

[48] J. H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1):1–141, March 1991. 58, 166

[49] K. Fukuda. cddlib reference manual. Web site, 2005. `ftp://ftp.ifor.math.ethz.ch/pub/fukuda/cdd/cddlibman/cddlibman.html`. 116

[50] J. Gablonsky. Direct version 2.0 userguide technical report. Technical Report CRSC-TR01-08, North Carolina State University, Center for Research in Scientific Computation, Raleigh, NC, 2001. 55, 119

[51] D. M. Gay. Hooking your solver to AMPL. Technical Report Technical Report 97-4-06, Bell Laboratories, Murray Hill, NJ, 1997. Available online as `http://www.ampl.com/REFS/HOOKING/index.html` and `http://www.ampl.com/REFS/hooking2.pdf`. 191

[52] T. Gerstner and M. Griebel. Numerical integration using sparse grids. *Numer. Algorithms*, 18(3-4):209–232, 1998. 105, 106

[53] R. Ghanem and J. R. Red-Horse. Propagation of probabilistic uncertainty in complex physical systems using a stochastic finite element technique. *Physica D*, 133:137–144, 1999. 53, 81, 101

[54] R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer-Verlag, New York, 1991. 53, 81, 101

[55] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright. User's guide for NPSOL (Version 4.0): A Fortran package for nonlinear programming. Technical Report TR SOL-86-2, System Optimization Laboratory, Stanford University, Stanford, CA, 1986. 55, 120

[56] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press, San Diego, CA, 1981. 15, 24, 129, 136, 297

[57] Philip E. Gill, Walter Murray, and Michael A. Saunders. SNOPT: An SQP algorithm for large-scale constrained optimization. *SIAM Review*, 47(1):99–131, 2005. 118

[58] D. Gilly. *UNIX in a Nutshell*. O'Reilly and Associates, Inc., Sebastopol, CA, 1992. 228

[59] A. A. Giunta. Use of data sampling, surrogate models, and numerical optimization in engineering design. In *Proc. 40th AIAA Aerospace Science Meeting and Exhibit*, number AIAA-2002-0538, Reno, NV, January 2002. 142

[60] A. A. Giunta and M. S. Eldred. Implementation of a trust region model management strategy in the DAKOTA optimization toolkit. In *Proc. 8th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, number AIAA-2000-4935, Long Beach, CA, September 6–8, 2000. 142, 153, 176

[61] A. A. Giunta, J. M. McFarland, L. P. Swiler, and M. S. Eldred. The promise and peril of uncertainty quantification using respone surface approximations. *Structure and Infrastructure Engineering*, 2(3-4):175–189, September-December 2006. 181

[62] A. A. Giunta and L. T. Watson. A comparison of approximation modeling techniques: Polynomial versus interpolating models. In *Proc. 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, number AIAA-98-4758, pages 392–404, St. Louis, MO, 1998. 58, 164

[63] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wessley Publishing Co., Inc., Reading, MA, 1989. 40

[64] G. A. Gray and T. G. Kolda. Algorithm 856: APPSPACK 4.0: Asynchronous parallel pattern search for derivative-free optimization. *ACM Transactions on Mathematical Software*, 32(3):485–507, September 2006. 54, 116

[65] J. D Griffin and T. G. Kolda. Nonlinearly-constrained optimization using asynchronous parallel generating set search. Technical Report SAND2007-3257, Sandia National Laboratories, Livermore, CA, 2007. 54, 116

[66] J. D. Griffin, T. G. Kolda, and R. M. Lewis. Asynchronous parallel generating set search for linearly-constrained optimization. *SIAM Journal on Scientific Computing*, 30(4):1892–1924, May 2008. 54, 116

[67] W. Gropp and E. Lusk. User's guide for mpich, a portable implementation of MPI. Technical Report ANL/MCS-TM-ANL-96/6, Argonne National Laboratory, Mathematics and Computer Science Division, 1996. 266

[68] W. Gropp, E. Lusk, and A. Skjellum. *Using MPI, Portable Parallel Programing with the Message-Passing Interface*. The MIT Press, Cambridge, MA, 1994. 256

[69] R. T. Haftka and Z. Gurdal. *Elements of Structural Optimization*. Kluwer, Boston, 1992. 15, 40

[70] A. Haldar and S. Mahadevan. *Probability, Reliability, and Statistical Methods in Engineering Design*. Wiley, New York, 2000. 15, 53, 90, 91, 185, 307, 308

[71] W. E. Hart. The coliny project. Web site, 2007. `http://software.sandia.gov/Acro/Coliny/`. 37, 40, 116

[72] A. S. Hedayat, N. J. A. Sloane, and J. Stufken. *Orthogonal Arrays: Theory and Applications*. Springer-Verlag, New York, 1999. 74

[73] J. C. Helton and F. J. Davis. Sampling-based methods for uncertainty and sensitivity analysis. Technical Report SAND99-2240, Sandia National Laboratories, Albuquerque, NM, 2000. 82

[74] J. C. Helton, J. D. Johnson, W. L. Oberkampf, and C. B Storlie. A sampling-based computational strategy for the representation of epistemic uncertainty in model predictions with evidence theory. *Comp. Methods in Applied Mechanics and Engineering*, 196:3980–3998, 2007. 111

[75] M. Hohenbichler and R. Rackwitz. Sensitivity and importance measures in structural reliability. *Civil Eng. Syst.*, 3:203–209, 1986. 175

[76] M. Hohenbichler and R. Rackwitz. Improvement of second-order reliability estimates by importance sampling. *J. Eng. Mech., ASCE*, 114(12):2195–2199, 1988. 94, 95

[77] H.P. Hong. Simple approximations for improving second-order reliability estimates. *J. Eng. Mech., ASCE*, 125(5):592–595, 1999. 94, 95, 308

[78] S. Hosder, R. W. Walters, and M. Balch. Efficient sampling for non-intrusive polynomial chaos applications with multiple uncertain input variables. In *Proceedings of the 48th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, number AIAA-2007-1939, Honolulu, HI, April 23–26, 2007. 106, 107

[79] D. Huang, T. T. Allen, W. I. Notz, and N. Zeng. Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of Global Optimization*, 34:441–466, 2006. 95, 146

[80] R. L. Iman and M. J Shortencarier. A Fortran 77 program and user's guide for the generation of latin hypercube samples for use with computer models. Technical Report NUREG/CR-3624, SAND83-2365, Sandia National Laboratories, Albuquerque, NM, 1984. 53, 82, 162

[81] D. Jones, M. Schonlau, and W. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 1998. 57, 95, 146

[82] A. Karamchandani and C. A. Cornell. Sensitivity estimation within first and second order reliability methods. *Struct. Saf.*, 11:95–107, 1992. 175

[83] B. W. Kernighan and D. M. Ritchie. *The C Programming Language*. Prentice Hall PTR, Englewood Cliffs, NJ, 2nd edition, 1988. 195, 198, 203, 253

[84] J. R. Koehler and A. B. Owen. Computer experiments. In S. Ghosh and C. R. Rao, editors, *Handbook of Statistics*, volume 13. Elsevier Science, New York, 1996. 73, 164

[85] N. Kuschel and R. Rackwitz. Two basic problems in reliability-based structural optimization. *Math. Method Oper. Res.*, 46:309–333, 1997. 308, 312

[86] L. D. Lathauwer, B. D. Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000. 144

[87] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Prentice–Hall, 1974. 139

[88] R. M. Lewis and S. N. Nash. A multigrid approach to the optimization of systems governed by differential equations. Technical Report AIAA-2000-4890, AIAA, 2000. 160

[89] A. Martelli. *Python in a Nutshell*. O'Reilly and Associates, Cambridge, MA, 2003. 233

[90] J. M. McFarland. *Uncertainty Analysis for Computer Simulations through Validation and Calibration*. PhD thesis, Vanderbilt University, Nashville, Tennesssee, 2008. available for download at http://etd.library.vanderbilt.edu/ETD-db/available/etd-03282008-125137/. 165

[91] M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979. 82

[92] J. C. Meza, R. A. Oliva, P. D. Hough, and P. J. Williams. OPT++: an object oriented toolkit for nonlinear optimization. *ACM Transactions on Mathematical Software*, 33(2), 2007. 55, 120, 130

[93] M. D. Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991. 77, 316

[94] R. H. Myers and D. C. Montgomery. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. John Wiley & Sons, Inc., New York, 1995. 79, 164

[95] A. Nealen. A short-as-possible introduction to the least squares, weighted least squares, and moving least squares methods for scattered data approximation and interpolation. Technical report, Discrete Geometric Modeling Group, Technishe Universitaet. 167

[96] F. Nobile, R. Tempone, and C. G. Webster. A sparse grid stochastic collocation method for partial differential equations with random input data. Technical Report Technical report TRITA-NA 2007:7, Royal Institute of Technology, Stockholm, Sweden, 2007. 105, 106

[97] F. Nobile, R. Tempone, and C. G. Webster. An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM J. on Num. Anal.*, 2008. To appear. 105, 106

[98] J. Nocedal and Wright S. J. *Numerical Optimization*. Springer Series in Operations Research. Springer, New York, 1999. 15, 136, 297

[99] W. .L. Oberkampf and J. C. Helton. Evidence theory for engineering applications. Technical Report SAND2003-3559P, Sandia National Laboratories, Albuquerque, NM, 2003. 111

[100] E. O. Omojokun. *Trust Region Algorithms for Optimization with Nonlinear Equality and Inequality Constraints*. PhD thesis, University of Colorado, Boulder, Colorado, 1989. 137

[101] M. J. L. Orr. Introduction to radial basis function networks. Technical report, University of Edinburgh, Edinburgh, Scotland, 1996. 167

[102] A. B. Owen. Orthogonal arrays for computer experiments, integration, and visualization. *Statistica Sinica*, 2(2):439–452, 1992. 76

[103] V. M. Pérez, M. S. Eldred, , and J. E. Renaud. Solving the infeasible trust-region problem using approximations. In *Proceedings of the 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Albany, NY, Aug. 30–Sept. 1, 2004. AIAA Paper 2004-4312. 137, 140

[104] V. M. Pérez, J. E. Renaud, and L. T. Watson. An interior-point sequetial approximation optimization methodology. *Structural and Multidisciplinary Optimization*, 27(5):360–370, July 2004. 136, 137, 140

[105] C. D. Perttunen, D. R. Jones, and B. E. Stuckman. Lipschitzian optimization without the Lipschitz constant. *J. Optimization Theory and Application*, 79(1):157–181, 1993. 116

[106] E. R. Ponslet and M. S. Eldred. Discrete optimization of isolator locations for vibration isolation systems: an analytical and experimental investigation. In *Proc. 6th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, number AIAA-96-4178, pages 1703–1716, Bellevue, WA, September 4–6 1996. Also appears as Sandia Technical Report SAND96-1169, May 1996. 117

[107] R. Rackwitz. Optimization and risk acceptability based on the Life Quality Index. *Struct. Saf*, 24:297–331, 2002. 175

[108] R. Rackwitz and B. Fiessler. Structural reliability under combined random load sequences. *Comput. Struct.*, 9:489–494, 1978. 91

[109] M. T. Reagan, H. N. Najm, P. P. Pebay, O. M. Knio, and R. G. Ghanem. Quantifying uncertainty in chemical systems modeling. *Int. J. Chem. Kinet.*, 37:368–382, 2005. 108

[110] T. D. Robinson, M. S. Eldred, K. E. Willcox, and R. Haimes. Strategies for multifidelity optimization with variable dimensional hierarchical models. In *Proceedings of the 47th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference (2nd AIAA Multidisciplinary Design Optimization Specialist Conference)*, Newport, RI, May 1–4, 2006. AIAA Paper 2006-1819. 144

[111] T. D. Robinson, K. E. Willcox, M. S. Eldred, and R. Haimes. Multifidelity optimization for variable-complexity design. In *Proceedings of the 11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Portsmouth, VA, September 6–8, 2006. AIAA Paper 2006-7114. 144

[112] J. F. Rodriguez, J. E. Renaud, and L. T. Watson. Convergence of trust region augmented lagrangian methods using variable fidelity approximation data. *Structural Optimization*, 15:1–7, 1998. 136

[113] M. Rosenblatt. Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23(3):470–472, 1952. 89, 91

[114] H. H. Rosenbrock. An automatic method for finding the greatest or least value of a function. *Computer Journal*, 3:175–184, 1960. 24

[115] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto. *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. John Wiley & Sons, 2004. 78, 79, 318

[116] K. Schittkowski. *More Test Examples for Nonlinear Programming, Lecture Notes in Economics and Mathematical Systems, Vol. 282*. Springer-Verlag, Berlin, 1987. 297

[117] K. Schittkowski. NLPQLP: A fortran implementation of a sequential quadratic programming algorithm with distributed and non-monotone line search – user's guide. Technical report, Department of Mathematics, University of Bayreuth, Bayreuth, Germany, 2004. 55, 119, 249

[118] P. R. Schunk, P. A. Sackinger, R. R. Rao, K. S. Chen, and R. A. Cairncross. GOMA – a full-newton finite element program for free and moving boundary problems with coupled fluid/solid momentum, energy, mass, and chemical species transport: User's guide. Technical Report SAND95-2937, Sandia National Laboratories, Albuquerque, NM, 1995. 293

[119] G. A. F. Seber and C. J. Wild. *Nonlinear Regression*. Wiley, New Jersey, 2003. 131

[120] G. D. Sjaardema. APREPRO: An algebraic preprocessor for parameterizing finite element analyses. Technical Report SAND92-2291, Sandia National Laboratories, Albuquerque, NM, 1992. 186, 188

[121] S.A. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. *Dokl. Akad. Nauk SSSR*, 4:240–243, 1963. 106

[122] M. Snir, S. Otto, S. Huss-Lederman, D. Walker, and J. Dongarra. *MPI: The Complete Reference*. MIT Press, Cambridge, MA, 1996. 256

[123] R. Sues, M. Aminpour, and Y. Shin. Reliability-based multidisciplinary optimization for aerospace systems. In *Proc. 42rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, number AIAA-2001-1521, Seattle, WA, April 16-19 2001. 310

[124] L. P Swiler, R. Slepoy, and A. A. Giunta. Evaluation of sampling methods in constructing response surface approximations. In *Proc. 47th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, number AIAA-2006-1827, Newport, RI, May 1-4 2006. 181

[125] L. P. Swiler and G. D. Wyss. A user's guide to Sandia's latin hypercube sampling software: LHS UNIX library and standalone version. Technical Report SAND04-2439, Sandia National Laboratories, Albuquerque, NM, July 2004. 71, 82, 185

[126] C. H. Tong. The psuade software library. Web site, 2005. `http://www.llnl.gov/CASC/uncertainty_quantification/#psuade`. 52, 71, 77

[127] C. H. Tong and J. C. Meza. DDACE: A distributed object-oriented software with multiple samplings for the design and analysis of computer experiments. Technical Report SAND##-XXXX, Sandia National Laboratories, Livermore, CA. Draft as yet unpublished, see also `http://csmr.ca.sandia.gov/projects/ddace/DDACEdoc/html/index.html`. 52, 71, 73, 162

[128] J. Tu, K. K. Choi, and Y. H. Park. A new study on reliability-based design optimization. *J. Mech. Design*, 121:557–564, 1999. 92

[129] L. Tvedt. Distribution of quadratic forms in normal space – applications to structural reliability. *J. Eng. Mech., ASCE*, 116(6):1183–1197, 1990. 308

[130] G. N. Vanderplaats. CONMIN – a FORTRAN program for constrained function minimization. Technical Report TM X-62282, NASA, 1973. See also Addendum to Technical Memorandum, 1978. 31, 54, 117

[131] G. N. Vanderplaats. *Numerical Optimization Techniques for Engineering Design: With Applications*. McGraw-Hill, New York, 1984. 15, 136, 139

[132] Vanderplaats Research and Development, Inc., Colorado Springs, CO. *DOT Users Manual, Version 4.20*, 1995. 54, 118

[133] K. W. Vugrin, L. P. Swiler, R. M. Roberts, N. J. Stuckey-Mack, and S. P Sullivan. Confidence region estimation techniques for nonlinear regression in groundwater flow: Three case studies. *Water Resources Research*, 43, 2007. 131

[134] L. Wall, T. Christiansen, and R. L. Schwartz. *Programming Perl*. O'Reilly & Associates, Cambridge, 2nd edition, 1996. 233

[135] R. W. Walters. Towards stochastic fluid mechanics via polynomial chaos. In *Proceedings of the 41st AIAA Aerospace Sciences Meeting and Exhibit*, number AIAA-2003-0413, Reno, NV, January 6–9, 2003. 106

[136] B. Walton. BPREPRO preprocessor documentation. Online document `http://bwalton.com/bprepro.html`. 236

[137] G. W. Wasilkowski and H. Woźniakowski. Explicit cost bounds of algorithms for multivariate tensor product problems. *Journal of Complexity*, 11:1–56, 1995. 106

[138] G. Weickum, M. S. Eldred, and K. Maute. Multi-point extended reduced order modeling for design optimization and uncertainty analysis. In *Proceedings of the 47th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference (2nd AIAA Multidisciplinary Design Optimization Specialist Conference)*, Newport, RI, May 1–4, 2006. AIAA Paper 2006-2145. 144

[139] N. Wiener. The homogoeneous chaos. *Amer. J. Math.*, 60:897–936, 1938. 101

[140] Y.-T. Wu, Y. Shin, R. Sues, and M. Cesare. Safety-factor based approach for probability-based design optimization. In *Proc. 42nd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, number AIAA-2001-1522, Seattle, WA, April 16–19 2001. 176, 179, 310

[141] Y.-T. Wu and P.H. Wirsching. A new algorithm for structural reliability estimation. *J. Eng. Mech., ASCE*, 113:1319–1336, 1987. 91

[142] B. A. Wujek and J. E. Renaud. New adaptive move-limit management strategy for approximate optimization, part 1. *AIAA Journal*, 36(10):1911–1921, 1998. 137

[143] B. A. Wujek and J. E. Renaud. New adaptive move-limit management strategy for approximate optimization, part 2. *AIAA Journal*, 36(10):1922–1934, 1998. 137

[144] D. Xiu and J.S. Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM J. Sci. Comput.*, 27(3):1118–1139 (electronic), 2005. 106

[145] D. Xiu and G. M. Karniadakis. The wiener-askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comput.*, 24(2):619–644, 2002. 101

[146] S. Xu and R. V. Grandhi. Effective two-point function approximation for design optimization. *AIAA J.*, 36(12):2269–2275, 1998. 58, 93, 162

[147] D. C. Zimmerman. Genetic algorithms for navigating expensive and complex design spaces, September 1996. Final Report for Sandia National Laboratories contract AO-7736 CA 02. 58, 165

[148] T. Zou, S. Mahadevan, and R. Rebba. Computational efficiency in reliability-based optimization. In *Proceedings of the 9th ASCE Specialty Conference on Probabilistic Mechanics and Structural Reliability*, Albuquerque, NM, July 26–28, 2004. 176