VALUE-BASED GLOBAL OPTIMIZATION

A Dissertation
Presented to
The Academic Faculty

by

Roxanne A. Moore

In Partial Fulfillment
Of the Requirements for the Degree of
Doctor of Philosophy in Mechanical Engineering

Georgia Institute of Technology

August 2012

VALUE-BASED GLOBAL OPTIMIZATION

Approved by:

Dr. Chris Paredis, Advisor
G. W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Dr. David Romero
Dept. of Mechanical & Industrial Engineering
University of Toronto

Dr. Bert Bras
G. W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Dr. C.F. Jeff Wu
School of Industrial & Systems Engineering
Georgia Institute of Technology

Dr. Michael Leamy
G. W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Date Approved: May 18, 2012

ACKNOWLEDGEMENTS

I am thankful for the many people in my life who have supported me in so many ways during the journey of writing this thesis.

Thank you to my advisor, Chris, for guiding me through the research process for the last five years. You have helped me to refine my technical skills, my coding skills, my analytical skills, and my communication skills. Thank you for not giving up on me during the rough patches. Making it to the end was truly a team effort.

Thank you to committee member Dr. David Romero for at times acting almost as a co-advisor in this work. Your helpful feedback and discussions truly helped this thesis come together.

Thank my other committee members: Drs. Michael Leamy, Jeff Wu, and Bert Bras. Your feedback and encouragement has been most helpful.

Thank you to the NSF Center for Compact and Efficient Fluid Power for funding this work.

I am grateful for the support of my fellow lab members through the years. Thank you to Stephanie, Ben, Aditya, Kevin, Rich, Alek, Sebastian, and Jackie. Thank you for your technical help, your emotional support, and the many laughs. And thanks, of course, for assisting me with my leftover cake.

I would also like to express my thanks to the undergrads I worked with during my years at Georgia Tech, especially Isabelle and Jayme, who worked on the hydraulic hybrid model used in this thesis. I enjoyed working with you!

I am thankful to have had the support and friendship of the Woodruff School Graduate Women. I cannot express how vital this group has been to my success at Georgia Tech. I have made many friends and swapped many stories. The opportunities for professional development and networking in a small group are unparalleled. I will miss you ladies!

My parents have played a vital role in my success for my entire life. They are supportive in any possible way and do everything they can to ensure that I have the resources I need to succeed. They sent me care packages all through grad school and even flew out to see my defense. I could not ask for better, more supportive parents!

Thank you to Craig for your unwavering support over the last 6 months. I know it has not always been easy, but I'm sure I could not have done it without your confidence in my abilities.

Finally, thank you to all the friends from all stages of my life who have supported me in so many ways on this journey. While I'm separated geographically from many of you, I am so grateful to have you in my life.

Words are not enough to express my gratitude.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Computational models and simulations are essential system design tools that allow for improved decision making and cost reductions during all phases of the design process. However, the most accurate models are often computationally expensive and can therefore only be used sporadically. Consequently, designers are often forced to choose between exploring many design alternatives with less accurate, inexpensive models and evaluating fewer alternatives with the most accurate models. To achieve both broad exploration of the alternatives and accurate determination of the best alternative with reasonable costs incurred, surrogate modeling and variable accuracy modeling are used widely. A surrogate model is a mathematically tractable approximation of a more expensive model based on a limited sampling of that model, while variable accuracy modeling involves a collection of different models of the same system with different accuracies and computational costs. As compared to using only very accurate and expensive models, designers can determine the best solutions more efficiently using surrogate and variable accuracy models because obviously poor solutions can be eliminated inexpensively using only the less expensive, less accurate models. The most accurate models are then reserved for discerning the best solution from the set of good solutions.

In this thesis, a Value-Based Global Optimization (VGO) algorithm is introduced. The algorithm uses kriging-like surrogate models and a sequential sampling strategy based on Value of Information (VoI) to optimize an objective characterized by multiple analysis models with different accuracies. It builds on two primary research contributions. The

first is a novel surrogate modeling method that accommodates data from any number of analysis models with different accuracies and costs. The second contribution is the use of Value of Information (VoI) as a new metric for guiding the sequential sampling process for global optimization. In this manner, the cost of further analysis is explicitly taken into account during the optimization process.

Results characterizing the algorithm show that VGO outperforms Efficient Global Optimization (EGO), a similar global optimization algorithm that is considered to be the current state of the art. It is shown that when cost is taken into account in the final utility, VGO achieves a higher utility than EGO with statistical significance. In further experiments, it is shown that VGO can be successfully applied to higher dimensional problems as well as practical engineering design examples.

CHAPTER 1: MOTIVATION FOR VALUE-BASED GLOBAL OPTIMIZATION

Systems design problems are often complex, involving many interactions between multiple subsystems [25, 56]. For example, if a designer or team of designers is attempting to design an airplane, hundreds of thousands of individual components must work together. On the engineering side of the spectrum, the mechanical systems, electrical systems, and control systems must all be seamlessly integrated. Other stakeholders are also involved; for example, how many passengers can the airplane seat and what will be their comfort level? Many variables come into play: the size of the fuselage, the size and placement of the seats, the size and placement of the engine, the engine acoustics, the in-flight amenities, etc. If one were to design an 'optimal' aircraft with respect to some high level objective, the problem would literally have hundreds of thousands of *design variables.*

In this dissertation, the *design space* is defined as the range and associated units of inputs or *design variables* from which a final *design artifact* is selected. The design space for systems engineering problems is potentially very large at all stages of the design process. From initial concept exploration to final sizing choices, there are many possible variables and combinations to consider. Naturally, it is not feasible from a time or cost perspective to prototype many potential design artifacts. Fortunately, modeling, simulation, and optimization can greatly assist with assessing the viability and performance of potential *design alternatives.* These tools have become vital to engineering design and decision making.

Models, while never perfectly accurate, are often the only option for design space exploration and informed decision making because they are significantly cheaper than physical experiments and prototypes. However, in systems engineering, the design space is often so big that even simulating so many possible alternatives not possible. This is particularly true for very *accurate* models-- while faster processors and more sophisticated simulation environments allow for very sophisticated modeling, added accuracy comes at a cost. For example, it is not unusual for a Finite Element Simulation to take several days. Even if a simulation only takes a few minutes, it may require tens of thousands of evaluations during an optimization or uncertainty analysis. Therefore, it is not necessarily pragmatic or even possible to explore an entire design space at a high level of detail due to time and computational costs incurred.

To counter the problem of computational cost, methods have been proposed for reducing the number of design variables by using screening methods [61] or by performing a sensitivity analysis [9]. These methods allow a designer to identify the design variables or *uncertain variables*, variables that are not directly controlled by the designer, that have the greatest effects on the overall system performance. Then, analyses can be run using only the most significant variables and neglecting those that have only marginal effects on system performance.

Alternatively, methods have been developed for approximating an otherwise computationally intractable functional relationship by using a simplified or *surrogate model* [9, 13, 30, 31, 46, 54, 55, 60, 61, 67]. For example, a complex simulation model might be approximated by a high-order polynomial surface, reducing the simulation time from over a minute to less than one second. By reducing the number of design variables

or approximating expensive functions with cheaper surrogates, evaluation of many more design alternatives can be performed at a more reasonable cost as compared to the original, high-dimensional analysis.

Design problems are frequently framed as optimization problems: the designer is seeking the best possible *design artifact* within the confines of the *design space* to meet his or her needs. For example, a designer may seek the design that will produce the most profit or achieve the highest performance specifications. In engineering applications, however, simply finding the mathematical optimum of a design optimization problem, if even possible, is not the primary objective. Even if sufficient computing resources were available, it would generally be impossible to prove global optimality analytically for complex, black-box simulation analysis models. Therefore, what is actually desired is not so much an optimal solution, but a sufficiently good solution that can be achieved at a reasonable design process cost. For example, an electronic widget that is 'optimally' designed with respect to performance may have no real value if it takes two years to find this particular design artifact and implement it. By then a competitor may have already produced another widget, or consumer demands may have shifted. Thus, the cost of the design process must be considered because it may greatly affect the utility of the final widget.

To that end, the Value-based Global Optimization (VGO) algorithm presented in this paper relies on the use of a utility function [65]. Utility functions are a mathematical means for comparing the relative profitability or goodness of a particular design artifact. However, many utility functions neglect the costs incurred during the design process itself. In the VGO approach, the costs of the process and analyses are modeled explicitly

and included in the utility function [63]. This provides a mathematically sound mechanism for considering the overall utility of the artifact including the cost of the design process.

In this chapter, the motivation and critical research issues associated with this algorithm are described. In Section 1.1, common trade-offs made in the systems design area are described. These trade-offs illustrate a need for variable accuracy modeling, which is addressed in Section 1.2. Variable accuracy modeling is a key contribution of the VGO algorithm. The critical issues that must be addressed in the VGO algorithm development are described in Section 1.3. The research questions and hypotheses, stemming from the critical issues, are then addressed in Section 1.4. Finally, this chapter concludes with a description of the remained of the thesis in Section 1.5.

1.1 THE DESIGNER'S DILEMMA

As discussed previously, modeling, simulation, and optimization have become increasingly important to the success of design and decision making endeavors in a variety of disciplines. This need for modeling and simulation is a product of the following problem characteristics: 1) large, high-dimensional design spaces 2) system complexity, and 3) the cost of prototypes and physical experimentation. Although no computational model can ever perfectly emulate a physical system, performing simulations is essential for making informed design decisions in the absence of physical prototypes. As modeling and simulation packages grow increasingly sophisticated, the error between model predictions and their physical counterparts has decreased. However, accuracy comes at a cost with respect to computation time. Thus, given limited

computing resources and limited time, designers are often confronted with the difficult choice between the following two extremes as is depicted in Figure 1.1:

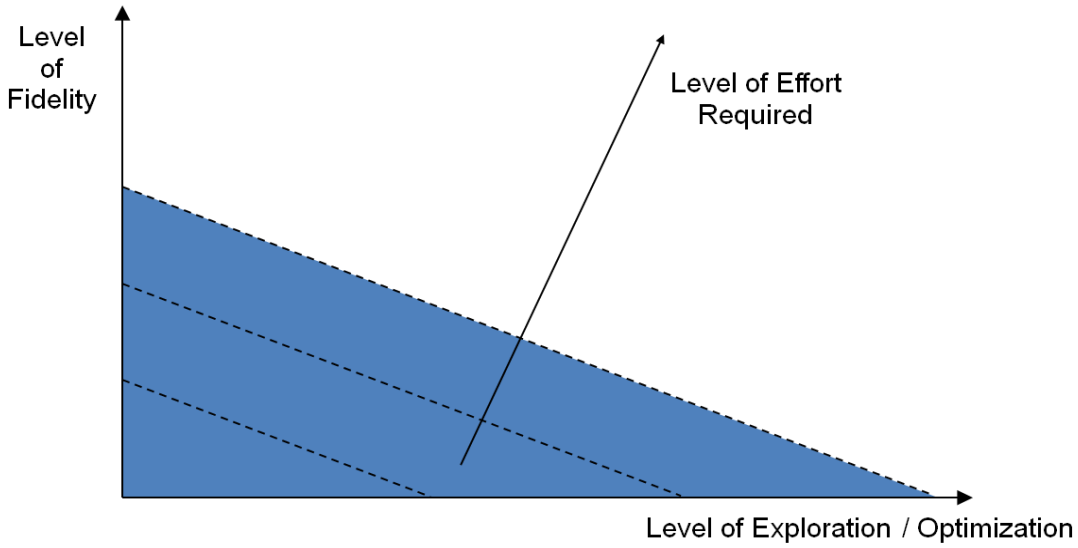

FIGURE 1.1: DESIGNER'S DILEMMA: LEVEL OF FIDELITY VERSUS LEVEL O F EXPLORATION

1) *Explore many design alternatives with an inexpensive, low-fidelity model*. This is known as broad exploration, or *global search*. With global search, a designer is unlikely to miss promising regions of the design space altogether. The drawback, is that models that are inexpensive enough to allow global, exhaustive search are unlikely to be accurate enough to allow accurate determination of the best solution from among the good solutions.

2) *Explore a smaller number of design alternatives with a very accurate but expensive high-fidelity model.* This is known as *local refinement.* Often, an accurate, expensive model is used. In this case, the best alternative is likely to be identified if it is among the small number of design alternatives considered, but there is no guarantee that an even better solution does not exist in the unexplored design space.

Given these two extremes, the logical question to ask is: does there exist a way to trade off inexpensive, broad exploration with high accuracy refinement in a way that does not compromise the quality of the solution? To achieve this, innovative use of models is required. Conceptually, this is illustrated in Figure 1.2. It can be seen that very little quality and very little exploration is sacrificed at the optimal point (designated by the star) but the level of effort required is reasonable. The question is: how is this compromise reached in practice?

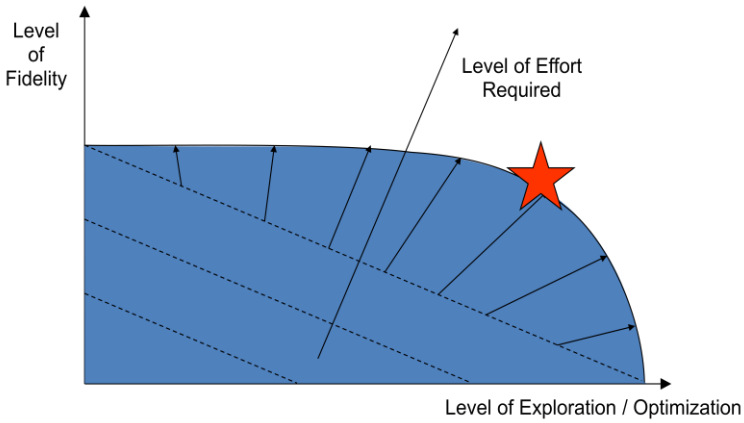

FIGURE 1.2:A DESIRABLE COMPROMISE BETWEEN EXPLORATION AND ACCURACY

Achieving this compromise is not feasible using a single analysis model with a particular cost and level of accuracy—we would again be forced to choose between broad exploration and local refinement, or between reasonable cost and high accuracy. Instead, we must use multiple models at different accuracies and costs to achieve this compromise.

1.2 CONCEPTUAL APPROACH: MODELS AT DIFFERENT ACCURACIES

One approach to reducing the cost of global search without sacrificing accuracy in the neighborhood of the optimum is to use multiple analysis models of differing levels of

accuracy [3, 15, 29, 38, 48, 59]. At this point, it is necessary to discuss what is truly meant by *accuracy.* Often in the literature the word *fidelity* is used interchangeably with *accuracy*; however, in this dissertation, the terms fidelity and accuracy are used with the following meanings.

Fidelity refers to the degree to which a model reflects the behavior of a real system being modeled [24]. It is a property of a *model*. One can state that model A has higher fidelity than model B if model A includes additional phenomena beyond all the ones included B. Note that this comparison between models A and B is a partial ordering; it is possible for A to include phenomena not included in B and vice versa. The term 'level of fidelity' must thus be used with caution because it is not a metric that can provide a full ordering of all models for a particular system.

Accuracy is different from fidelity in that it applies only to *simulations* (i.e., experiments performed on models [8]). It characterizes the degree of closeness of a prediction to its actual, true value. Only in the context of a specific simulation can one assess accuracy. Depending on the context of the experiment, a model of a particular fidelity can produce very different levels of accuracy.

When considering how accuracy is achieved in practice, it is also useful to consider the definitions for *resolution* and *abstraction*. *Resolution* is a special type of fidelity characterization that refers specifically to the level of discretization of a model or simulation in either space or time. For instance, a finite element model has a higher resolution if the mesh is denser, meaning that the discretization intervals are smaller.

Similarly, a dynamic simulation has a higher resolution of the differential equations are solved using a smaller time step.

Finally, *abstraction* or *level of abstraction* refers to the level of information content of a model. As is true for fidelity, it is a property of a model rather than of a simulation. Through a process of abstraction (or generalization), certain system properties are removed from a model so that one can no longer obtain information about these properties in an experiment or simulation [18].

Based on these definitions, the term variable fidelity modeling, though widely used in the literature [3, 4, 15, 38, 45, 49, 68], is somewhat of a misnomer for two reasons. First, in the context of design optimization one is interested primarily in the *accuracy* of a model prediction, not its fidelity. Even though varying the level of fidelity is one way to influence the accuracy of a prediction, the level of fidelity does not directly characterize the accuracy. Therefore, in this dissertation the term *accuracy* is used most often to characterize model quality.

Secondly, the term *variable* or *varying* is also something of a misnomer when referring to models. Models do not change their accuracies dynamically, at least not in a way that is meaningful or controllable by designers. Models typically do not have a tuning knob to make them more or less accurate with ease, though that would be exceedingly useful. Given the current state of the art, the only way a model might be considered truly 'variable' with respect to accuracy is that simulations may be run at higher or lower resolutions to yield more or less accurate analyses. In practice, however, these are treated as different models. Models of 'varying levels of accuracy' or 'different levels of

accuracy' or 'variable fidelity' all refer to the same thing: a collection of models and their associated simulations all of which are characterized as having different levels of accuracy.

Using multiple models at differing accuracies enables computational resources to be used more effectively by relying on the most accurate — and most costly — analysis models only when we approach the optimum. This idea is illustrated in Figure 1.3; assuming a maximization problem, it is necessary to tighten the accuracy bounds as we get closer to the optimum. A very accurate assessment of a bad solution's inadequacy is a waste of computational resources. When a solution is not promising, what is needed is a model with just enough accuracy to enable us to identify the general direction in which better solutions can be found. However, an inaccurate model by itself cannot discern the best solutions from the set of good solutions, so in the neighborhood of the optimum we must rely on our most accurate models. In this fashion, combining models with multiple accuracies enables global exploration at reasonable cost while still ensuring high accuracy in the neighborhood of the optimum.

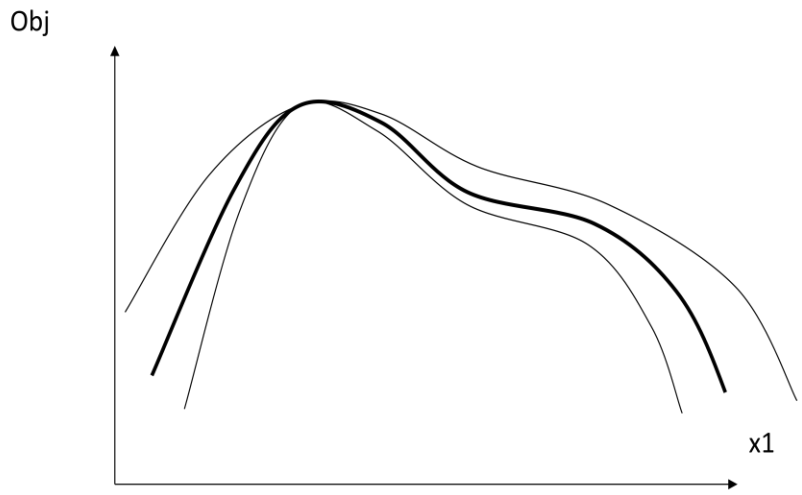

FIGURE 1.3: AN OBJECTIVE FUNCTION AND ITS DESIRED ACCURACY BOUNDS

1.3 CRITICAL ISSUES AND MODEL MANAGEMENT

In the previous section, it was illustrated that conceptually, models at different levels of accuracy can enable design space exploration in a cost efficient and effective manner. In this section, more practical considerations are addressed. Specifically, what algorithmic infrastructure is required for managing information obtained from each of the models?

Past approaches to variable accuracy modeling tend to be limited to only two models. Clearly, two models are superior to one, but what is more desirable is if any and all available models could be leveraged to aid in the design space exploration and optimization. In addition to this weakness, past approaches to variable accuracy modeling and optimization generally do not explicitly account for the cost of the analyses used during the optimization process [2, 4, 15, 38, 44, 52, 53, 68].

What is needed, therefore, is a global optimization algorithm which leverages simulation outcomes from any number of models at different accuracies while accounting for their associated costs explicitly. In this fashion, all of the relevant simulation data can be used

to make cost-effective decisions throughout the optimization process. In particular, the most accurate and costly models should only be used when it is valuable to do so, and less accurate models should be used to ensure sufficient global exploration at low cost. Thus, a method for using multi-accuracy data in a meaningful way is a critical component of a new global optimization algorithm, as well as quantifying value and selecting the most valuable analysis action at every step in the optimization process.

Thus, the critical issues that must be addressed in this thesis are as follows:

1. A method for combining multi-accuracy predictions from any number of models is needed so that better alternatives can be found at lower cost.

2. A mathematically sound method for trading off cost and solution quality in the optimization context must be developed to determine the most valuable analysis at every step in the optimization.

The critical issues identified must both be addressed in order to develop and characterize a meaningful algorithm for solving engineering design optimization problems in a cost effective manner. While each of these issues could be viewed as separate contributions, together they are much more powerful. The result will be an automated way to balance global search and local refinement during optimization and find good design artifacts based on whatever data and analyses are available. Another particularly useful outcome of explicitly considering costs is the ability to stop the optimization process when the point of diminishing returns is reached—that is, when the cost of further evaluation outweighs the potential benefit of further refinement. These issues map directly to the research questions presented in the next section.

1.4 RESEARCH QUESTIONS AND HYPOTHESES

The goal of this dissertation is to address the following research question:

Primary Research Question: *How can designers perform design optimizations at a reasonable cost without sacrificing solution quality?*

Hypothesis: *A Value-Based Global Optimization (VGO) algorithm will allow designers to achieve good solutions (design artifacts) at better costs than can be achieved with comparable existing algorithms.*

This primary hypothesis is that a new optimization algorithm (VGO) is needed in order to best leverage the available resources. By leveraging models of varying accuracies, accounting for costs explicitly, and adopting a value-driven strategy for selecting additional analyses during the optimization process, VGO will allow designers to achieve good solutions (design artifacts) at better costs than can be achieved with comparable existing algorithms.

The contributions of this algorithm must be two-fold: to allow for multiple models at multiple accuracies and costs, and to provide a meaningful method for *sequential sampling.* Sequential sampling will be discussed in greater detail in Section 2.3.2, but the basic premise is that new analyses are added dynamically during the optimization process. In this case, we want to add analyses only where it is most *valuable* to do so, where value can be formally calculated. Each contribution is addressed in a separate secondary research question.

Secondary Research Question 1: *How can data from multiple models of varying levels of accuracy be used advantageously during the design optimization process?*

In the context of optimization, merely running various simulations throughout the design space is not useful, unless a more 'brute force' method of dense sampling is being used. What is more efficient is making use of correlations in the data and fitting a surface to the data to gain further insights as to where the best solutions might be found. Recall that in Chapter 1, the idea of *surrogate modeling* was briefly introduced. This concept will be detailed in Section 2.2, and a detailed overview of current surrogate modeling approaches will be provided. For conceptual understanding at this stage, it is sufficient to understand that a surrogate model provides a mathematically tractable (cheap to evaluate) surface that is fit to the available simulation data. Fitting a surface to the data allows an optimizer to make use of valuable properties such as gradient information, and in some cases the uncertainty associated with the predicted surface.

In VGO, we seek to leverage both surrogate modeling capabilities and variable accuracy modeling to achieve low cost, high accuracy optimization. Many different approaches to both surrogate modeling and variable accuracy modeling have been described in the literature, sometimes even in combination. However, none provides a surrogate functional structure that allows for any number of models to be used without restriction as to where in the design space data can be added. Specifically, many have the restriction that the most accurate models can only be evaluated at *design sites* where the less accurate models have already been evaluated [29, 35].

A *design site* is a point in the *design space*, or a particular combination of input variables, where analyses have been performed. Each design site has an associated *observation* of the objective function from a simulation model, or an associated output that was determined from analysis of the design site using a particular model. A *design site* and its associated *observation* make up a *sample site.* These terms are captured pictorially in Figure 1.4.

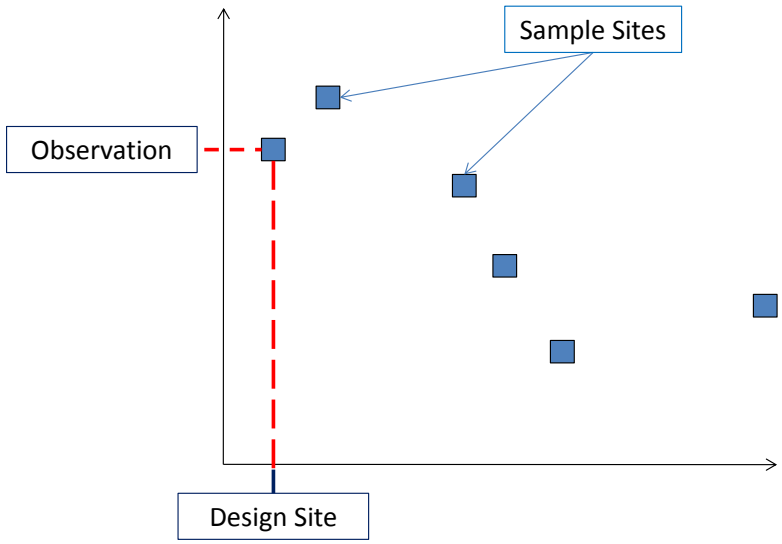

FIGURE 1.4: A DESIGN SPACE POPULATED BY SAMPLE SITES

The hypothesis for this question is a proposed method for creating surrogate models using data from any number of models at different accuracies with no sampling restrictions regarding where *design sites* from a particular model can be taken.

Secondary Hypothesis 1: *A Gaussian process-based surrogate model, similar to a kriging model, can be derived mathematically to accommodate multi-accuracy observations from any number of different models.*

To formulate a surrogate model based on multi-accuracy data, a Gaussian process modeling approach is used. This approach allows for the resulting surface to make use of statistical properties, such as correlation between the simulation observations. This Gaussian process-based surrogate model is fit to the design sites and their associated observations of the objective function in the design space. This model can then be used to make predictions about the objective function at any point in the design space, regardless of sampling density. The resulting functional structure is similar to that of a classic kriging model [55].

Kriging modeling, which has been used widely in the field of computational experiments, assumes zero uncertainty at all design sites and observations, resulting in an interpolator. Because computational experiments are generally deterministic, this is often a desirable property. However, the approach proposed in this thesis relaxes this assumption and allows for model uncertainty; this relaxation is necessary in order to accommodate data at different accuracies. In this thesis, it is assumed that the uncertainty in each model, or specifically, the amount by which the model differs from the true experimental value, can be adequately characterized by a zero mean Gaussian process with known variance. The result is that the surrogate surface, rather than being an interpolator, instead is weighted by the accuracy of the individual sample sites to which the model is fit. This allows for more 'weight' to be given to the most accurate models and eliminates any restrictions on where sample sites can be added. That is, the surrogate model will simply tend closer to

points from more accurate models and further from the less accurate observations. The conceptual approach for this surrogate model is discussed with more rigor in Section 3.4 and the mathematical derivation of this surrogate model is given in Section 4.1.

This surrogate modeling approach is only one of the two critical issues associated with the VGO algorithm. The second critical issue is concerned with sequential sampling and assessing the value of a particular analysis at each step in the optimization process. This leads to the second Secondary Research Question:

Secondary Research Question 2: *How can the most valuable design site and analysis be dynamically selected at each step in the optimization process?*

Secondary Hypothesis 2: *Maximizing the Value of Information (VoI) provides a metric for choosing the next design site and associated analysis model at each step in the optimization process.*

For the VGO optimization algorithm, the surrogate modeling approach from *Secondary Research Question 1* is used in conjunction with the Value of Information (VoI) metric for sequential sampling. This means that at each iteration in the optimization, a new sample site is added to further guide the optimization process. VoI allows for effective selection of this new point and analysis by providing a mathematical means for calculating value for any site in the design space for any of the available models. Then, it is merely a matter of picking the design site and analysis with the maximum value.

The VoI metric comes from information theory but lends itself readily to design optimization problems. In the context of VGO, VoI automatically takes into account the

likelihood of finding an improved solution, given the cost of performing an analysis and the model accuracy. A very powerful property of the VoI metric is that it naturally balances global search with local refinement without the specification of any user-defined tuning parameters. Sometimes a very cheap analysis that reduces the uncertainty in the global space may be of more value than a very costly analysis in the vicinity of the current optimum. Another unusual property of VoI as compared to other sequential sampling criteria is that it provides an intuitive stopping criterion requiring no user-defined parameters. The VGO algorithm stops when the VoI is less than or equal to zero; that is, no more additional analyses are performed when the potential benefit is outweighed by the cost. Because of the above properties, the VoI metric sequential sampling, in conjunction with the Gaussian process surrogate for multi-accuracy data, provides an effective way to navigate the design space during optimization.

1.5 THESIS OVERVIEW AND ROADMAP

In this chapter, the motivation for Value-Based Global Optimization (VGO) was presented. The critical issues of managing multi-accuracy data and selecting cost effective analyses during optimization were defined. This led to a general, proposed approach and associated research questions and hypotheses. To summarize briefly, a new global optimization algorithm is needed because existing global optimization approaches do not allow for multi-accuracy analyses from any number of models without restriction as to where design sites can be added. Additionally, current global optimization approaches rarely account for analysis cost during the sequential sampling process. Also, current global optimization approaches generally rely on user-defined stopping criteria rather than stopping when it is no longer valuable to continue sampling. These

limitations prevent existing global optimization approaches from being as efficient and cost-effective as possible. We hypothesize the VGO can outperform comparable global optimization algorithms by leveraging multi-accuracy data and a value-based sequential sampling strategy.

In the next chapter, a critical review of related work is given. Background for this algorithm covers several different disciplines, including relevant literature in design and systems engineering, design optimization, surrogate modeling approaches, sampling approaches, variable accuracy modeling approaches, information theory, and global optimization algorithms. Particular attention will be given to similar global optimization algorithms so that theoretical and experimental comparisons can be made.

In Chapter 3, the theory behind Value-Based Global Optimization is introduced and the conceptual approach is given. The goal of this chapter is to provide a relatively simple illustration of each of the components of the algorithm and how they interact. Starting with a pictorial representation of the algorithm and a simple pseudo-code representation, explanations are then given regarding the two main contributions: the Gaussian process-based surrogate model for multi-accuracy data and the Value of Information (VoI) metric for sequential sampling during the optimization process. Background information and basic mathematical descriptions are provided, particularly for the VoI. This chapter also addresses the problem characteristics and the basics of selecting an appropriate objective function and models to solve the problem.

In Chapter 4, the focus is on the technical details of the VGO algorithm. The first contribution of this chapter is a detailed mathematical description and derivation of then

Gaussian process-based surrogate model which can accommodate data from any number of models at different accuracies. The second contribution is a detailed description of the mathematical formulae used to calculate the Value of Information during the sequential sampling step of the algorithm. Mathematical details regarding some particular calculations are given, and the problem setup is discussed in greater detail. Problem initialization, costs, accuracies, and nested optimizations are addressed. After presenting the mathematical details of the algorithm, an illustration of the working algorithm is provided. Finally, Chapter 4 is concluded with a summary of contributions and a description of characterizations needed to compare the VGO algorithm to other global optimization approaches. .

Chapter 5 is focused heavily on demonstration and characterization of the VGO algorithm. Several experiments are presented and the results explained. Specifically, a method for creating a meaningful test suite for comparing VGO to other global optimization algorithms is provided. Then, the defined test suite is used to run comparisons between VGO and the Efficient Global Optimization (EGO) algorithm. The results are analyzed with respect to solution quality and costs incurred, and it is shown that VGO is superior for the particular test suite with statistical significance. After this comparison, more experiments were run to show the versatility and scalability of VGO. The VGO algorithm is used to solve two known global optimization problems in three and six dimensions. Finally, based on some experimental results, qualitative discourse is provided for selecting meaningful combinations of models in terms of accuracy and cost in the context of VGO.

While Chapter 5 deals primarily with theoretical test problems, the focus of Chapter 6 is a more advanced engineering example. In Chapter 6, multiple models of a hydraulic hybrid car are used in conjunction with VGO to perform a profit optimization. The intention of this example is to show that the VGO algorithm is viable and effective in real world applications. In this chapter, we walk through the model construction for the hybrid models of different accuracies as well as a brief description of demand modeling in order to calculate the expected profit for a particular design artifact. Selection of appropriate costs and model characterization with respect to accuracy are discussed, and the chapter concludes with a discussion of the results and associated computational time and cost.

The thesis concludes with Chapter 7 where a summary of the research questions, hypotheses, and contributions are provided along with a critical review of the achievements and potential shortcomings of the VGO algorithm. Opportunities for future work are delineated.

CHAPTER 2: LITERATURE REVIEW

In the previous chapter, the context, motivation, and problem were defined, leading to the critical issues and research questions regarding the VGO algorithm. In this chapter, the focus is on surveying the relevant literature and providing a critical review of similar work.

To adequately address all of the aspects of VGO, many different genres of literature need to be surveyed. The chapter will proceed as follows. First, relevant literature from decision making in design and systems engineering will be reviewed. This section helps to frame the context from which the motivation for VGO is drawn as well as common methodologies for setting up and solving design problems. Next, a detailed discussion of surrogate modeling is provided with a specific focus on kriging modeling techniques, followed by a detailed discussion of sampling strategies for constructing surrogates. Following the surrogate modeling and sampling strategy discussion, a detailed description of variable accuracy modeling is provided along with many different examples from the literature. Finally, global optimization algorithms comparable to VGO are presented; many of these algorithms also leverage surrogate modeling or multi-accuracy modeling techniques in the context of the algorithms. The chapter concludes with a return to the thesis overview and roadmap.

2.1 DECISION MAKING IN DESIGN

In chapter 1, some of the challenges of designing and engineering large, complex systems were discussed [25, 57]. Regardless of the system complexity, the engineering design

process generally involves the transformation of design requirements and objectives into a design artifact. That artifact is often iteratively refined in order to best meet the requirements and preferences of the designer and stakeholders [47]. When large systems are designed, there are often several distinct stages to this process and several iterations of design artifacts.

The engineering design process may have several stages, but for the purpose of this thesis it is useful to distinguish between two phases of the process. The first phase is often referred to as conceptual design. During conceptual design, basic design concepts are generated. More formally, system architectures are abstracted in terms of subunits and their interactions. For example, a designer needs to create a man-powered transportation vehicle. In the conceptual design phase, he might generate some concepts at a high level of abstraction, some of which might resemble a bicycle, a tricycle, a scooter, a wheelchair, etc. The mere act of generating basic solution structures is non-trivial, as creativity is required as well as an understanding of how the interactions of the various subsystems will affect the success of the final design. As one might imagine, this phases becomes infinitely more challenging as the systems being designed increase in number of parts, number of subsystems (e.g. electrical, controls, mechanical). This is the more open-ended phase of the design process, and is not the focus of this thesis.

This thesis is targeted more toward the second phase of the design process. This phase occurs once the overall system architecture is known and the sizing and refinement stage of the design process can begin. For example, the decision maker has already decided that he is going to design a bicycle, but there are still many decisions to be made. He still needs to select a wheel diameter, a frame length, and gear ratios. In this phase, modeling

and simulation are invaluable aids in making final parameter selections because full scale prototypes can often be circumvented.

Now the problem becomes a matter of selecting the 'best' sizes and parameter values for the particular system architecture. Selection of the 'best' parameter values depends completely on the decision maker's preferences. These individual preferences may be based on corporate or consumer objectives, but ultimately the decision maker's preferences must be captured in an objective function in order to rigorously determine the value of a particular design artifact. That is, we seek to maximize the overall *utility* of the final design artifact, but doing so requires understanding the tradeoffs that are being made. In the next section, the use and formulation of utility functions for evaluating and comparing design alternatives is discussed.

2.1.1 Utility Functions

In order to unambiguously determine if one design alternative is preferable to another, a mathematically rigorous objective function is required. To that end, this thesis is written from a Decision-Based Design (DBD) perspective [26, 39, 64]. In DBD, it is assumed decisions are best made using mathematically sound methods derived from decision theory; in particular, the use of a *utility function* to quantify a design artifact's goodness is a critical element.

While there are many approaches for formulating a utility function, the goal is to capture a design artifact's goodness from a life-cycle perspective with due consideration given to risk preferences and uncertainty. A meaningful utility function allows a decision maker to understand the tradeoffs being made by choosing one design alternative over another.

A common approach for utility function construction in design applications is to use Multi-attribute utility theory (MAUT) [33], which is an extension of von Neumann and Morgenstern's utility theory [65]. An *attribute* is a measurable design outcome which affects the artifact's perceived goodness. For example, in designing a car, a decision maker may select several relevant attributes that affect the utility, including maximum speed, fuel economy, and maximum acceleration. Note that attributes are not design variables; rather, they are affected by the choice of design variables. In theory, MAUT provides a mathematically sound way of considering different attributes of a design with different units into a single function with a consistent set of units. However, MAUT can easily lead to meaningless and inconsistent results when more than two attributes are being traded off.

The complications associated with MAUT have led others in the design research field to pursue a single criterion method that adequately represents an artifact's utility. A logical choice is to use economic value or profit to characterize the goodness of a particular design alternative [66]. Using expected profit as a utility function eliminates many of the complications associated with MAUT, but is not without its own share of challenges. In order to predict the profitability of a particular design artifact, it is necessary to characterize the demand for that artifact. How many people will buy this particular artifact, and at what price? Predicting profit requires a mapping between the attributes and expected sales and must account for the cost of creating the design artifact and the price at which it is sold. These issues are addressed in the context of the engineering example of designing a hydraulic hybrid car presented in Chapter 6.

A very positive upside to using profit as a utility function is the fact that the units are always currency. In this thesis, where analysis cost is being explicitly accounted for, it is relatively straightforward to account for costs incurred in the final utility function. Recall that engineers are mainly interested in good solutions at reasonable cost, not mathematically optimal solutions. A profit function lends itself to this type of formulation quite naturally. Very simply, the expected utility of a design artifact is the expected profit from that artifact less the analysis costs incurred during the design process to select that particular alternative.

2.1.2 The Role of Optimization in Design

As discussed in Chapter 1, design problems are often formulated as optimization problems. In the previous section, utility functions for quantifying a design alternative's goodness were discussed; these utility functions serve as objective functions that a designer seeks to maximize. Logically, a designer would choose the design alternative with the maximum expected utility. An optimization scheme is often useful in this context.

Optimization is a very mature field in the engineering and mathematical communities and can be used throughout the entire design process to aid in the decision making process. Depending on the stage of the design process (e.g. conceptual or refinement) as well as the nature of the design problem, optimization takes on many different functional forms. During conceptual design process, there is often a high cost for generating alternatives, as well as a lot of uncertainty. The number of distinct concepts considered by designers at this stage is typically small; so, the 'optimization' taking place is often done by brute force, Pugh selection [50], quality function deployment [1], or by calculating a utility

associated with each alternative [42]. These are not typical optimization schemes in the sense that the design space is not continuous, nor are the examples relevant for mixed integer optimization approaches. In these cases, optimization is merely a more formal way of selecting which alternatives to explore further as compared to a designer making an ad hoc decision.

As we move into an age where generating design concepts and associated models of said designs is not so expensive and is becoming increasingly automated, a greater number of candidates can be considered at early stages in the design process. However, as the number of potential candidates grows, there are some inherent difficulties with performing a thorough evaluation of all potential design candidates. First, depending on the domain, performing a rigorous simulation or developing a detailed model of each candidate may be computationally prohibitive. Additionally, even if each candidate could be modeled in a reasonable amount of time, this type of design space is often discrete and multi-modal, so conventional gradient based optimization is often not applicable.

Because of the discrete nature of some design spaces, attention has been given to evolutionary algorithms and other stochastic optimization algorithms. These algorithms are often capable of obtaining near global optimality even in noisy design spaces exhibiting multi-modality and/or discontinuities [7, 46]. Much success has been achieved in solving complex engineering problems using evolutionary techniques [17, 71], but more study is needed for using these techniques during the conceptual design process.

Optimization is more often employed in the refinement stages of the design process when the overall system architecture is known. A more classical optimization scheme can be used to determine final parameter sizes (e.g. diameters, gear ratios, engine sizes, etc.) based on a utility function. In this later stage of the design process, the design space is more often continuous, making classical, gradient-based optimization methods fairly easy to apply.

However, even in these cases where the solution structure is already known, the cost of running an optimization can be high. The models used to evaluate the system architecture can be computationally expensive, but beyond that, optimization requires many function evaluations and optimization under uncertainty involves even more function evaluations. Even if the system model only takes a couple minutes to run the optimization may be prohibitively expensive. For example, if an optimization under uncertainty requires 1000 steps and 1000 Monte Carlo samples per step, at two minutes per simulation, this optimization would take 3.8 years. Therefore, improving the efficiency of the optimization process without sacrificing solution quality is a very important research issue in design.

One approach to managing the cost of optimization is to use low fidelity approximations of high fidelity models that can either assist or take the place of these expensive simulations during the optimization process. This idea is discussed in detail in the next section, along with several specific approaches.

2.2 SURROGATE MODELING

A *surrogate model* is a mathematically tractable function that is used to approximate a complex or black-box model. Consider the optimization example from the previous section where a particular model took two minutes to evaluate and the optimization would have taken 3.8 years. If the simulation model were replaced with a surrogate model taking only .2 seconds to evaluate, the optimization would be complete in just 55 hours.

Many different types of surrogate models are prevalent in the engineering and optimization literature [9, 13, 30, 31, 46, 54, 55, 60, 61, 67]. Surrogate models are not only used for optimization; they can also be used for design space exploration and visualization because many observations of the surrogate can be made at low cost. The primary objective of a surrogate model is always to create a computationally tractable approximation of an otherwise expensive model so that the cost of a particular process can be reduced while still retaining an acceptable degree of accuracy.

To create these models, sample sites of the high accuracy model are required. It is impossible to circumvent the use of expensive, high accuracy models altogether, but the amount of function evaluations to *seed*, or fit, the surrogate model is far fewer than would be required to run an entire optimization. At this point, it is useful to review some terms from Chapter 1. Recall that a *sample site* of a model includes a *design site* and an associated *observation,* that is, a sample site includes both the x-coordinates or inputs and the calculated output from the model. These sample sites are used to seed the surface, and an approximating function is fitted to the given data. The resulting approximation is computationally inexpensive to evaluate at any point in the space, so it can be used for

optimization or other exploration very efficiently. A fairly detailed taxonomy of surrogate surfaces is presented in [61] and in [30], and some common types include polynomial approximations, radial basis functions [12, 17, 46, 71], spline approximations [14], and kriging models [9, 10, 16, 23, 41, 54, 55, 61, 67]. While all approaches have their advantages and drawbacks in terms of computational complexity and accuracy, the remainder of this thesis draws on kriging modeling. Kriging modeling is particularly well-suited for applications in computer experiments and is explained in detail in the next section.

2.2.1 Kriging Modeling

Kriging modeling is an interpolation technique that has its origins in geostatistics literature [10] but is now very prolific in engineering and computational experiments. Mathematically, kriging models consist of a sum of a regression model and a realization of a zero-mean Gaussian process realization. This model construction allows for a lot of flexibility in model properties and fits. The basic premise of kriging modeling is shown in Figure 2.1. We see the same set of sample sites shown in Chapter 1, only this time we show the basic components for fitting a kriging model to this data. First, a regression model is fit; in the picture, this is a zero-order regression model equal to the mean of the observations. Then, a zero-mean Gaussian process realization makes up the difference between the regression model and the design sites, and this is added to the regression model to get the resulting kriging fit. While this is a very high level explanation of kriging modeling, this basic representation should allow for better understanding of the relevant kriging literature and discussion.

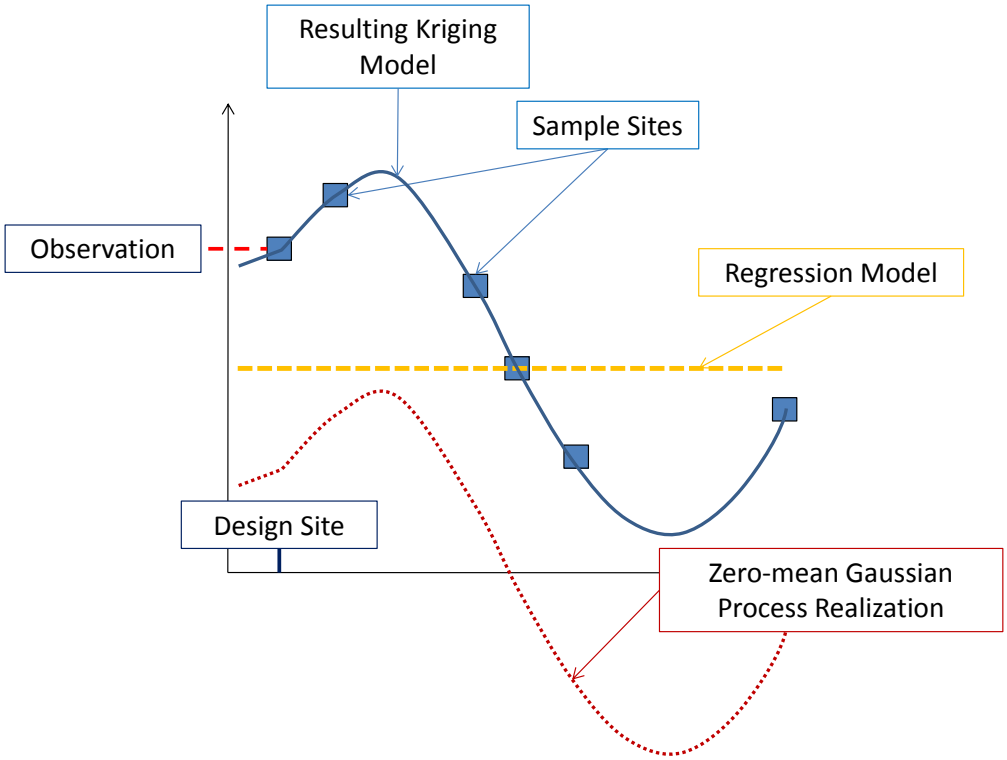

FIGURE 2.1: KRIGING MODELING OVERVIEW

2.2.1.1 The Advantages of Kriging Modeling

Kriging modeling was first applied to computer experiments by Sacks *et al* [55] because of certain desirable properties. The first desirable property of kriging models is that they are interpolators; a kriging predictor evaluated at a design site always returns the exact observation value at that design site. This is particularly useful for computer experiments (as opposed to physical experiments), because computer experiments are often deterministic and lack measurement noise. Generally, there is no need to assume any uncertainty associated with the sample sites.

The second property of kriging models that make them desirable for computer experiments is that they are rooted in statistics. Many of the assumptions made in the

formulation of the kriging predictor are good assumptions in design and engineering domains; the resulting surfaces are reasonably smooth and observations taken from nearby design sites are often highly correlated. This means that kriging models can result in a good fit for design and engineering problems. In addition, because kriging models are statistical in nature, the predictor parameters can be determined without user input by using Maximum Likelihood Estimation (MLE). That is, kriging models do not require any user-defined tuning parameters; everything can be determined mathematically by leveraging the statistical nature of the model.

A final property of kriging models that causes them to be very prolific in the engineering literature is that kriging models also have the advantage of providing an equation for the mean-squared error (MSE) of the predictor. The MSE provides a single metric for assessing the kriging model's accuracy at any location in the design space. While the model being approximated is assumed to have zero error, there is error between the kriging predictor and the underlying model, particularly at points in the design space that are far from any design sites. This allows the designer to make decisions knowing how much confidence he or she can have in the kriging model.

2.2.1.2 The Shortcomings of Kriging Modeling

While kriging models have many appealing attributes for computational experiments, they do also have some shortcomings. One limitation of kriging models is that they can be computationally expensive to generate. There are two aspects to this expense: the first is the expense of generating the sample sites to which the kriging model is fit, and the second expense is the actual fitting of the kriging model. The cost of evaluating the sample sites is not a property of the kriging modeling technique and will be addressed

later. However, it is important to discuss the expense associated with fitting the actual kriging model.

Generation time for kriging models increases exponentially with the number of design sites, so particularly for large space with many sample sites, the cost of fitting a kriging model is non-negligible. Some effort has been made to reduce the computational effort involved in fitting a model to a very large design space. In an optimization context, for example, adaptive kriging methods can be used. In an adaptive method, only the nearest neighbors of the design site under consideration by the optimizer are used to construct a localized kriging model [9]. By choosing a relatively small number of nearest neighbors (<1000), the computation time for constructing the localized kriging model is small. However, since only localized models are constructed, this technique is not appropriate for visualizing the entire design space.

The remaining shortcomings of kriging predictors discussed here have something to do with its interpolation properties. This may seem inconsistent, as interpolation was just described as a desirable property for many applications. There are some applications, however, where interpolating the sample sites is not the most desirable outcome and poses a restriction on the types of problems to which kriging can be applied. In particular, the interpolation assumption does not explicitly allow for model error. That is, it must be assumed that the underlying model is perfectly accurate; the error at the design sites where the original model has been sampled is identically zero. If this assumption is valid, then it is logical to interpolate the given data. However, should it become necessary to recognize model error, as will be seen in this thesis, then the classical kriging formulation will not suffice.

An extension of the interpolation limitation imposed by kriging is that interpolating the sample points is only desirable if the design sites are only drawn from one coherent set of data from one model. Combining data from multiple models of varying accuracy and interpolating the design sites would not result in a meaningful fit. More importantly, having multiple data points at the same input, even if they came from different models, would cause a classic kriging fit to fail due to a singularity. To handle data from multiple models with uncertainty, as is done in this thesis, a modified kriging approach will be required.

2.2.1.3 Modified Kriging Modeling Approaches

In the literature there have been some proposed modifications to the classic kriging approach in order to either relax the interpolation constraint and/or to accommodate data from multiple models at multiple accuracies.

One modified kriging formulation is known as stochastic kriging [5]. This kriging formulation does relax the interpolation constraint to allow for *model error*. However, this model error is not the same as *model inadequacy*. Model error, in this context, is due the stochastic nature of discrete simulations; these simulations are not deterministic, so a single observation is not representative of the expected value of the simulation at a design site. To adjust the kriging fit accordingly, an uncorrelated zero-mean Gaussian error term, or white noise term, is added to the kriging model beyond the regression and Gaussian process realization terms. The result is that the kriging model does not interpolate the sample sites exactly but rather has a bit of a smoothing effect. Model inadequacy, on the other hand, is a key factor in the VGO algorithm presented in this thesis. Model inadequacy refers to the difference between a particular deterministic model and reality;

the error itself is deterministic but unknown. Model inadequacy cannot be addressed the same way as model error in the stochastic kriging formulation.

Other modified kriging algorithms have been developed to accommodate model inadequacy and more specifically, to accommodate sample sites from any number of different models at different accuracies [29, 34, 35, 51]. These kriging modeling approaches each have their own shortcomings. For example, in [51], design sites from the higher fidelity models can only be added where the available lower fidelity model has already been sampled. This means that the design sites for the high fidelity model are a subset of the design sites for the low fidelity model. Some versions of kriging modeling for multi-accuracy data assume that the different models are correlated, which requires estimation of a large number of hyperparameters as compared to the VGO surrogate modeling approach presented in Chapters 3 & 4 [29].

2.2.1.4 Kriging Modeling Conclusions

This section provided a relatively lengthy discussion of kriging literature and variations on classic kriging modeling approaches. Many advantages and shortcomings were identified; some of these will be re-addressed in the derivation of the surrogate modeling approach for VGO found in Chapter 4. While other researchers have provided some similar kriging modeling approaches, the approach used in VGO seems to be unique in that there is no restriction on number of models sampled or where samples can be added and from which models. Additionally, the assumed correlation structure is different in VGO from most of the other formulations; the resulting model is more intuitive with fewer parameters to estimate.

As for any surrogate model, the accuracy of a kriging model, modified or not, is going to depend on the number and quality of sample sites that are used to seed the model. In the next section, *sampling strategies* are discussed.

2.3 SAMPLING STRATEGIES

A sampling strategy is a method by which design sites are selected for the purpose of fitting a surrogate model. Surrogate models of all varieties are fit to some sample sites obtained by evaluating a particular model at a particular design site. Depending on the number and quality of sample sties, the quality of the surrogate models will differ. Generally, larger numbers of sample sites lead to more accurate surrogate models.

There are two general strategies for sampling. One option is that a surrogate model can be fit to a predetermined set of samples selected using either a design of experiments or other selection criteria for a given number of samples. That is, the design sites are strategically selected a priori, evaluated, and then the surface is fit to the samples. Alternatively, design sites can be added incrementally using information from the current iteration of the surrogate model, with the model being dynamically updated as new samples are added. If the intent is to get a reasonably accurate view of the entire design space, then a global, fixed sampling approach is likely to be a good choice. On the other hand, if optimization is the primary concern, it is not economical to sample the entire design space evenly or even to achieve uniform accuracy throughout the space; for optimization, high accuracy is only needed near the optimum. A detailed discussion of fixed sampling strategies is given here, followed by a description of several incremental sampling strategies.

2.3.1 Fixed Sampling Strategies

Fixed sampling approaches require the a priori selection of design sites using proven statistical techniques. These design sites are then evaluated using the underlying computational model, and the surrogate model is fit to the sample sites; once the surrogate is fit it is not generally modified. Fixed sampling approaches are most commonly used when it is necessary to accurately model the entire design space; that is, the objective is to visualize a complex function while saving costly function evaluations by using a surrogate.

One if the most common approaches to fixed sampling is to use a space-filling design of experiments to select the design sites prior to sampling and fitting. One logical choice would be to simply use a Full Factorial sample and create a grid, but this approach is neither very efficient nor very strategic. Other fixed sampling techniques include Latin Hypercube Sampling (LHS) [43], Quasi-Latin Hypercube Designs (Quasi-LHD) [69], Cascading Latin Hypercubes [22], Orthogonal Arrays [62], and maxi-min variants of the above [32].

Latin hypercube sampling (LHS) [43] is a very commonly used stratified sampling technique that has been used in conjunction with kriging models [67]. While Latin hypercube designs ensure that the model is sampled throughout the range of all the input variables, it has been shown that LHS is often too sparse for capturing all of the details of complex models [55]. While simply increasing the number of samples may reduce the error of the kriging model, adding samples in areas in which the model is already sufficiently accurate is a waste of resources. Because kriging models include a closed form solution for MSE, this parameter can be leveraged to come up with more strategic

space filling sampling approaches. This leads to some different types of designs for fitting a kriging model where the samples are not equally spaced or rigorously stratified, but rather, spaced in a way that guarantees optimal accuracy with respect to a particular metric for a fixed number of samples.

There are many different metrics that have been used for selecting optimal design site configurations in fixed sampling situations with a predetermined number of design sties [55]. These metrics are specifically tailored for kriging modeling in that they rely on the estimate of MSE given by the kriging model. One such metric for selecting design sites is known as minimizing Integrated Mean Squared Error (IMSE). In this method, the design sites are selected such that the integral of the MSE function throughout the design space is minimized. It is useful to recall that the MSE at a design site for a classic kriging model is identically zero. Logically, MSE increases with distance from design sites. This metric seeks to minimize the overall integral of MSE; the best solution will have the small MSE consistently throughout the design space. This metric is appropriate when global accuracy is desired.

Another metric presented for fixed sampling is Maximum Mean Squared Error (MMSE). In MMSE, design sites are selected such that the maximum of the MSE in the design space is minimized. While this metric is still appropriate for global accuracy, the emphasis of MMSE is different from that of IMSE; for MMSE the maximum MSE present at any point in the design space is minimized, whereas IMSE does not bound the error at any one point, but rather seeks to minimize the total amount of error in the design space. One last fixed sampling metric worthy of mention is an entropy criterion, whereby

a Bayesian design approach is taken to select the design sites to minimize the expected posterior entropy.

While all of these methods are appropriate in a global exploration context, none of them is ideal for design optimization. First, by using a fixed sampling strategy, the number of design sites is pre-specified without any mathematical justification as far as the cost incurred and the desired quality of the resulting surrogate. Additionally, all of these methods are generally targeted towards global accuracy; none of these methods allow for accuracy to vary throughout the space. If the accuracy of the surrogate model is not allowed to vary throughout the design space, then no resources are specifically channeled toward finding the optimum. In fact, resources are simply allocated evenly throughout the design space, even in areas that lack promising solutions. As was discussed in Chapter 1, it makes sense from an engineering design optimization perspective to allocate a lot of resources in promising regions but not to waste time accurately assessing poor solutions. To accomplish this non-uniform accuracy surrogate model, it is important to examine incremental approaches for adding design points so that the current state of knowledge dictates where it is most valuable to expend additional computational effort.

2.3.2 Sequential Sampling Strategies

While fixed sampling strategies are simple to implement and can ensure reasonable global accuracy, they are largely not conducive to channeling the computational resources toward finding the optimum. In this case, it makes more sense to start with a smaller number of samples to fit an initial surface and then incrementally add design sites to increase accuracy in promising areas or in areas with high uncertainty.

Metrics for incremental sampling are present in the literature, primarily in situations where the emphasis is on optimization rather than global exploration. However, it should be noted the individual metrics used to select additional design sites are not specific to global optimization. It is perfectly plausible to implement an incremental sampling scheme with the intention being global accuracy. For example, the same criteria presented in the previous section (IMSE, MMSE, posterior entropy) can be employed in incremental schemes. The difference is that rather than deciding how many points to sample beforehand, the same metrics can be used to guide the selection of only one additional design site or even a few design sites at a time. After the new sites are added, the surface is updated and additional design sites can be selected for evaluation and added to the surrogate model in the next iteration.

In the global optimization literature, several metrics are presented for selecting the next design sites based on the current kriging fit [30]. While these metrics are all used in incremental sampling schemes, the emphasis for many of these metrics is strictly optimization. Unlike IMSE and MMSE, these sequential sampling metrics are designed to favor finding the best solution at high accuracy, rather than focusing on the overall accuracy of the entire space.

One commonly used sequential sampling metric is maximizing the Probability of Improvement (PI). The PI metric is a function for calculating the probability of finding an improvement over the current best design site in the next iteration. To calculate PI, the user must provide a pre-specified target value. For example, at any candidate design site in the design space, a user might want to know the probability of improving over the current best by 5% or more, where 5% is the target value. This target value is quantified

as a percentage improvement over the current best, and it is also possible to use multiple target values at once. Conceptually, this is an appropriate metric for global optimization; samples are added if and only if they have the highest probability of resulting in improvement over the current best point based on the current kriging model. Additionally, sampling stops when the probability of improving by some percentage is equal to zero for every candidate design site in the design space. However, the most obvious disadvantage of this approach is that the target value must be specified by the user. The selection of an appropriate target value is likely to have a high impact on the quality of results achieved. Also, while the stopping criterion is intuitive, the cost of performing additional function evaluations is not taken into account. It is likely that situations would arise when the cost of running an additional analysis would outweigh the improvement that could be achieved over the current best.

Another incremental metric presented in the global optimization literature is Expected Improvement (EI) [19, 29-31]. This metric is similar to the PI approach, but in this metric the expected *amount* of improvement over the current best is maximized. Rather than targeting a design site that has the highest probability of yielding improvement, here the design site with the greatest amount of expected improvement is selected. Again, this criterion is appropriate in global optimization; the focus is not on global exploration and uniform accuracy, but on finding the global optimum. There is a simple stopping criterion as well; stop when the expected improvement is smaller than some constant. The obvious drawback is that the user must define this constant and, much like PI, the quality of results achieved are sensitive to this choice of constant.

There is also a multi-point version of this criterion [19] in which multiple design sites are selected for evaluation in only one iteration, which reduces the cost of fitting a kriging surface after each new point is added. The shortcomings of this metric are that uncertainty and model accuracy are not taken into account, nor is the cost of analyzing the new design site.

Finally, the incremental sampling strategy most similar to VoI that has been presented in the literature is the augmented expected improvement metric, or augmented EI. Modified EI does account for the cost of additional function evaluations [29]. However, modified EI also includes a number of user defined parameters that can affect the searching behaviors and the stopping criteria. Specifically, augmented EI includes three user defined parameters that reflect the reduction in value for less accurate analysis models, the reduction in value due to random error, and the reduction in value due to cost of analysis. However, these user-defined parameters are not transparent to the user. While much of the same data is incorporated in both the proposed VoI metric and the augmeted expected improvement metric (risk preferences, uncertainty, etc.) they are not mathematically equivalent and VoI is hopefully a bit more transparent to the user. Perhaps most importantly, the augmented EI metric, just like the original EI metric, requires a user-defined stopping criterion, and the success and efficiency of the optimization is likely to be sensitive to the choice of this stopping criterion. Because the augmented EI function was specially designed to work in conjunction with models at varying accuracies, it seems logical to now discuss the literature regarding variable accuracy modeling.

2.4 VARIABLE ACCURACY MODELING

Variable accuracy modeling implies the use of multiple models of differing accuracies and costs for the purpose of creating more accurate surrogate models and/or performing optimizations efficiently and effectively. The idea is that less accurate, inexpensive models can be used to perform global search while more accurate, more expensive models are reserved for local refinement in areas of interest. The results achieved using models of varying accuracy should, in theory, be more accurate than densely sampling using only a low fidelity model and much less costly than sampling with only a high fidelity model. That is, variable accuracy modeling should afford a solution to the designer's dilemma presented in Chapter 1 by enabling both broad exploration and high accuracy while incurring only modest computational expense.

Specific examples of variable accuracy modeling from the literature will now be discussed. Recall that *fidelity* and *accuracy* are often used interchangeably in the literature, but in this thesis it is assumed that fidelity is a property of a model reflecting the amount of knowledge captured in that model, while accuracy describes the closeness between reality and a simulation outcome. Variable accuracy modeling has two different aspects to it: one is how to create the individual models to be sampled, and the other is how to leverage the multi-accuracy data acquired from the available models. The literature contains many approaches to both aspects of the problem.

The idea of using multi-accuracy data in the optimization process for engineering design dates at least to the seventies [59]. One of the simplest approaches to multiple accuracy modeling in an optimization context is to create feasibility constraints that can be tested quickly using an inexpensive model. If certain conditions are not met, then no further

function evaluations take place. This type of strategy is used in conjunction with an optimization framework by Paredis [48], and by Gurnani et al [20, 21].

Other approaches to multiple accuracy modeling include the space mapping approach [6], which attempts to create a mapping or correction between a coarse (low accuracy) design space and a fine (high accuracy) design space that yields the same computational outcome.

Seminal work in variable accuracy modeling has been done by Alexandrov *et al*. [2-4]. In one of their examples, an aerodynamic optimization is performed using the Euler equations over variable mesh sizes, effectively changing the resolution of the model. In another approach, variable-fidelity physics models are used, where the high-fidelity model is the Navier-Stokes equation and the low fidelity model is the Euler equation. In both cases, the method of correlation for the low and high accuracy data is a first order error function in a given trust region using augmented Lagrangian methods, which have been shown to converge to a Karush-Kuhn-Tucker (KKT) feasible point for constrained minimization problems [52]. Using the low fidelity model and this corrective factor, nested optimizations are performed on the low fidelity model, and then the trust region is adjusted based on the performance of the high fidelity model. While this method requires relatively few function calls to the high fidelity model, the method is restricted to derivative based optimization approaches.

Other similar works [15, 38, 53, 68] apply a very similar trust region optimization technique successfully using one of the low fidelity model types presented by Alexandrov, or by using a surrogate model as the low fidelity model. All of the examples

presented in these works are in the aerospace domain, typically dealing with optimization of airplane wings or other control surfaces for aerial vehicles.

Much work has focused on the use of surrogate models as low fidelity models. Instead of replacing a complex function with a surrogate, the idea is to leverage both during optimization. Some variable-fidelity optimization literature exploits the use of local response surface approximations (RSA's) with a variety of different sampling and interpolation techniques; this is particularly common in the Multi-Disciplinary Optimization (MDO) domain. One approach is to use response surface approximations with high-fidelity sampling in conjunction with the trust region methodology presented by Alexandrov *et al* [49]. In [52, 53], a concurrent subspace optimization technique is used and a comparison of RSA constructions is provided. Other surrogate assisted work includes the use of radial basis function surrogates during the optimization process [17, 46, 71]. Zang provides an overview of the surrogates and optimization techniques applied to the MDO domain [70].

Most of the prior work presented thus far in this section makes use of gradient-based optimization techniques; the infrastructure in these multi-accuracy modeling is often specifically tailored towards those gradient-based techniques. Some other related work, however, does make use of stochastic optimizers in conjunction with variable fidelity models [17, 20, 21, 46, 68, 71]. Most of these approaches are similar to those previously presented in that they still use local interpolation or surrogate surfaces.

As illustrated by the prolific literature on the subject, models at multiple levels of accuracy can aid in the optimization or design space exploration processes by enabling

inexpensive search coupled with high accuracy. While there are many approaches to managing multi-accuracy data during optimization, many of the previously delineated approaches have some shortcomings. Several approaches to multiple accuracy modeling are limited to only two analysis models, and moreover, do not explicitly account for the cost of the analyses used during the process [2, 4, 15, 38, 44, 52, 53, 68]. Therefore, what is needed is a method to combine predictions from any number of models of different accuracies so that all of the relevant information can be used and weighted according to its accuracy in a cost effective manner.

So far in this chapter, systems design, surrogate modeling, sampling strategies, and variable accuracy modeling have all been discussed in isolation. In the next section, specialized global optimization algorithms that incorporate some or all of these techniques are presented. Special attention is given to algorithms that are comparable to VGO.

2.5 GLOBAL OPTIMIZATION ALGORITHMS

Optimization is a very mature field in the engineering and mathematical communities. In Section 2.1.2, the role of optimization in design and systems engineering was discussed. In this section, the focus is not on individual optimizers so much as on global optimization algorithms, which are really more like optimization suites. These algorithms are tailored more directly to design and computer experiment scenarios where computationally expensive models are often involved. Often, these algorithms leverage surrogate models, models of varying levels of accuracy, and/ or sequential sampling strategies along with classical optimization algorithms to solve global, multi-modal problems involving costly models and simulations [3, 29-31, 34, 35].

While it is theoretically possible to solve design problems by using classical optimization techniques, it is usually not practical to do so. The problem with applying traditional optimization techniques to systems design is the number of function evaluations required per iteration. For a gradient based optimizer in three dimensions, each step of the optimizer would require four function evaluations, one at the current location, and three to compute the gradient. Because design problems can be multi-modal, a simple gradient-based optimizer is not likely to find the global optimum with only one starting point. If the underlying model is expensive to evaluate, it can be computationally prohibitive to evaluate a high fidelity model multiple times at each step in the optimization process. Finally, classical optimizers are not designed to take the cost of optimization into account. While optimization techniques are very useful and are used in the optimization schemes presented in this section, it is important to frame engineering design problems in a practical way that favors good solutions achieved at a reasonable cost rather than mathematical optima.

Approaches to global optimization in this domain have tended to be based either on kriging modeling or on models with differing accuracies, but not both. The most closely related approach to the VGO algorithm is the Multi-Fidelity Sequential Kriging Optimization (MFSKO) algorithm, based on the augmented expected improvement metric [29]. The MFSKO uses a multi-fidelity kriging modeling technique but assumes that the higher fidelity design sites are subsets of the lowest fidelity design sites. Based on the MFSKO predicted mean and MSE, the augmented expected improvement metric is then used for sequential sampling. The augmented EI does account for cost and includes three user defined parameters that reflect the reduction in value for less accurate

analysis models, the reduction in value due to random error, and the reduction in value due to cost of analysis. However, the user-defined parameters are not transparent to the user. Our approach uses the VoI metric which includes much of the same information (risk preferences, uncertainty, etc.) but is mathematically different from augmented EI and more transparent for the user. Perhaps most importantly, the augmented EI metric (much like the EI metric presented in EGO) requires a user-defined stopping criterion and does not necessarily stop when it is no longer valuable to continue to run analyses. In the next two sections, we describe in detail our approach for variable fidelity surrogate modeling and incremental sampling for global optimization.

2.5.1 Efficient Global Optimization

Efficient Global Optimization (EGO) is a global optimization algorithm presented by Jones *et al* [31]. EGO relies on a kriging surrogate model coupled with a rigorous metric for selecting additional design sites called Expected Improvement (EI). Seminal work in the area of using stochastic models for global optimization was done by Kushner in the 1960s [36]. The idea was that stochastic models could be used to aid in the optimization of multi-peak, multi-modal functions that are often encountered in engineering design, and that statistical information could be leveraged to select new design sites.

The Expected Improvement (EI) metric used in EGO leverages information from the kriging model to balance global search, indicated by areas with high uncertainty or MSE, and local refinement, indicated by areas where the surface has a promising objective value [28]. To calculate the expected improvement at a site x which has not yet been sampled, information from the kriging predictor and the MSE from the kriging predictor are used to provide insight about the unknown simulation output, $y(x)$. While $y(x)$ is not

random, it can be modeled as having a Gaussian distribution with the mean and standard deviation characterized by the predicted value from the kriging model and its associated uncertainty. The expected value of the amount by which $y(x)$ is predicted to improve over the current best observation is known as the Expected Improvement. This sampling metric is computationally inexpensive and fairly intuitive. The stopping criterion imposed states that the optimization stops when the EI is less than some constant—in this thesis, this constant is referred to as a. The problem is that this constant must be defined by the user, so the success and efficiency of the optimization will be dependent on a good choice of this constant.

Aside from the problem of selecting a good constant for the stopping criteria, this global optimization algorithm has a couple other shortcomings. First, the EI metric is completely independent of simulation cost; the expected improvement is based solely on the predictions about the simulation outcome and is not adjusted to reflect the costs incurred by performing the simulation. The second shortcoming is that EI is incompatible with models of varying accuracies. Specifically, the calculated EI would be exactly the same for each x regardless of the simulation to be used to obtain the sample; the accuracy of the model is not taken into account in the EI metric.

Overall, the EGO algorithm has many positive properties due to its statistical nature; it has been applied successfully in the engineering design domain [58]. However, EI is best applied when only one, relatively inexpensive model is available for sampling. In cases where multiple models can be leveraged, particularly if one or more of the models is very costly, then a sampling metric weighting cost and accuracy combined with a surrogate modeling technique which can handle multiple models would be more effective.

2.5.2 Multi-Fidelity Sequential Kriging Optimization

Multi-Fidelity Sequential Kriging Optimization (MFSKO) is a global optimization algorithm with its roots in EGO but with some significant modifications [29]. MFSKO addresses some of the shortcomings of EGO delineated at the end of the previous section. In particular, MFSKO uses a surrogate model that does accommodate variable accuracy models and an augmented Expected Improvement (augmented EI) metric that takes into account model cost and uncertainty. This algorithm is probably the most similar algorithm to VGO currently available in the literature, but there are some significant differences in both the multi-fidelity surrogate modeling approach and the sequential sampling metric.

The kriging-based multi-accuracy surrogate modeling technique used in MFSKO is derived from the surrogate modeling techniques of Kennedy and O'Hagan [34, 35]. There are some slight simplifications in MFSKO, but the overall idea is the same. Unlike some of the previously introduced methods for variable accuracy modeling, this technique allows for any number of models of varying accuracies as opposed to just two. One of the main assumptions is that there must be correlation and some general similarities between all the models used during the optimization process. The resulting correlation structure is more complex than the one for VGO, and typically there are more hyperparameters which must be determined when fitting the model.

The augmented EI function used for sequential sampling is the same as the EI function from the EGO algorithm but with three multiplicative factors. The first factor reflects the correlation between the model being considered for simulation at the untried point x and the most accurate simulation. The factor is equal to one for the highest fidelity model

and zero if the model is completely uncorrelated with the most accurate model. It is effectively a discount factor for using a less accurate model and addresses model inadequacy. The second factor is an adjustment for when the model contains random errors, different from model inadequacy. It reflects an updated posterior for repeated samples and is equal to 1 when then variance for the model is zero. In VGO, it is assumed that the models are deterministic, and the posterior mean and variance is calculated in a more sophisticated manner. The final factor is a ratio between the cost of the model being evaluated and the highest fidelity model; it is literally a discount for using a cheaper model, and is equal to one for the highest fidelity model.

While the augmented EI function captures much of the same information as the VoI metric used in VGO, these factors are less transparent to the user. Also, because the analysis cost is set up as a discount factor and not subtracted from the augmented EI, the optimization does not automatically stop when the expected improvement is equal to the cost incurred. The authors address this, stating that it was indeed a design choice. This setup affords the user to have an objective function that is not necessarily expressed in dollars; with VoI it is assumed that the utility function is a profit function and that costs translate directly to that utility. However, VoI circumvents the need for specifying an arbitrary stopping constant, the same way as is prescribed with EGO.

MFSKO is one of the most sophisticated global optimization schemes available and it leverages surrogate modeling, variable accuracy modeling for any number of models, and sequential sampling strategies. It takes into account model accuracy, random error, and cost. The assumption that utility be expressed in dollars or profit is not necessary in MFSKO. However, the surrogate model is more complex, the augmented EI function

contains discount factors that are not transparent to the use, and it relies on the specification of an arbitrary stopping constant.

2.6 THESIS ROADMAP

In this chapter, the relevant literature from several different genres was reviewed. The chapter began with an overview of systems engineering and design, specifically focusing on the role of utility functions and optimization during the different design stages. Next, surrogate modeling techniques were surveyed with a section devoted specifically to kriging modeling. The surrogate modeling section was followed by a review of different sampling strategies for seeding the surrogate models, including both fixed sampling approaches and sequential sampling approaches. The next section was devoted to variable accuracy modeling techniques, which we established as a way of solving the designer's dilemma presented in Chapter 1. The chapter concluded with a review of a few select global optimization algorithms that leverage some combination of surrogate models, variable accuracy modeling, sampling strategies, and optimization techniques. The global optimization algorithms surveyed are specifically targeted toward optimization problems that rely on expensive computational simulations, and these algorithms are in some way comparable to VGO. In the next chapter, VGO is discussed in detail on a conceptual level.

CHAPTER 3: CONCEPTUAL APPROACH FOR VALUE-BASED GLOBAL OPTIMIZATION

In this chapter, the VGO algorithm is presented. The focus in this chapter is purely conceptual; the mathematical and implementation details are presented in Chapter 4.

The next section in the chapter is dedicated to the problem setup and certain assumption that have been made about the objective function and the available models. The overall approach for the VGO algorithm is then presented in section 3.2; the goal is to show how all of the different aspects of VGO come together to solve design optimization problems. In Section 3.3, the pseudo-code is presented, and is used to guide the discussion for the remainder of the chapter.

In Section 3.4, the approach taken for the variable accuracy Gaussian process-based surrogate modeling technique is discussed. In Section 3.5, Value of Information is explained, from its roots in decision theory to its application in VGO.

Section 3.6 is a review of the general properties, advantages, and disadvantages of VGO and in Section 3.7 the research questions are revisited. This is followed by a discussion of how VGO theoretically addresses the research questions and research gap. Finally, the chapter concludes in Section 3.8 with a return to the thesis roadmap.

3.1 PROBLEM CHARACTERISTICS AND SETUP

First, it is necessary to describe the context and some assumptions in greater detail. To use VGO, it is required that one or more models be available from which to sample or observe outcomes at different design sites. This may seem trivial, but there are some

restrictions on the relationship between the models. Models at different accuracies are not models of different aspects; this thesis is not concerned with interfacing between geometric models, mechanical simulations, control models, etc. While this is in itself an important problem, it would not be meaningful to fit a surface to models of different aspects. What is needed are models with comparable inputs and outputs, some of which are more accurate (and costly) than others. The outputs must be a measure of the same *attribute*.

VGO takes this constraint even one step further. Not only must the different models map to the same attribute space, but the resulting objective function to be maximized must be a *profit* or other utility function measured in *dollars.* This is why maximization is assumed throughout the thesis, and allows us to compute an overall utility function that takes into account both the profitability of a particular artifact as determined by model observations as well as the cost of performing the analyses. Having the utility function in dollars negates the need for any unit conversions between the costs incurred during the process and the resulting profitability of the selected artifact.

Requiring the objective function to be a profit metric may seem like a significant limitation of VGO. Huang *et al* [29] deliberately circumvented this constraint by using a cost ratio discount factor in their augmented EI function instead of subtracting the cost directly. However, given the merits of a single metric utility function described in Section 2.1.1 and the nature of design decision making, it can be argued that considering other attributes instead of profit is misleading. It is difficult to argue that any other factor is a more fundamental objective than profit. For example, if one were designing a hydraulic hybrid vehicle as will be described in Chapter 6, the decision maker might

consider maximizing fuel economy while minimizing production costs. While this is not an incorrect idea necessarily, does it not simply reduce to maximizing profit? If there is no market for a particular vehicle, then there is no point in creating such a vehicle. Cost effective, strategic decisions can generally be formulated as financial decisions, merely a consequence of maximizing profits while minimizing costs. Consequently, the use of profit as a utility function is not considered a shortcoming in this thesis, but rather a way of encouraging financially sound decision making practices. That is, it is assumed that the externalities that would influence the utility of the final artifact have been quantified and internalized into a profit function.

Now that the requirements for the objective function and models have been delineated, other problem characteristics can be addressed. VGO does not require any particular smoothness or monotonicity in the design space; in fact, it is assumed that in most cases the designer will not know what the design space looks like. If it were possible to visualize the design space and know that it were monotonic, then a simple gradient-based optimizer could be employed on a reasonably accurate, not too expensive model, and good results could be achieved. Additionally, there is no restriction on the accuracy or correlation of the individual models used. While it is meaningless to use a completely inaccurate model, as long as the models are adequately characterized by their standard deviations, then VGO will automatically behave as appropriate. This model characterization will be discussed in further detail in the VoI sections of this chapter and in Chapter 4. The conceptual approach for VGO will now be explained.

3.2 CONCEPTUAL APPROACH

In this section, the overarching concepts of VGO are presented. Working from the top down, the overall concept is first presented, and then some detail about the individual elements is provided.

To provide a pictorial description of the algorithm, the approach of the VGO algorithm is illustrated in Figure 3.1. The first image represents a potential starting point for the algorithm. Starting with samples from any number of models at different accuracies, a Gaussian process-based surrogate model is fit to the data, as is depicted in the second image. The solid line represents the predictor, and the dashed line might represent the uncertainty, or mean squared error (MSE) of the predictor. Then, given the costs and accuracies of the available analysis models and the current prediction of the truth, the next design site and analysis model are selected using the Value of Information (VoI) selection criterion, as shown in the third image. The star represents what might be the sample site added after VoI was maximized. That is, the next design site and model are chosen to be the most valuable considering the cost of the analysis, its accuracy, reduction of uncertainty, and potential for improvement in the objective function.

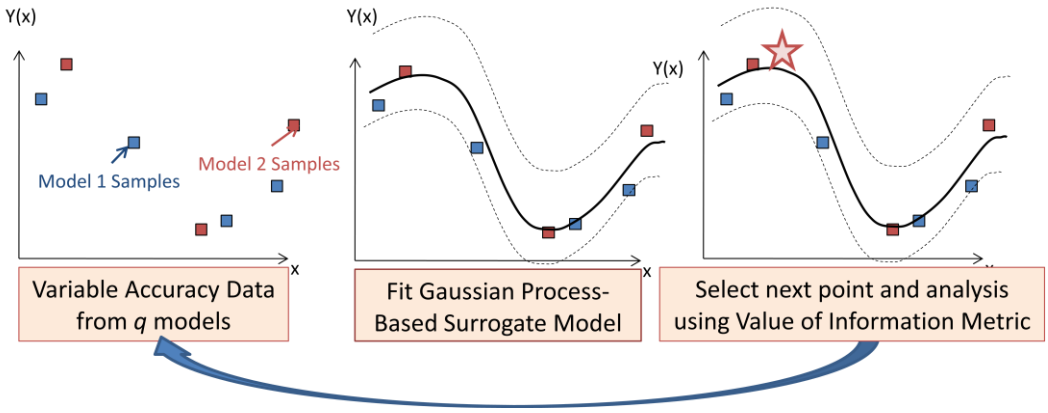

FIGURE 3.1: VGO APPROACH FOR COST EFFECTIVE OPTIMIZATION

USING MODELS OF DIFFERING ACCURACIES

After the new design site is selected, it is evaluated using the appropriate model, and added to the design space as a sample site. At this point, the process repeats until it is no longer valuable to do so, as is determined using the VoI metric.

For the sake of clarity, some terms will be discussed in more detail. When the VGO algorithm begins, it is seeded with what is known as an *initial sample*. There are many potential options for initial samples which will be discussed in greater detail in Chapter 4. The initial sample could also be called a *fixed sample* in that it is determined prior to running any optimizations or fitting the initial prediction; recall that fixed sampling strategies were discussed in Chapter 2. After the surrogate model is fit, *sequential sampling* is used for the remainder of the global optimization routine. Value of Information is the metric for sequential sampling, whereby new samples are added to the design space and the surrogate model is updated dynamically at each step in the optimization process.

It is important to mention at this point that selecting the best new design site/ analysis combination requires an explicit maximization of the VoI function over the range of the design space. The details of how this maximization is implemented are provided in Chapter 4, but it does exist as a nested optimization in every step of the overall global optimization. This is no different from maximizing Expected Improvement, or minimizing the maximum MSE; most sampling criteria require a nested optimization.

One of the most attractive features of VGO is the stopping criterion. VGO stops when there is no value in performing additional analyses to refine the solution. VGO naturally

balances local refinement and global search; there is value both in trying to refine the surrogate surface in the most promising areas as well as reducing uncertainty in areas that may have few or no sample sites. Once the cost incurred in performing additional analyses exceeds the potential gains of global search and local refinement, no further sample sites are added. At this point, a local gradient-based maximization is run on the current surrogate surface starting from the best sample site to determine the surface maximum.

3.3 PSEUDO-CODE

The overarching description and pictorial representation were provided in the previous section. Here is a pseudo-code outline of the VGO algorithm:

Value-Based Global Optimization (VGO)

```
Initialize:
  set S := GenerateLHS(numSamples)
  set Y := AnalyzeLowFidelityModel(S)
Iterate:
  while forever
    set ŷ(x):=GenerateVarAccKriging(S,Y)
    set ŷ_max:= max(ŷ(s_11),…, ŷ(s_qn_q))
    for i=1 to numModels
      set [s_max(i), maxVoI(i)]:=MaximizeVoI(ŷ(x), ŷ_max, C_i)
```

```
    end
    set [maxVoI, index]
     :=argmax(maxVoI(1),…,maxVoI(numModels))
    if maxVoI<0
      break while loop
    end if
    set S:=S ∪{s_max(index)}
    set Y:= Y ∪{AnalyzeModel(s_max(index)),index)}
  end while
Terminate:
  set globalMax:=Maximize(ŷ(x))
```

Note that the initial sample does not have to be LHS (Latin Hypercube Sampling), nor do the initial samples necessarily need to be evaluated using the lowest fidelity model; however, in the examples presented later in this thesis, this is how the algorithm was seeded. The decision to use LHS is very common in surrogate modeling; it is a way to ensure reasonably global coverage without densely sampling. As far as using the lowest fidelity model, this is in line with the VoI concept: use the less expensive models for global exploration and save the most accurate ones for local refinement. Experimental results will be shown and discussed in more detail in Chapter 5, but as long as the low fidelity model can provide some information about where good solutions might be found

and is characterized appropriately, then VGO is not overly sensitive to the fixed sampling strategy.

The VGO algorithm, as implemented, naturally trades off global search and local refinement. The variance of the simulation prediction drives up value proportionally to uncertainty, which results in global search. On the other hand, the neighborhood of good solutions is an attractive search area due to the high predicted mean. In addition, VGO tends to use the high fidelity model only in the local refinement stages. That is, only when a region seems promising and it is cost effective does the algorithm tap into the more expensive resources. Finally, the VGO algorithm has an intuitive stopping criterion that does not require user input; stop when the VoI is less than or equal to zero, that is, when it is no longer valuable to search further.

3.4 GAUSSIAN PROCESS MODELING FOR VARIABLE ACCURACY DATA

The previous sections focused on problem setup, problem characteristics and the overview of the VGO approach. Now that VGO has been explained at a high level, this section is focused on a conceptual explanation of the surrogate modeling technique employed in the algorithm.

In Chapter 2, surrogate modeling and particularly kriging modeling were discussed at length [9, 10, 16, 23, 41, 54, 55, 61, 67]. There is no shortage of existing surrogate modeling techniques, but there are very few that can accommodate variable accuracy data from any number of models, and none with the same assumptions as VGO. In traditional kriging modeling, zero uncertainty is assumed at all design sites, resulting in an interpolator. This is generally a desirable property for deterministic computer

experiments from a single underlying model. In VGO, this assumption is relaxed, allowing for model inadequacy, or error between the available analysis models and the true behavior of a system. This relaxed assumption allows observations from *n* analysis models of varying levels of accuracy to be considered simultaneously. The end result of is a smooth, continuous surrogate surface that tends closest to the most accurate samples in the design space. That is, the closeness with which the surface tends to a particular design site is weighted by its accuracy. To understand how this is accomplished, it is necessary to understand the assumptions made in defining the model.

Assumption 1: *The truth, or the objective function resulting from the physical system that is being simulated, can be modeled as the sum of a polynomial regression and a zero-order Gaussian process.*

This is an important distinction between the VGO surrogate modeling approach and the classic kriging approach. In kriging modeling, it is assumed that the underlying *simulation* can be modeled as the sum of a polynomial regression and a zero-order Gaussain process. By assuming that the *truth* can be modeled as such, many of the same properties of kriging modeling are preserved while necessitating an extra term to compensate for the difference between *observations* of the truth taken from simulation data and the truth itself, or a term to represent *model inadequacy*. This is illustrated in Figure 3.2. In classic kriging, the resulting surrogate would interpolate the model observations exactly, negating the need for an additional term in the model to compensate for the difference between the observations and the truth.

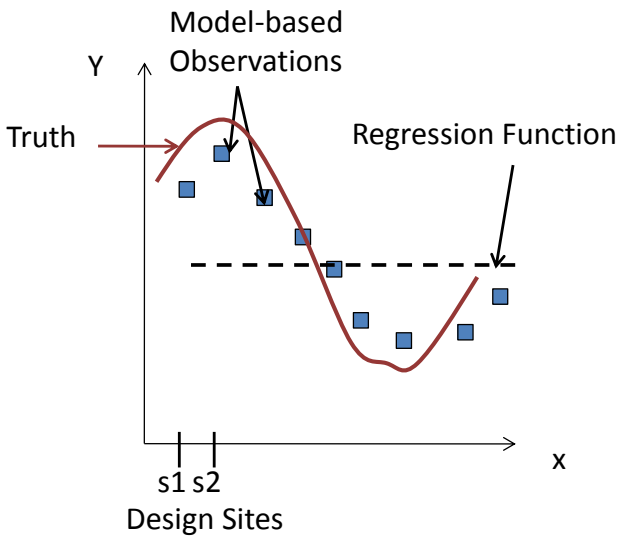

FIGURE 3.2: ASSUMED MODEL OF THE TRUTH

Assumption 2: *The error between the truth and the simulation observations for a given simulation can be modeled as a zero-mean Gaussian process.*

In the first assumption, the need for an additional term to compensate for model inadequacy was established. In this assumption, it is stated that this term can be modeled as a zero-mean Gaussian process. This model inadequacy term is quite different from the model error term used in stochastic kriging, as presented in [5]. In stochastic kriging, the error is assumed to be due to the stochastic nature of discrete event simulations. Because these simulations are nondeterministic, a white noise error term is added to the regression and Gaussian process terms. In contrast, in this thesis, we consider model inadequacy, which is an error term that is deterministic but unknown. This error is therefore reflected in the model as a distinct correlated Gaussian process. That is, if an observation is greater than the truth, then it is assumed that in some neighborhood of that observation, other observations will also be greater than the truth. The size of that neighborhood is captured using the *correlation strength.* The correlation strength need not be specified;

this can be determined mathematically using Maximum Likelihood Estimation (MLE), which will be explained and justified in the mathematical derivation presented in Chapter 4. For now, the important assumption is that for each model, the difference between the truth and observations from a particular model can be represented as a correlated Gaussian process.

Some possible realizations of model error are shown in Figure 3.3. In this figure, Model 1 might represent a low fidelity model; its error has a relatively high variance but is very smooth. This type of phenomenon might occur when certain physics are abstracted from a model, for example. Model 2, on the other hand, is much more accurate in that the variance is small, but the error is much rougher, indicating that the model might have some higher frequency content or uncertainty associated with it. These are just possible correlated Gaussian process realizations which might represent model error; the requirement is that models classified as more accurate exhibit less variance with respect to the truth than their less accurate counterparts.

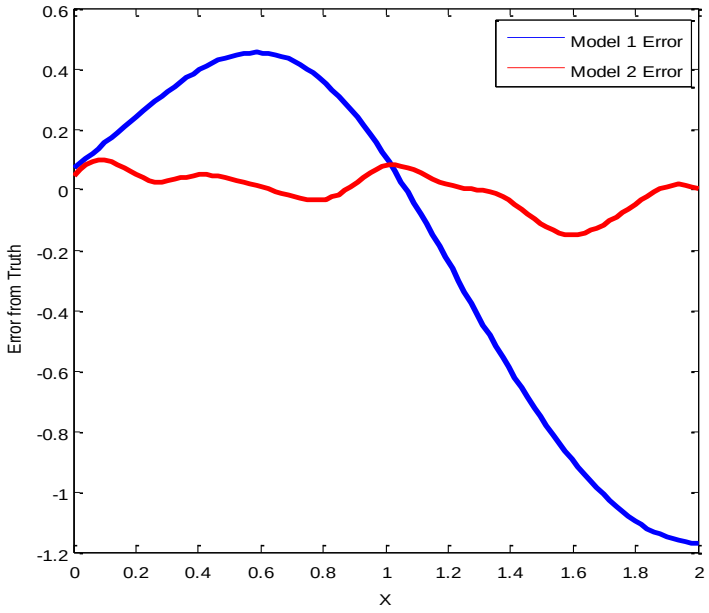

FIGURE 3.3: POSSIBLE GAUSSIAN PROCESS REALIZATIONS FOR THE ERROR BETWEEN MODELS AND THE TRUTH

Assumption 3: *The error between a given model and the truth is assumed to be uncorrelated to the error between any other model and the truth.*

In the previous assumption, it was explained that the error between the truth and observations from a particular model are assumed to be represented by a correlated Gaussian process. In this assumption, it is established that while the individual model error models are correlated within themselves, they are not correlated with each other. Referring back to Figure 3.3, the error of Model 1 and the error of Model 2 do not exhibit any explicit correlation with each other. In practice, this means that if a Model 1 observation is above the mean (the truth), this bears no indication on whether or not an observation of Model 2 at or near the same design site will also be above the mean.

While it is assumed that the individual models bear some correlation with the truth function and therefore probably exhibit some of the same behaviors, it is not necessary to assume that the errors of the individual models are correlated with each other. This simplifies the math in the derivation significantly, but it is also logical. If the models are constructed with different physics, with different grid sizes, and/or with different resolutions, the resulting errors will likely have different roughness and different standard deviations.

By adopting the above assumptions, kriging modeling can be tailored to accommodate sample sites from any number of models at different accuracies. Classic kriging models are a combination of a regression model and a zero-mean Gaussian process that captures the error between the regression model and the observations. The VGO modeling technique is a combination of a regression model, a zero-mean Gaussian process that captures the error between the regression model and the observations, and an additional zero-mean Gaussian process realization that captures the difference between the observations and the truth. This will be demonstrated and explained mathematically in Chapter 4, but the resulting model is very similar to the classic kriging model; only one matrix has a different construction. Consequently, many of the desirable properties of kriging models are retained while mainly relaxing the interpolation constraint.

One such desirable property of kriging models is that the hyperparameters can be determined without user input using Maximum Likelihood Estimation (MLE). MLE is used to determine the Gaussian process that is statistically most likely to yield the given realization. This is what makes the statistical basis of kriging and the Gaussian process modeling assumptions so attractive. Using a straightforward maximization of the

likelihood function, which is basically a multivariate Gaussian probability density function, the hyperparameters for the kriging model can be determined automatically without any tuning from the user.

The other attractive property of kriging models that is leveraged in the VGO approach is that kriging models explicitly model the uncertainty in their predictions in terms of a mean squared error (MSE). The MSE provides a statistical means for bounding the accuracy of the kriging model at any point in the design space. A large MSE would indicate high uncertainty. In classic kriging formulations, the MSE at the design sites is identically zero. In the modified VGO approach, the MSE at the design sites is equal to the variance of the model from which the samples were drawn.

In summary, VGO uses a modified kriging modeling technique for fitting a surrogate model to sample sites from different models at different accuracies. While VGO has unique assumptions, the use of the statistical basis of the kriging model allows for many of the advantageous properties of kriging models to be maintained. In VGO, any model can be sampled at any design site irrespective of prior function evaluations; this allows us to select the most valuable analysis at any particular iteration. Additionally, it is assumed that the different model inadequacies are not correlated with each other, which keeps the number of hyperparameters to a minimum. The result is a continuous, smooth surrogate surface which is weighted by the accuracy of the sample sites to which the model is fit. The mathematical details and further explanation will be provided in Chapter 4.

3.5 VALUE OF INFORMATION AS A SAMPLING STRATEGY

The second component of VGO warranting further detail is the use of Value of Information (VoI) as a sequential sampling strategy. Up until this point, the use of sequential sampling to update the predicted surface has been justified; adding sample sites in a meaningful way can help designers to optimize expensive functions without densely sampling the entire space. There are certain regions of the design space that are beneficial to model very accurately. However, in other regions it is only important to ensure that the uncertainty is low enough that the designer can be confident that he did not miss a promising area. Sequential sampling allows for design sites to be added for either local refinement or global exploration in areas where the samples are most needed.

The need for a novel sequential sampling strategy is rooted in the belief that the utility of a particular design is not merely a function of the final artifact utility but also a function of the analysis costs incurred during the design process to arrive at that particular artifact. Therefore, what is desired is a sequential sampling metric that accounts for the analysis costs incurred when a new design site is added to the design space before a decision is made about whether or not to run the simulation. However, the costs do not tell the whole story; it is also necessary to consider the prior and posterior uncertainty (before and after analysis) to determine whether a particular analysis is *valuable* at a given design site. To do this, the Value of Information metric is presented. While the VoI metric is an established metric from decision theory, it has not previously been used as a sampling criterion for global optimization algorithms. The idea of using VoI, or specifically, maximizing VoI was introduced conceptually in Chapter 1, but up until this point has not been rigorously defined.

The Value of Information metric is a measure of the expected benefit of gathering additional information prior to making a decision [27, 37]. . This metric provides a mathematically sound mechanism for trading off solution quality, potential for improvement, and the cost of performing additional analyses. By combining the VoI metric for finding the next sites for evaluation with the Gaussian process surrogate model for multiple analysis models, it is possible to navigate a design space of multi-accuracy data in a cost-effective manner.

Value of Information comes from decision theory, and to fully understand the concept of value of a particular information source it is necessary to frame the design optimization problem as a series of *decision* problems. At any point in the optimization, there is an option to stop analyzing and select the current best solution. There is also the option to perform an additional analysis at any given design site in the design space prior to making a selection. The value of information is defined as the expected difference in payout of a decision made with or without performing additional analyses. For example, assume that there exists some prior belief about the objective that we hope to maximize, specifically $\hat{y}$. Given the sample sites (information) already obtained (S, Y), there is a belief about the current best solution,

$$\hat{y}_{max} = \max_{s_{ij} \in S} \hat{y}(s_{ij}) \tag{3.1}$$

After performing some additional analysis at a point s_{new}, in the design space, a simulation outcome y_i is obtained, and the incorporation of that new piece of information

into the knowledge state results in some posterior belief, $\boldsymbol{y}|y_i(x)$, where $\boldsymbol{y}$ is a random variable representing our belief about the updated truth surface. This posterior belief might have a higher objective function value than $\hat{y}_{max}$ and it would then be selected, else $\hat{y}_{max}$ from the prior would be selected. The ex-post value of analysis y_i at any $x = s_{new}$ is then:

$$\boldsymbol{v}(x) = \max\left(\mathrm{E}\left[\boldsymbol{y}\middle|y_i(x)\right], \hat{y}_{\max}\right) - \hat{y}_{\max} \tag{3.2}$$

In the case where an analysis is performed, the equation above represents the difference between the payoff given the choice $\max(\mathrm{E}[\boldsymbol{y}|y_i(x)], \hat{y}_{\max})$ resulting after observing the simulation outcome $y_i(x)$, and the payoff that would have resulted under the prior action $\hat{y}_{\boldsymbol{max}}$, which would have certainly been selected had the simulation *y* not been performed. This value can be positive or zero; the simulation outcome $y_i(x)$ might not lead to a change in action, in which case the value of information would be zero. If an improvement in the objective function is achieved because of the additional analysis, then the value of information is positive.

Now, if the cost of obtaining the simulation outcome y is considered, this leads to the realized incremental gain:

$$G = v(x, y) - C = \max\left(\mathrm{E}\left[\boldsymbol{y}\middle|y_i(x)\right], \hat{y}_{\max}\right) - \hat{y}_{\max} - C \tag{3.3}$$

where $\mathcal{C}$ is the cost of analysis to obtain the simulation outcome $y_i(x)$. Therefore, if the cost $\mathcal{C}$ is equal to or exceeds the value of the source, then the realized gain is 0 or negative, implying that performing the additional analysis would be a waste of resources.

This is precisely the principle applied when considering whether to perform a particular analysis on a design site.

For any design site, there are multiple analyses or models available from which to gather data. The value of information is source specific; that is, the value of an analysis of model *i* at a potential design site $x = s_{new}$ is different from the value of an analysis from model *j* for the same *x*. The value of the particular information source depends on the available alternatives and the quality of the available sources. If after performing an analysis, the selected design remains unchanged, then the value of information is *zero*; the additional analysis did not result in any improvement in our objective function. If the analysis causes a change in the selected artifact and therefore an improvement in the objective function, then the value of information is *positive*.

The problem is that the value of information or realized incremental gain is only useful from the decision making perspective *prior* to running the additional analysis and yet can only truly be known *after* the simulation is performed. Thus, what is used in practice is the *expected* value of information. If the analysis model has no uncertainty associated with it, that is, the posterior variance is zero, then the expected value of perfect information can be calculated. When using models of varying accuracy, however, it is assumed that the models exhibit model inadequacy and have some positive uncertainty associated with them and thus the expected value of *imperfect* information is used. This quantity represents the difference between the expected utility with the new analysis included and the expected utility without the added analysis. To ensure that cost is accounted for, VGO actually uses the expected incremental gain to represent VoI, and the

terminology is used interchangeably since the two quantities differ only by a constant cost for a particular analysis.

It was explained previously that in VGO, VoI is used as a sequential sampling strategy. This means that at each step in the optimization process, the expected value of information or expected incremental gain for each analysis is maximized within the bounds of the design space. It does not matter mathematically speaking if the optimization is run on the VoI strictly speaking or on the incremental gain; the two differ only by a constant and therefore have the same mathematical optimum. This must be done for *each* available analysis. Then, the maximum of individual maxima for the models is determined, and this point and analysis are selected to be added to the sample sites. This continues until the maximum expected *gain* for each model is less than zero. For this stopping criteria, it makes sense to use the incremental gain calculation to factor in the cost of the analysis to ensure that the potential payoff is significant enough to outweigh the cost of analysis incurred by adding the new design site.

In terms of global optimization, this is a novel approach. The Value of Information and realized incremental gain give a mathematically sound, rational approach for determining when to stop sampling based on model cost, accuracy, and potential for improvement. Additionally, the VoI metric naturally balances local refinement with global search without any additional tuning factors from the user. Sometimes an inexpensive analysis in an area with high uncertainty provides more value than a very expensive analysis near the optimum, and vice versa. That is, a simple, inexpensive model that can reduce the uncertainty in an area with very sparse sample sites might be the most valuable analysis during a particular iteration, or a more accurate analysis may be justified to refine the

surface near other promising sample sites. VoI will also naturally screen out analysis models that are not cost effective or that are dominated by other analysis models (e.g., there exists a model of the same cost with better accuracy).

In this section, the conceptual approach to Value of Information was explained. Further details are provided in Chapter 4, where the full mathematical derivation is provided, along with specific calculations related to the individual terms.

3.6 PROPERTIES OF VGO

Now that the conceptual approach to VGO has been explained, along with more detail on the surrogate modeling and sequential sampling techniques, some of the resulting properties of this algorithm will be summarized, along with some of the primary advantages and shortcomings.

VGO is a global optimization algorithm that is aimed specifically at design problems relying heavily on the use of models and simulations. The goal of VGO is not truly mathematical optimality, but rather good solutions at reasonable cost. This is a key distinguishing factor from other global optimization algorithms; the formulation of the overall utility is a function of both the solution quality and the costs incurred in reaching that solution.

There are two main aspects to VGO: one is a surrogate modeling method that can accommodate simulation data at varying accuracies, and the other is the use of Value of Information for sequential sampling. These two contributions allow for the leveraging of all available models in the most meaningful and cost-effective way while still allowing for highly accurate solutions.

The main advantages to VGO (conceptually) are its specificity to engineering design problems, its general lack of need for user input, and its simple and logical stopping criterion. Thus far in the presentation of the algorithm, the only disadvantage is the requirement that the utility function be expressed in profit. While this may seem limiting, it can also be viewed as a guideline; it is generally a good idea to approach systems engineering and management problems with a clear understanding of the bottom line without obfuscating it with less meaningful intermediate attributes. More advantages and disadvantages will be discussed in the experimental characterizations presented in Chapter 5.

It should also be noted at this point that the size of the design problem that can be accommodated using VGO is somewhat limited. In Chapter 1, we conceptually discuss the difficulties of designing very large systems with thousands of variables; in fact, there are likely large, complex systems that could conceivably include up to a million design variables. The size of the problem that can be accommodated by VGO is approximately 10 dimensions (design variables) or less. That is, VGO can handle similarly sized problems as other similar global optimization suites. There is some mathematical reasoning for this size restriction. For kriging modeling techniques and variants thereof, there is a high risk of ill-conditioning during matrix inversion for a very large number of sample sites, which is likely to occur as the dimensionality of the problem increases.

In spite of the limitations on the number of design variables, VGO is still a useful tool for aiding in the design of large systems. In the thesis, we assume a relatively early stage of the design process. It is likely that for such large problems, the variables included at later stages of the design process will be dependent on many decisions made early in the

design process. For example, if an electric hybrid were selected over a hydraulic hybrid in the early design stages, then accumulator volume would not be a meaningful design parameter in the more detailed design phase. However, it should also be noted that, in this thesis, all of the design variables are assumed to be continuous. For a very large system, there are likely to be several discrete variables. It is therefore a good idea to pare down the number design options using simpler models of relatively small dimensionality to narrow the search space before trying to include additional variables. Considering too many variables all at once is likely to be inefficient, regardless of the choice of optimization algorithm. A common approach to limiting the number of design variables is to perform a screening test to check which variables have the highest impact on the objective function. Alternatively, a designer may work with smaller subsystem problems for the overall system. For these smaller sub-problems, VGO is likely to be a useful algorithm, particularly in the early design stages before too many detailed parameters are known. As computing power continues to increase, it will be possible to consider larger numbers of design variables, but it must still be efficient to do so before applying any optimization algorithm.

3.7 ADDRESSING THE RESEARCH QUESTIONS

At this time, it is logical to return the research questions presented in Chapter 1 and discuss the conceptual contributions presented thus far.

Recall the primary research question for this thesis:

Primary Research Question: *How can designers perform design optimizations at a reasonable cost without sacrificing solution quality?*

Hypothesis: *A Value-Based Global Optimization (VGO) algorithm will allow designers to achieve good solutions (design artifacts) at better costs than can be achieved with comparable existing algorithms.*

So far, the VGO algorithm has been presented in concept only. Some conceptual comparisons have been made with other algorithms, but more explicit comparisons will be made in Chapter 5. So far, it has been established that VGO can leverage any number of available models, takes into account the cost of analysis, and allows for selective sampling in regions of the design space that are either highly promising or highly uncertain. Assuming a proper utility function, the flexibility and cost-effectiveness of VGO make it very attractive compared to other algorithms currently available in the literature. This question will be revisited again in Chapter 7 after all of the experimental results have been presented.

Secondary Research Question 1: *How can data from multiple models of varying levels of accuracy be used advantageously during the design optimization process?*

Secondary Hypothesis 1: *A Gaussian process-based surrogate model, similar to a kriging model, can be derived to accommodate multi-accuracy observations from any number of different models.*

The first secondary research question relates directly to the surrogate modeling technique described in Section 3.4. So far, only the conceptual approach has been presented. The mathematical foundations will be derived in Chapter 4. It has been explained that, in principle, by relaxing the interpolation constraint generally imposed by classic kriging

modeling techniques, it is possible to create a surrogate surface that is weighted by the accuracy of the individual samples to which it is fit. This allows the optimization to continue while leveraging all of the available sample sites with the understanding that some of the information is more certain than other information. This also creates a situation where varying levels of accuracy of the surrogate surface can be achieved in different regions. It was established in Chapter 1 that there is no need to precisely know how bad a poor solution is, only the general direction in which better solutions can be found. On the other hand, it is necessary to model the most promising regions of the design space with high accuracy in order to discern the best solution from the space of good solutions. This variable accuracy modeling technique allows for the predictor accuracy to vary appropriately, allowing us to achieve both broad exploration and high accuracy.

Secondary Research Question 2: *How can the most valuable design site and analysis be dynamically selected at each step in the optimization process?*

Secondary Hypothesis 2: *Maximizing the Value of Information (VoI) provides a metric for choosing the next design site and associated analysis model at each step in the optimization process.*

The second secondary research question pertains directly to the VoI sequential sampling criterion presented in Section 3.5. By calculating value as function of cost, accuracy, uncertainty, and potential for improvement, VoI provides a novel means for sequential sampling and surface updating during the optimization process. VoI leverages the most cost-effective analyses and naturally balances global search and local refinement. The

VoI function complements the surrogate surface modeling technique described in the first secondary research question in that it helps to identify and refine the promising regions on the surface and reduces uncertainty in the areas that have only been sparsely sampled. The fact that it also takes into account cost means that VoI enables the creation of a cost-effective, accurate predictive surface that can be used for optimization.

3.8 THESIS ROADMAP

In this chapter, the conceptual approach for VGO was described. The problem setup and assumptions were delineated, followed by an overview of the algorithm and the associated pseudo-code. This was followed by a more detailed discussion of the Gaussian process-based surrogate modeling technique used in VGO to leverage multi-accuracy data. The relationship with classic kriging modeling was discussed along with a discussion of the new or adjusted assumptions that make this surrogate modeling technique different from the others presented in the literature. Following the surrogate modeling discussion, the basic tenets of Value of Information were described. By maximizing the expected gain, VoI provides a mathematically sound mechanism for sequential sampling by accounting for uncertainty, cost, and potential for improvement of the objective function. The overall properties, advantages, and disadvantages encountered in this conceptual realization were recapped, and the preliminary responses to the research questions were discussed. The basic foundations for addressing the research questions were laid, but the experimental results to validate the hypotheses are not presented until Chapter 5. In the next chapter, the theoretical foundations for VGO are laid. This chapter includes all of the mathematical derivations related to the Gaussian process-based surrogate model and the actual mathematics of calculating the expected

value of information. Once the algorithm theory has been derived in detail, the algorithm is characterized using experimental results in Chapter 5, and is illustrated on a practical example in Chapter 6.

CHAPTER 4: THEORETICAL FOUNDATION OF VALUE-BASED GLOBAL OPTIMIZATION

This chapter provides all of the mathematical derivations relevant to the VGO algorithm. In the previous chapter, the focus was on the conceptual approach, assumptions, and contributions with respect to the research questions. Now that the approach is understood, all of the underlying mathematics must be rigorously presented. In Section 4.1, the Gaussian process-based surrogate model for multi-accuracy data is derived. In Section 4.2, the full formulation for expected value of information and all of the calculations relevant for determining the individual terms are provided. The remaining sections pertain to initialization, intermediate optimizations, and the stopping criterion. This chapter provides a complete theoretical description of the VGO algorithm.

4.1 MATHEMATICAL FORMULATION FOR MULTI-ACCURACY GAUSSIAN PROCESS-BASED SURROGATE MODEL

One of the primary contributions of the VGO algorithm is a novel surrogate modeling technique which can accommodate seed data from any number of simulations of different accuracies. From the overview of the algorithm presented in Chapter 3, the first step after the initial fixed sample and at each step in the optimization after a new sample site is added is to fit a surrogate model to the available data. As explained previously, a novel surrogate modeling approach is required to fit to data from analysis models of varying accuracies with the assumptions delineated in Chapter 3. In this section, the mathematical derivation is presented.

The approach taken in this thesis is a modification of classic kriging modeling, and the initial steps of the derivation follow the kriging derivation. Following the seminal work of Sacks *et al.* [54], we adopt a model:

$$y(x) = f^{\mathrm{T}}(x)\beta + z(x), \qquad x \in \mathbb{R}^p \tag{4.1}$$

This model consists of a regression term, $f^{\mathrm{T}}(x)\beta$, and a zero mean Gaussian process term, $z(x)$. In classic kriging modeling approaches, this model would represent the outcome of a particular simulation, or an objective function based directly on the simulation data. However, in this approach, it is assumed that this expression is a model of a *truth* or a true objective function that cannot be observed directly. That is, while the outcomes of the computer models are deterministic, they are all merely approximations of some true objective that is in essence unknown. The difference between a simulation model and the true objective is called model inadequacy [34], which is assumed to be normally distributed:

$$y(s_{ij}) = y_{ij} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \sigma_i) \tag{4.2}$$

where y_{ij} is the simulation outcome for model i at design site s_{ij}, and ε_{ij} is the normally distributed model inadequacy for model i. It is important to note that in this approach, the ε_{ij} terms are potentially correlated with inadequacy errors for the same model at different design sites, whereas in stochastic kriging they are uncorrelated [28]. This means that for a particular model i, ε_{i1} is correlated with ε_{i2} if s_{i1} and s_{i2} are in the same

neighborhood. In stochastic kriging, on the other hand, the ε_{ij} terms are assumed to be uncorrelated resulting in a white noise term.

A surrogate model, by definition, is a function of the seed data to which it is fit. Therefore, to determine the unknown parameters for the regression, $f^{\mathrm{T}}(x)\beta$, and the Gaussian process, $z(x)$, the available seed data from the analysis models are used. Since VGO is specifically targeted toward engineering design optimization, it is assumed that the *design space* is the p dimensional space defined by the optimization variables and their ranges. It is also assumed that there exists a single, real objective function that we seek to maximize, generally for VGO this objective function will be *profit* or another utility function capturing monetary gains. Finally, it is assumed that this design space is populated by evaluating each of q unique analysis models n_q times for a total of $m = \sum n_q$ design sites. Recall that in this context a *design site* refers to a point in the design space at which one of the underlying analysis models has been evaluated and each design site has a corresponding *observation*, which is the predicted value of the objective function predicted by an analysis model.

The design sites are captured in a matrix S and the observations are captured in a vector Y_s. To simplify the mathematical derivation, it is assumed that the observation space is one-dimensional, restricted to $\mathbb{R}^1$.

$$\begin{aligned} \mathrm{S} &= \left[\underbrace{\mathrm{s}_{11},\ldots,\mathrm{s}_{1\mathrm{n}_1}},\underbrace{\mathrm{s}_{21},\ldots,\mathrm{s}_{2\mathrm{n}_2}},\ldots,\underbrace{\mathrm{s}_{\mathrm{q}1},\ldots,\mathrm{s}_{\mathrm{qn}_\mathrm{q}}}\right],\ \mathrm{s}_{\mathrm{ij}} \in \mathbb{R}^{\mathrm{p}} \\ \mathrm{Y_s} &= \left[\overbrace{\mathrm{y}_{11},\ldots,\mathrm{y}_{1\mathrm{n}_1}}^{\text{Model 1}},\overbrace{\mathrm{y}_{21},\ldots,\mathrm{y}_{2\mathrm{n}_2}}^{\text{Model 2}},\ldots,\overbrace{\mathrm{y}_{\mathrm{q}1},\ldots,\mathrm{y}_{\mathrm{qn}_\mathrm{q}}}^{\text{Model q}}\right],\ \mathrm{y}_{\mathrm{ij}} \in \mathbb{R}^{1} \end{aligned} \tag{4.3}$$

Based on the previous discussion, we assume that the *observations* of the analysis model at the design sites can be modeled as follows:

$$Y_s = F^T\beta + \mathrm{Z} + \mathrm{Z_m} \tag{4.4}$$

Again, $F^T\beta$ is a regression term, usually a zero-order or in some cases first or second order polynomial in p dimensions. In this model, Z is a Gaussian process *realization* representing the error between the regression function and the unknown truth. The remaining term Z_m is also Gaussian process realization, but this time representing the *model inadequacy* of the individual analysis models. That is, it is assumed that the observation sites Y_s can be modeled as the sum of a regression model and two Gaussian process realizations, one representing the difference between the regression model and the truth, and the other capturing the difference between the truth and the model observations.

Z_m is the only term in this model formulation that is different from the traditional formulation presented in [54]. For a design site s_{ij}, Z_m represents the model inadequacy error for model i. We assume that the model inadequacy error at site s_{ij} is correlated with the model inadequacy error at site s_{ik} for the same model, i, but uncorrelated with the model inadequacy error for any other model, l. It is also assumed that the model inadequacy errors, Z_m, are uncorrelated with the error, Z, in the regression model. These assumptions about correlation are increasingly important as the derivation progresses.

Now that the assumed models have been established, what is needed is a way to predict the objective function values for any point in the design space. A *predictor* is the desired

outcome of the surrogate modeling strategy, a way of inexpensively determining an estimate on the objective for any value in the design space given the data that has already been acquired.

Following the traditional kriging derivation, we now consider the linear predictor:

$$\hat{y}(x) = c(x)^{\mathrm{T}} Y_s, \text{ with } c(x) \in \mathbb{R}^m. \tag{4.5}$$

The error between the surrogate model and reality is then the difference between the linear predictor, $\hat{y}(x)$, and the truth, $y(x)$, from equation 4.1. Recall that the truth is assumed to be the sum of a regression term $f^{\mathrm{T}}\beta$ and a Gaussian process *z*. Consequently, the error can be described as:

$$\hat{y}(x) - y(x) = c(x)^{\mathrm{T}} Y_s - y(x) \tag{4.6}$$

From 4.4 and 4.1:

$$\hat{y}(x) - y(x) = c(x)^{\mathrm{T}}(F^T\beta + \mathrm{Z} + \mathrm{Z_m}) - (f(x)^{\mathrm{T}}\beta + z(x)) \tag{4.7}$$

To keep the predictor unbiased we impose the constraint:

$$F^{\mathrm{T}} c(x) - f(x) = 0 \tag{4.8}$$

This means that the difference in the expected value between the predictor and the assumed model for the simulation observations is zero. Because all of the Gaussian processes used in the derivation are zero-mean by definition, this constraint need only be imposed on the regression terms.

Using Equation 4.8, the error now becomes:

$$\hat{y}(x) - y(x) = c(x)^T(Z + Z_m) - z(x) \tag{4.9}$$

Up until this point in the derivation, we have been looking strictly at error. The goal is to minimize this error, but in order simplify the expression, the *mean squared error* (MSE) of the predictor is introduced, which is the expected value of the error, squared.

$$\varphi(x) = \mathrm{E}\left[\left(\hat{y}(x) - y(x)\right)^2\right] \tag{4.10}$$

$$= \mathrm{E}[c^T Z Z^T c + 2c^T Z Z_m^T c + c^T Z_m Z_m^T c - 2c^T Z z - 2c^T Z_m z + z^2] \tag{4.11}$$

In order to simplify this expression, the assumptions made about the nature of the Gaussian processes described in Chapter 3 must be drawn upon. These assumptions aid in determining the covariance structure for each term.

First Term Covariance:

$$\mathrm{E}[ZZ^T] = \sigma^2 R \tag{4.12}$$

This term is defined as in [54]. For *Z*, the covariance is assumed to be $\sigma^2 R$ where the correlation *R* is defined as

$$R_{ij} = \mathcal{R}(\theta, s_{ij}, s_{kl}) \tag{4.13}$$

representing the correlation between the Gaussian process realizations at the design sites. Recall that, depending on the design sites' relative proximity, it is assumed that there exists some correlation between their observation values. In this notation, θ is the

roughness parameter characterizing the width of the correlation kernel $\mathcal{R}$. As θ increases, the neighborhood for one design site to be correlated with another gets smaller, and the surface gets rougher. As θ decreases, the neighborhood gets larger, resulting in a smoother surface.

There are many candidate correlation kernels which may be appropriate depending on the nature of the function being approximated. Some options include linear, exponential, and Gaussian. One of the most commonly used correlation kernels and the one used throughout this thesis is the Gaussian correlation kernel for stationary correlations. For two vectors x and x' in p dimensions, this correlation structure is defined as:

$$\mathcal{R}(\theta, x, x') = \prod_{j=1}^{p} \exp(-\theta_j \left| x_j - x_j' \right|^2) \tag{4.14}$$

This correlation structure looks like a simple, normal distribution centered at a particular design site. The correlation is identically one at the design site—a design site is always perfectly correlated with itself. The drop off in correlation strength is smooth and continuous, asymptotically approaching zero as the distance between the design sites increases. The standard deviation of the distribution increases as θ decreases, and vice versa. The result is a smooth, continuous surface.

Second Term Covariance:

$$\mathrm{E}[Z Z_m^T] = 0 \tag{4.15}$$

It was assumed that the Gaussian process accounting for the difference between the regression model and the truth, *Z*, is modeled separately from the Gaussian process accounting for the difference between the truth and the observations of a particular model, Z_m. Because the model inadequacy errors, Z_m, are modeled as independent Gaussian process realizations, they are uncorrelated with *Z*, and the resulting covariance is 0.

Third Term Covariance:

$$\mathrm{E}[Z_m Z_m^T] = \sum_{i=1}^{q} \sigma_i^2 R_i \tag{4.16}$$

This term represents the covariance structure for the model inadequacy errors, Z_m. Because there exists a Z_m for each analysis model, two different correlations must be considered: the correlation between observations from different analysis models and the correlations between samples from the same model. It is assumed that the model inadequacy errors, Z_m, are correlated for design sites of the *same* analysis model but uncorrelated for design sites of different analysis models. This modeling assumption is different from previous multi-accuracy kriging modeling approaches [29, 35]. This assumption is based on the belief that the model inadequacy between one model and the truth is completely independent of the inadequacy between any other model and the truth. The models may be based on different underlying physics, different assumptions, different meshes, or different abstractions; for this reason, the nature of the model inadequacy for a given model cannot be assumed to be correlated with another analysis

model. Note that this not imply that the simulation results, y_{ij}, themselves are uncorrelated, only that their error with respect to the truth is.

The covariance structure for an individual model i for correlations between samples from the same model *i* is $\sigma_i^2 R_i$. Here the correlation structure R_i is defined as in Equation 4.14, but including only the design sites from model i and with a separate *p*-dimensional vector θ_i of roughness parameters.

While the σ^2 parameter for the covariance for Z is assumed to be unknown (and is calculated using Maximum Likelihood Estimation), σ_i^2 is assumed to be a known, user-supplied parameter characterizing the accuracy of each individual model i; there is one, 1-dimensional σ_i^2 for each of the q analysis models available. A very inaccurate model has a large variance with respect to the truth, while a more accurate model has a smaller variance. For a new prediction at a site x which has not been previously sampled, the correlation of every design site for each of the individual analysis models must be taken into account. Therefore, the final covariance structure for Z_m is a block diagonal matrix:

$$\mathrm{E}[Z_m Z_m^T] = \begin{bmatrix} \sigma_1^2 R_1 & [0] & [0] & [0] \\ [0] & \sigma_2^2 R_2 & [0] & \vdots \\ [0] & [0] & \ddots & [0] \\ [0] & \dots & [0] & \sigma_q^2 R_q \end{bmatrix} \tag{4.17}$$

$$= \sum_{j=1}^{q} \sigma_i^2 [R_i]_{m \times m}$$

$$=\sigma_1^2\begin{bmatrix} R_1 & [0] & [0] & [0] \\ [0] & [0] & [0] & \vdots \\ [0] & [0] & \ddots & [0] \\ [0] & \dots & [0] & [0] \end{bmatrix} + \sigma_2^2\begin{bmatrix} [0] & [0] & [0] & [0] \\ [0] & R_2 & [0] & \vdots \\ [0] & [0] & \ddots & [0] \\ [0] & \dots & [0] & [0] \end{bmatrix} + \dots + \sigma_q^2\begin{bmatrix} [0] & [0] & [0] & [0] \\ [0] & [0] & [0] & \vdots \\ [0] & [0] & \ddots & [0] \\ [0] & \dots & [0] & R_q \end{bmatrix}$$

Fourth Term Covariance:

$$\mathrm{E}[Zz] = \sigma^2 r(x) \tag{4.18}$$

Here, z is defined as the Gaussian process describing the error between the regression model and the truth, and Z is the Gaussian process *realization* representing this same error with respect to the observations, Y_s. The correlation structure between Z and z, is then as defined above as in [54] with:

$$r(x) = [\, R(\theta, s_{11}, x), R(\theta, s_{12}, x), \dots, R\left(\theta, s_{qn_q}, x\right)]^T \tag{4.19}$$

as the correlation between a *prediction* site x and the design sites The prediction site is an arbitrary location in the design space where we seek to predict the value of the objective function, while s_{ij} is a design site where an observation of an analysis model already exists. The function $r(x)$ describes the correlation between the prediction site *x* and all of the previously sampled design sites in the design space as a function of the distance between *x* and all of the available design sites.

Fifth Term Covariance:

$$\mathrm{E}[Z_m z] = 0 \tag{4.20}$$

Z_m, the model inadequacy Gaussian process realization, is assumed to be uncorrelated with *z*, the Gaussian process representing the difference between the regression model and the truth. Just as it is assumed that there is no correlation between the realization capturing the difference between the regression and the truth and the realization capturing model inadequacy, there is also assumed to be no correlation between the model inadequacy realization and the actual Gaussian process assumed to model the difference between the true regression and truth captured in the underlying model defined in Equation 4.1.

Sixth Term Covariance:

$$\mathrm{E}[z^2] = \sigma^2 \tag{4.21}$$

This is by definition a property of the Gaussian process *z*, and is the same as in the derivation of classic kriging in [54]. This term completes the six terms in the MSE expression given in Equation 4.11.

Combining all these terms, the MSE can now be rewritten as:

$$\varphi(x) = \sigma^2 c^{\mathrm{T}} R c + c^{\mathrm{T}} \left(\sum_{i=1}^{q} \sigma_i^2 R_i \right) c - 2\sigma^2 c^{\mathrm{T}} r(x) + \sigma^2 \tag{4.22}$$

$$= \sigma^2 (1 + c^{\mathrm{T}} R c - 2c^{\mathrm{T}} r) + c^{\mathrm{T}} \left(\sum_{i=1}^{q} \sigma_i^2 R_i \right) c \tag{4.23}$$

At this point the remaining unknown parameters must be determined. These are determined in two steps: first, the MSE is minimized, and second, the likelihood is

maximized. The MSE, as explained previously, represents the uncertainty of the error between the prediction and the unknown truth. Clearly, it is desirable for the discrepancy between the two to be minimized in order for the predictor to be as accurate as possible. In the classic kriging formulation, the MSE at the design sites is identically zero because the underlying simulation is assumed to be perfectly accurate. In this formulation, the MSE at the design sites is σ_i^2, capturing the model inadequacy of a particular observation. The MSE increases with distance from the design sites; since there is little to no correlation with the other observations when far from any other design sites, the uncertainty about the accuracy of the model is the highest. Overall, it is desirable to minimize this potential error.

After the MSE is minimized, the maximum likelihood estimation (MLE) is used to determine the Gaussian process which is *most likely* to yield a particular realization. This is one of the primary advantages of kriging modeling. Using MLE, it is possible to determine all of the unknown model parameters automatically while leveraging the statistical nature of the underlying model. Given the sample sites, MLE aids in the determination of the underlying Gaussian process that would most likely yield that particular set of samples. Specifically, it helps in the determination of the roughness and variance parameters that define the predictor.

First, the MSE must be minimized. Recall from Equation 4.3 that

$$Y_s = \left[\overbrace{y_{11}, \cdots y_{1n_1}}^{\text{Model 1}}, \overbrace{y_{21}, \cdots y_{2n_2}}^{\text{Model 2}}, \cdots \overbrace{y_{q1}, \cdots y_{qn_q}}^{\text{Model q}}\right] \; y_{ij} \in \mathbb{R}^1$$

To solve for c, we begin by minimizing the MSE with respect to c,

$$\min_{c} \varphi(x) = \sigma^2(1 + c^{\mathrm{T}}Rc - 2c^{\mathrm{T}}r) + c^{\mathrm{T}}\left(\sum_{i=1}^{q} \sigma_i^2 R_i\right)c \tag{4.24}$$

Subject to the same unbiased constraint previously imposed in Equation 4.8

$$F^{\mathrm{T}}c - f = 0 \tag{4.25}$$

The details of this minimization can be found in [54], with the difference that the covariance matrix now includes an additional term corresponding to the covariance of the model inadequacy:

$$K = \sigma^2 R + \sum_{i=1}^{q} \sigma_i^2 R_i \tag{4.26}$$

In the original kriging formulation, the covariance K would simply be equal to $\sigma^2 R$ because the model is assumed to be perfectly accurate. By introducing model inadequacy and in order to use models at different accuracies, the second term is introduced, as is defined in Equation 4.17.

From the first order necessary conditions for optimality, minimizing MSE subject to the unbiased constraint, we get:

$$\begin{bmatrix} K & F \\ F^{\mathrm{T}} & 0 \end{bmatrix}\begin{bmatrix} c \\ -\lambda/2 \end{bmatrix} = \begin{bmatrix} \sigma^2 r \\ f \end{bmatrix} \tag{4.27}$$

with solution:

$$c(x) = K^{-1}\left(\sigma^2 r(x) + F\frac{\lambda(x)}{2}\right) \tag{4.28}$$

$$\lambda(x) = -2(F^{\mathrm{T}}K^{-1}F)^{-1}(F^{\mathrm{T}}K^{-1}\sigma^2 r(x) - f(x)) \tag{4.29}$$

Substituting 4.28 and 4.29 into Equation 4.5, and knowing that K and K^{-1} are symmetric, the prediction, $\hat{y}$, is then:

$$\hat{y}(x) = c(x)^{\mathrm{T}}Y_s = \left(\sigma^2 r(x) + F\frac{\lambda(x)}{2}\right)^{\mathrm{T}} K^{-1}Y_s \tag{4.30}$$

$$= \sigma^2 r(x)^{\mathrm{T}}K^{-1}Y_s - (F^{\mathrm{T}}K^{-1}\sigma^2 r(x) - f(x))^{\mathrm{T}}(F^{\mathrm{T}}K^{-1}F)^{-1}F^{\mathrm{T}}K^{-1}Y_s \tag{4.31}$$

To simplify this expression, a generalized Least Squares fit is used. To implement this, the regression problem $F\beta \cong Y_s$ is solved, where Y_s is modeled the realization of a stochastic process.

The generalized least squares fit becomes

$$(F^T K^{-1}F)\beta^* = F^T K^{-1}Y_s \tag{4.32}$$

$$\beta^* = (F^T K^{-1}F)^{-1}F^T K^{-1}Y_s \tag{4.33}$$

and the predictor can be rewritten as

$$\hat{y}(x) = \sigma^2 r(x)^T K^{-1}Y_s - (F^T K^{-1}\sigma^2 r(x) - f(x))^T\beta^* \tag{4.34}$$

$$= f(x)^T\beta^* + r(x)^T\gamma^* \tag{4.35}$$

where

$$\gamma^* = \sigma^2 K^{-1}\,(Y_s - F\beta^*). \tag{4.36}$$

This formulation of the predictor is nearly identical to the classic kriging predictor, only γ^* is instead a function of the expanded covariance matrix K.

The MSE then becomes

$$\varphi(x) = \sigma^2 + u(x)^{\mathrm{T}}(F^{\mathrm{T}}K^{-1}F)^{-1}u(x) - \sigma^2 r(x)^{\mathrm{T}}K^{-1}\,(\sigma^2 r(x)) \tag{4.37}$$

where $u = F^{\mathrm{T}}K^{-1}\,(\sigma^2 r(x)) - f(x)$ and $\sigma^2, \theta,$ and θ_i are found from maximum likelihood estimation.

To maximize the likelihood, the most probable Gaussian probability density function is selected:

$$\max_{\sigma,\theta,\theta_i} L = \max_{\sigma,\theta,\theta_i}\left(\frac{1}{(2\pi)^{m/2}|K|^{1/2}}\, e^{-\frac{1}{2}(Y_s - F\beta^*)^{\mathrm{T}}K^{-1}\,(Y_s - F\beta^*)}\right). \tag{4.38}$$

To make the problem more computationally tractable, the natural logarithm of L is maximized,

$$\max_{\sigma,\theta,\theta_i} \ln(L) = \max_{\sigma,\theta,\theta_i}\left(-\ln\left(|K|^{1/2}\right) - \frac{1}{2}(Y_s - F\beta^*)^{\mathrm{T}}K^{-1}\,(Y_s - F\beta^*)\right). \tag{4.39}$$

Since the individual σ_i are assumed to be *known* as a metric of model accuracy, L is maximized over the remaining unknown parameters. Therefore, we need

$$\frac{\partial L}{\partial \sigma} = 0, \frac{\partial L}{\partial \theta} = 0, \frac{\partial L}{\partial \theta_i} = 0, \quad \text{for } i = 1, \dots, q \tag{4.40}$$

.

Because of the complexity of K it is not possible to find a closed form expression for any of the partial derivatives. This is different from the derivation presented in [54], where it is possible to develop an analytical expression for σ^2 and the optimization reduces to one dimension. In this case, the maximization must be done numerically over all unknown parameters.

4.2 VALUE OF INFORMATION IMPLEMENTATION

In the previous section, the theoretical foundations for the Gaussian process surrogate modeling component of VGO were presented. In this section, the VoI theory is developed. This section builds on the conceptual approach provided in Section 3.5.

VoI is most understandable when it is applied at the individual decision level. To set the context, assume that at any step in the optimization process or for each step of the VGO algorithm, there exists a choice to add an additional analysis and subsequently sample site at a point x in the design space. We want to assess how much value there is in performing a given analysis at that point. This metric is targeted specifically toward optimization, unlike many other sequential sampling strategies. The value is added when the result is some improvement in the objective function, that is, when the particular artifact that would be selected *changes* from what would have been selected prior to the analysis because the new artifact is more profitable.

To determine in advance what the expected value of an analysis is (even though the true value cannot be known until *after* analysis), a utility function is constructed over the range of possible simulation outcomes. These possible outcomes are compared to the current best available; the current best is defined as the artifact that would be selected if

the optimization process were to stop prior to any additional analyses. The expected value of information from analyzing a point *x* is defined as

$$\text{VoI}(x)=\int_{-\infty}^{\infty} f_{\tilde{y}}(\eta)\max(\text{E}[u(\boldsymbol{y}|\eta - C)],\text{E}[u(\boldsymbol{y}(s_{max}) - C)])d\eta - \text{E}\left[u\left(\boldsymbol{y}(s_{max})\right)\right] \quad (4.41)$$

where *x* is the point where the analysis is to be performed and $y_i(x)$ is its associated observation or simulation outcome. In this equation, η is a dummy variable over which the integration is performed. The quantity $f_{\tilde{y}}(\eta)$ represents the probability distribution of possible outcomes and is computed as a Bayesian update of the prior beliefs about the truth at point *x* and the posterior of the simulation results and is defined in Equation 4.45.

$\text{E}[u(\boldsymbol{y}(s_{max}) - C)]$ is the expected utility of choosing the current best design site, $\hat{y}_{\max}$. This quantity is fixed irrespective of the simulation outcome, $y_i(x)$, and is the basis for comparison between the alternatives. For the risk neutral case,

$$\text{E}[u(\boldsymbol{y}(s_{max}) - C)] = \hat{y}_{\max} - C \quad (4.42)$$

In this expression, $\hat{y}_{\max}$ is the current best solution, and C is the cost incurred by performing a particular analysis at site *x*. This formulation is in line with the assumption that utility, *u*, is defined as the difference between the artifact utility and the analysis cost incurred in achieving that utility.

$\text{E}[u(\boldsymbol{y}|\eta - C)]$, on the other hand, is the expected utility of choosing the new design site after analysis, and this quantity very much depends on the simulation outcome $y_i(x)$. Specifically, it depends on how the new predictive surface behavior after the new sample site $(s_{new}, y_i(s_{new}))$ is added to the design space.

In this expression, $\boldsymbol{y}|\eta$ represents the updated predictor at site *x* given the new observation, $y_i(x)$. Again, *C* is the cost of analyzing *x*.

Clearly only one choice will be made; if the utility of selecting the new design site is higher than that of selecting current (that is, the information from the analysis causes the designer to change her selection) then the new design site will be selected; otherwise, the current best achieved prior to the additional analysis, will be selected.

Since only one decision will be made to select the new design site or the current best (whichever has the higher utility), the integral from 4.41, or the expected utility of analyzing *x* with a particular model, can be divided into a sum of two integrals as follows:

$$E[u(\text{Analyze } x)] = \int_{-\infty}^{T^*} f_{\tilde{y}}(\eta) u(\boldsymbol{y}(s_{max}) - C) d\eta + \int_{T^*}^{\infty} f_{\tilde{y}}(\eta) u(\boldsymbol{y}|\eta - C)\, d\eta \qquad (4.43)$$

where T^* is the point at which the expected utility of selecting the new design site is equal to the expected utility of selecting the current best; that is, the information provided by analyzing *x* results in the decision maker being indifferent to selecting the new or old solutions. The remaining terms are defined as follows, assuming risk neutrality:

- **Current Best:** $\hat{y}_{\max}$

  This quantity is computed as the predicted value of the surrogate model $\hat{y}_{\max} = \max_{s_{ij} \in S} \hat{y}(s_{ij})$ at the design site with the maximum observation value. This calculation will be discussed in more detail in Section 4.2.5.

- **Cost:** C

  This is the cost of running the particular simulation. This quantity only varies with the choice of model *i* and is not affected by the prediction site *x*.

- **Prior of the truth:** $\sim \boldsymbol{N}(\hat{y}, \sqrt{\varphi})$

  This is the predicted mean and standard deviation of the surrogate model, which is prediction the value of the *truth*. The mean is equivalent to the predictor, and the standard deviation is simply the square root of the MSE. This is representative of the current fit before a new sample is potentially added.

- **Prior of the simulation:** $\sim \boldsymbol{N}(\mu_{si}, \sigma_{si})$ for each of the i analysis models.

  These quantities have not yet been introduced, but come into play when predicting the posterior on the truth in Equation 4.44. This mean and standard deviation representing the prior on the *simulation* is distinct from the prior on the truth. Our belief about the outcome of a particular simulation is only a function of the other data we have from that simulation. Similarly, the uncertainty associated with the simulation outcome is different from the uncertainty about the truth; specifically, there is no model inadequacy assumed in the model. Model inadequacy is a characterization with respect to the truth—if a model has been sampled, we know with 100% certainty what the simulation outcome is, and the uncertainty, σ_{si}, goes to zero at the sample sites.

- **Outcome of the simulation:** $\sim \boldsymbol{N}(y_i, \sigma_i)$

This is the simulation outcome (which is not known until the analysis is run) and its associated model inadequacy with respect to the truth.

These terms are used to compute the probability density function representing the distribution of possible outcomes.

To determine $f_{\tilde{y}}(\eta)$ for a particular analysis, we must ignore the observations from other analysis models and consider only the observations from the model currently under consideration. It should be noted once again that VoI must be calculated separately for each model available, and then the maximum of the maximum VoI's from each model is taken to determine the next site and analysis. Therefore, if VoI for Model 1 is being calculated:

$$f_{\tilde{y}}(\eta) = \frac{1}{\sqrt{2\pi(\sigma_{s1}^2)}} e^{-\frac{(\eta-\mu_{s1})^2}{2\sigma_{s1}^2}} \tag{4.44}$$

where μ_{s1} and σ_{s1}^2 are the predicted mean and variance at *x* given the observations from model 1 only.

4.2.1 Calculating the Prior Mean for Each Model

In the previous section, μ_{si} was introduced as a way to assess the prior mean on a particular simulation outcome at a point *x* in the design site with respect to a particular model. This is distinct from $\hat{y}$, the prediction of the truth in that μ_{si} takes into consideration only design sites from model *i*. If the prediction site *x* is very close to a design site s_{ij} from model *i*, then it is expected that the simulation outcome *y* will be very highly correlated to the observation at s_{ij}. On the other hand, if *x* is not close to any

design sites from model *i*, then there will be high uncertainty about the outcome of the simulation, and the prediction will converge to that of the predicted truth.

To derive this expression, it is assumed that

$$y_i = y + z_i \tag{4.45}$$

where y_i is the Gaussian process that perfectly models the behavior of model *i*, *y* is the truth and z_i is the Gaussian process that models the error between the two.

Additionally,

$$\hat{y}_i = \hat{y} + \hat{z}_i \tag{4.46}$$

This is the prediction of the mean given only the data from model *i* and can be expressed as the sum of the Gaussian process realizations used to predict the truth and the error between the truth and model *i*.

The mean corresponding to the prior belief about the outcome of a particular analysis model is calculated from the following predictor:

$$\mu_{si} = \hat{y}_i = \hat{y} + \hat{z}_i = \hat{y} + r_i^{\mathrm{T}} \gamma_i^* \tag{4.47}$$

where $\hat{y}$ is the prediction of the truth, $\hat{z}_i$ is the predicted error between the truth and model *i*, and

$$\gamma_i^* = R_i^{-1}\left(Y_{Si} - \hat{y}(s_i)\right) \tag{4.48}$$

Here, Y_{Si} are the observed design sites from Model *i*, $\hat{y}(s_i)$ is the prediction of the truth at the design sites from model *i*, and R_i^{-1} corresponds to the correlation matrix for design sites only from Model *i*, that is R_1 is R_i for $i = 1$.

Making the substitution that $\hat{y}(s_i)$=$c^{\mathrm{T}}(s_i)Y_{Si}$, we get

$$\hat{y}_i = \hat{y} + b_i^{\mathrm{T}} Y_{Si} \tag{4.49}$$

Where

$$b_i = r_i^{\mathrm{T}} R_i^{-1}(I - \tilde{C}) \tag{4.50}$$

Here, r_i^{T} corresponds to the distance between the prediction site *x* and the design sites from model *i only* and not the distances between *x* and all of the design sites in *S*. As the distance from *x* to the design sites from model *i* increases, the correlations go to zero. This means that, per Equation 4.48, as r_i^{T} goes to zero, the predicted μ_{si} converges to $\hat{y}$. That is, if there is insufficient about the simulation outcomes from a particular model, the best estimate we have about the model outcome is the prediction of the truth.

In equation 4.49, $\tilde{C}$ is defined as follows:

$$\tilde{C} = \begin{bmatrix} c^{\mathrm{T}}(s_{i1}) \\ \vdots \\ c^{\mathrm{T}}(s_{in_q}) \end{bmatrix} \tag{4.51}$$

Recall that *c* was defined in Equation 4.28 as

$$c(x) = K^{-1}\left(\sigma^2 r(x) + F\frac{\lambda(x)}{2}\right)$$

And from Equation 4.29,

$$\lambda(x) = -2(F^{\mathrm{T}}K^{-1}F)^{-1}(F^{\mathrm{T}}K^{-1}\sigma^2 r(x) - f(x))$$

4.2.2 Calculating the Predicted Variance for Each Model

In the previous section, a method for calculating the prior mean with respect to a particular simulation was defined. It is necessary to define the variance or MSE associated with this mean. The variance for a prediction site *x* given the observations of model *i* only is notated as σ_{si}^2, and is determined from

$$\mathrm{var}(\hat{y}_i) = \sigma_{si}^2 = \mathrm{E}[(y_i - \hat{y}_i)^2] \tag{4.52}$$

It is again assumed from Equation 4.45 that $y_i = y + z_i$ where y_i is the Gaussian process that perfectly models the behavior of model *i*, *y* is the truth and z_i is the Gaussian process that models the error between the two. Again, from Equation 4.46, $\hat{y}_i = \hat{y} + \hat{z}_i$, which is the prediction of the mean given only the data from model *i* and the Gaussian process realizations used to predict the truth and the error between the truth and model *i*. Thus,

$$\sigma_{si}^2 = \mathrm{E}[(y - \hat{y})^2 + (z_i - \hat{z}_i)^2] \tag{4.53}$$

The full derivation for this expression can be found in Appendix A.

The resulting expression for σ_{si}^2 is:

$$\sigma_{si}^2 = \sigma^2 + \sigma_i^2 + (c + b_i)^{\mathrm{T}} K (c + b_i) - 2\sigma^2 (c + b_i)^{\mathrm{T}} r - 2\sigma_i^2 (c + b_i)^{\mathrm{T}} r_i \tag{4.54}$$

In the above expression, c is defined in Equation 4.28 and b_i is defined in Equation 4.50. Similar to the prior mean derived in the previous section, in the absence of samples from model *i*, the prior variance will converge to the variance of the predicted truth.

Some sample results for mean and variance with respect to a particular model for a one-dimensional problem are captured in Figure 4.1. The predicted mean interpolates the design sites from Model 1, and the variance is identically zero at the design sites because the model is deterministic. This is different from the estimate of the truth where there exists a non-zero variance at the design sites due to the recognition of model inadequacy.

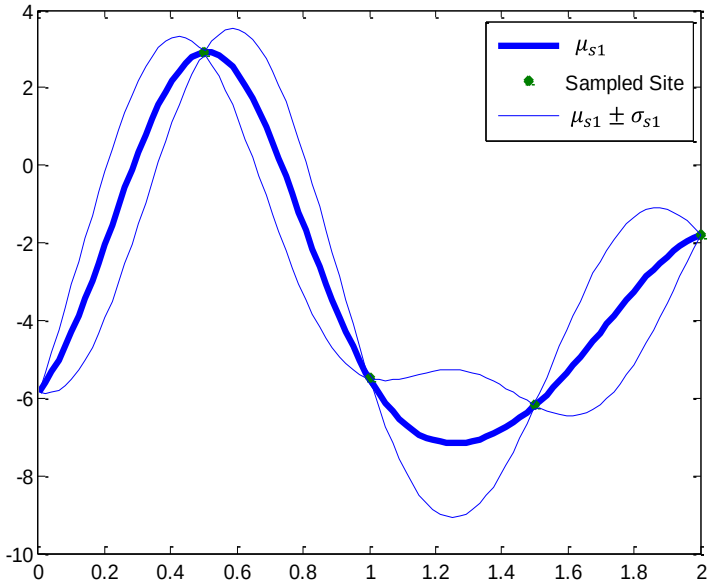

FIGURE 4.1: PRIOR MEAN AND VARIANCE WITH RESPECT TO MODEL 1

4.2.3 Calculating the Posterior on the Truth

The posterior on the estimate of the truth, $\boldsymbol{y}|\eta$, is a normal distribution needed to evaluate the VoI integral. The mean of this distribution is a new predictor $\boldsymbol{y}|\eta$ that would result once a new design site s_{m+1} and its associated observation *y* are added to the current vector of design sites, *S*, and the current vector of observations, Y_S, respectively. The currently unknown simulation outcome *y* is assumed to be normally distributed, which means that the resulting posterior for the truth estimate will also be normally distributed. The structure of the predictor $\boldsymbol{y}|\eta$ is the same as in derived in Section 4.1, but one additional *design site* is added. That is, for a particular candidate evaluation site *x*, the design site is known with certainty: $s_{m+1} = x$. Its associated observation, *y*, is not known until after the analysis is performed, but we assume that it is normally distributed. To calculate the posterior mean for the truth prediction, we use the same Gaussian process-based surrogate model presented in 4.1. The variance of the posterior of the truth estimate is then simply the MSE associated with the predictor. Thus, the resulting posterior distribution for the truth prediction is normal, with mean and variance described as follows:

$$\boldsymbol{y}|\eta \sim N(\sigma^2 \bar{r}^T \bar{K}^{-1} \bar{Y} - (\bar{F}^T \bar{K}^{-1}\, \sigma^2 \bar{r} - f)^T (\bar{F}^T \bar{K}^{-1}\, \bar{F})^{-1} \bar{F}^T \bar{K}^{-1}\, \bar{Y},$$
$$\sigma^2 + u^{\mathrm{T}} (\bar{F}^{\mathrm{T}} \bar{K}^{-1}\, \bar{F})^{-1} u - \sigma^2 \bar{r}^{\mathrm{T}} \bar{K}^{-1}\, (\sigma^2 \bar{r})) \qquad (4.55)$$

The mean is derived from Equations 4.33 and 4.34, and the variance is derived from Equation 4.37. In this expression,

$$\bar{\mathrm{Y}} = \begin{bmatrix} Y_S \\ y \end{bmatrix} \tag{4.56}$$

where $[\mathrm{Y}_S]_{m\times 1}$ is defined from Equation 4.3. $[\bar{Y}]_{(m+1)\times 1}$ is comprised of the current set of simulation outcomes with the new simulation outcome *y* appended to it. Similarly,

$$\bar{K} = \begin{bmatrix} K & \tilde{k} \\ \tilde{k}' & \tilde{k}_{m+1} \end{bmatrix} \tag{4.57}$$

Where $[K]_{m\times m}$ is the covariance matrix used to predict the truth at the current iteration of the algorithm. $[\bar{K}]_{(m+1)\times(m+1)}$ has one additional row and column appended to it containing the covariance information for the new design site, s_{m+1}. The remaining terms $[\bar{r}]_{(m+1)\times 1}$ and $[\bar{F}]_{(m+1)\times p}$ are comprised of the current fit parameters $[r]_{m\times 1}$ and $[F]_{m\times p}$, respectively, but contain an additional entry to accommodate the new design site. Finally, $[f]_{p\times 1}$ remains unchanged from the prior truth estimate to the posterior, as the dimensionality of the problem does not change.

This distribution associated with $\hat{y}|y$ on the whole represents the updated surrogate model when the new design site is added to the space. Only the mean of this quantity, however, is needed to compute VoI for the risk neutral case, as is presented in this thesis. This means we must compute

$$\boldsymbol{y}|\eta = \sigma^2 \bar{r}^T(x)\bar{K}^{-1}\bar{Y} - (\bar{F}^T\bar{K}^{-1}\,\sigma^2\bar{r}(x) - f(x))^T(\bar{F}^T\bar{K}^{-1}\,\bar{F})^{-1}\bar{F}^T\bar{K}^{-1}\,\bar{Y} \tag{4.58}$$

Normally, the predictor model is grouped by *x* so that the same form of the equation can be used for different values of *x;* that is, the same model can be used to calculate a

prediction for any *x* in the space. Here, however, we want to evaluate the model for the same *x* design site but for different values of *y,* the posterior on the simulation. The design site in *x* is fixed but the resulting model will vary depending on the observation from the analysis model at site *x,* specifically the posterior outcome *y*. Thus, the posterior on the truth given *y* can instead be represented as follows:

$$\boldsymbol{y}|\eta = a_1 Y_S + a_2 y \tag{4.59}$$

where a_1 and a_2 are constants *if* it is assumed that the MLE parameters remain unchanged with the addition of the new design site. Rather than recalculate and optimize the MLE for each evaluation of VoI, which would be computationally prohibitive, it is assumed that the addition of one new sample site will not have a drastic effect on the fit parameters. Consequently, the parameters of the current surrogate surface are used and the maximization is skipped. This assumption expedites the calculation significantly.

4.2.4 Calculating a_1 and a_2 using Block Matrix Inversion

To further expedite the calculation of $\hat{y}|y$, a_1 and a_2 can be computed as functions of the terms from the previous fit and the update terms separately. By employing a block matrix inversion technique, there is no need to calculate $\overline{K}^{-1}$ explicitly, which saves significant computation time.

In the previous section, it was mentioned that

$$\overline{K} = \begin{bmatrix} K & \tilde{k} \\ \tilde{k}' & \tilde{k}_{m+1} \end{bmatrix} \tag{4.60}$$

Where $[K]_{m\times m}$ is the already known covariance matrix. By constructing the matrix in this fashion, it is only necessary to calculate the entries for the appended row and column, rather than recalculating the entire matrix. Additionally, we can view it as a block matrix setup:

$$\overline{K} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{4.61}$$

This setup allows us to compute the inverse of $\overline{K}$ as

$$\overline{K}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix} \tag{4.62}$$

These calculations are simplified by realizing that:

$$A^{-1} = K^{-1} \tag{4.63}$$

Also, $(D - CA^{-1}B)^{-1}$ is a scalar. Everything else can be broken down into the known components from the current fit, $[r]_{m\times 1}$, $[F]_{m\times p}$, and $[f]_{p\times 1}$ and the 'update' terms appended to them. By substituting the block matrix formula from 4.62 with the proper individual terms into Equation 4.58, a_1 and a_2 can be computed with minimal additional calculations and no new matrix inversions. The full details of the derivation are given in Appendix B.

4.2.5 Computing the Current Best

In Equation 4.42, $\hat{y}_{\max}$ was introduced as the current best solution, or more accurately, the design that would be selected if no further analysis occurred. Determination of this parameter is not altogether straight forward, however.

The most likely candidate for determining the current best is to maximize the current predictor of the objective. There are two problems with this approach. One, an extra maximization is incurred. The other problem is more conceptual; in early steps of the optimization when the design space is only sparsely populated, it is possible that the predicted maximum is not in the neighborhood of one of the sample sites and is therefore an unrealized, predicted gain.

Another alternative would be to use the observation value at the most promising design site. This approach, however, does not take into account the accuracy of the prediction or the predicted value of the truth.

The approach used in this thesis is to select the value of the truth predictor at the most promising design site. That is, we select the maximum value of the truth incurred at a previously sampled site. This saves the need for an additional maximization of the surrogate model, and prevents the current best from being too speculative.

4.2.6 Final VoI Integration and Computation

Now that all of the individual terms have been defined, we return to the integral presented in Equation 4.44:

$$\mathrm{E}[u(\text{Analyze } x)] = \int_{-\infty}^{T^*} f_{\tilde{y}}(y) \cdot u(\hat{y}_{\max} - C)dy + \int_{T^*}^{\infty} f_{\tilde{y}}(y) \cdot u(\hat{y}|y - C)dy \tag{4.64}$$

It was mentioned previously that T^* represents an indifference point; the point at which the decision maker is indifferent between selecting the current best or the new design site given the simulation outcome.

T^* can be found as

$$T^* = \frac{\hat{y}_{\max} - a_1}{a_2} \tag{4.65}$$

Given the individual terms, the result of the integral for $E[U(\text{Analyze } x)]$ can be computed analytically using any available mathematical software with symbolic manipulation capabilities.

Before using the VoI calculation in context, it should be noted that

$$\text{VoI} = \mathrm{E}[u(\text{Analyze } x)] - \hat{y}_{\max} \tag{4.66}$$

We are interested in the potential for improvement over the current best. Knowing that the value of information is zero or positive, and when cost is included, the minimum value of information is $-C$ (cost). When $\mathrm{E}[u(\text{Analyze } x)]$ is calculated, the overall utility is considered, including the case where no change is made and the profit incurred is that of $\hat{y}_{\max}$. However, to determine if there is value in performing an additional analysis and therefore any potential for changing the current best design, the utility of $\hat{y}_{\max}$ must be deducted before VoI is maximized. This simple subtraction does not

change the location of the maximum, but does make the stopping criterion more intuitive. It is the subtraction of cost and this current best that allow VGO to stop when VoI is less than zero.

4.2.7 Maximizing VoI

In the context of VGO, it is not sufficient to merely compute VoI; rather, it must be maximized over the entire range of the design space. This is a multi-modal global optimization problem unto itself. While the VoI surface tends to be smooth and continuous, it is also prone to flat regions—unpromising regions where the VoI is identically the cost incurred running an analysis, because there is effectively no chance of exceeding the current best in that region.

There are multiple optimization options for this type of problem. In this thesis, the approach taken is to use a multi-start gradient optimization. Gradient optimizations are started on a Latin Hypercube grid that scales with the number of dimensions. That is, more samples are used as the dimensionality of the problem increases. In addition, optimizations are started at the vertices of the design space hypercube, and at the current best design site. Because VoI must be maximized for *each* available analysis, this is a reasonably significant number of function evaluations.

There may be more efficient approaches for maximizing VoI. In the future, it may be a good idea to derive an analytical expression for the gradient to prevent the optimizer from getting 'stuck' in flat regions, but this approach could still could cause tuning problems in the optimization. For example, a relatively small tolerance on the change in the function value would be needed for the optimizer to step through a flat region and reach a peak;

however, it is not necessary to determine the mathematical maximum very accurately, which would incur unnecessary computational expense. Another alternative is to start fewer gradient-based optimizations at better chosen starting points; for example, the peaks of VoI from the previous iteration are likely to be good starting points, as are points that lie at maximum distances from the current set of design sites. Improving the efficiency of this optimization is left for future work.

4.3 VGO INITIALIZATION

In the previous sections, the multi-accuracy surrogate modeling technique and the VoI theoretical foundations were established. In this section and the remaining sections, the remaining details of the VGO algorithm are addressed.

In order to initialize the VGO algorithm, a fixed sampling of design sites must be selected and evaluated. While a particularly dense sampling is unnecessary and in fact unadvisable, it is necessary to get some global coverage of the design space so that reasonable VoI calculations can be made. There are basically two mechanisms that can cause the VoI to be underestimated, both of which can be triggered by under-sampling at initialization: either the predicted mean of the simulation outcome is underestimated, or the uncertainty of the truth prediction (MSE) can be underestimated.

In the first case, because VoI relies heavily on the prior mean of the simulation outcome as means for assessing promising areas of the design space, having too few samples can result in understated VoI calculations in highly uncertain areas. Because of the statistical properties of the surrogate surface, when it is very sparsely sampled and the design sites are uncorrelated, the surface converges to the mean of the data as the sample sites get

further away. This results in a lot of flat, seemingly unpromising areas. While the high uncertainty in these areas does increase VoI, it also helps to have a reasonable estimation of the expected mean of the simulation outcome.

In the second case, under-sampling can effectively cause aliasing in the surrogate model with respect to the models and truth. Not only is an erroneous initial fit misleading to the optimizer, but the estimated MSE could be underestimated. If the predicted surface is (erroneously) very smooth, then the MSE is unlikely to capture the true uncertainty of the prediction if the simulation models exhibit high frequency behavior. When the MSE is underestimated, the VoI will subsequently be underestimated.

To avoid these pitfalls, it is necessary to begin with an initial sample of sufficient size. If the model behavior is known up front or if some expert knowledge is available, then the Nyquist criterion should be followed. If the model behavior is not known, it may be wise to err on the side of too many samples from the lowest fidelity model, particularly if it is inexpensive to evaluate. Another option would be to run the VGO algorithm more than once with different initial sample sizes to verify solution quality.

While beginning with a sufficient global sample will help to ensure the success of the VGO algorithm, it is also not necessary to over-sample, especially with expensive analyses. To do so would undermine the intent of using VGO in the first place; it is desirable to reserve your computational resources for valuable analyses. Therefore, it is advisable to do the initial fixed sample with a sufficient sampling of low accuracy analyses and then allow VoI maximization to determine how best to sample going forward.

In this thesis, the initialization approach taken is to use a Latin Hypercube design with a mini-max criterion. That is, an LHS design is selected that minimizes the maximum distance between two design sites. The LHS design sites are then evaluated with the lowest fidelity model available, and these sample sites are used to seed the initial surrogate model. The size of the LHS sample depends on the size and dimensionality of the design problem, and will be addressed on a case by case basis in Chapters 5 and 6. This is by no means the only acceptable initialization scheme, and some experiments in Chapter 5 show that the performance of VGO is reasonably robust with respect to the size of the initial sample.

4.4 VGO STOPPING CRITERION

As has been mentioned several times previously, one of the most attractive features of VGO is its intuitive stopping criterion. VGO stops when the maximum VoI for all models is less than zero. Specifically,

$$\max(\text{VoI}) = \max_{x}\left(E[u(\text{Analyze } x)] - \hat{y}_{max}\right) \; < 0 \tag{4.67}$$

This means that the maximum expected value added by analyzing a new design site x with the cost of analysis taken into account does not exceed the current best solution for any x.

While this stopping criterion requires no tuning parameters from the user, it does rely on the analysis cost and accuracy data provided by the user. Determination of these parameters is discussed in further detail in Section 6.2. If the models are not appropriately characterized in terms of their costs and standard deviations relative to the truth, then VGO may stop prematurely or run longer than desired. This idea of relative

accuracy and cost with respect to the truth will be discussed again in Section 5.2, but in this context, if the relative costs are too low and/or the relative accuracies are too high, then it will be valuable to allow analyses to continue for a long time. On the other hand, if the relative cost is too high and/or the relative accuracy is too low, further analysis will be less valuable very early on in the optimization.

4.5 FINAL MAXIMIZATION

After the sequential sampling process is completed and it is no longer valuable to perform additional analyses, a final optimization is run on the surrogate model of the truth. In the VGO implementation presented in this thesis, this maximization is a single start, gradient-based optimization starting from the best design site. The result of this maximization is returned as the best design artifact, along with the predicted utility, and the VGO algorithm is complete. This section completes the theoretical description of VGO. In the next section, a brief illustration of the VGO algorithm is provided.

4.6 VGO ILLUSTRATION

The VGO algorithm, including the multi-accuracy kriging model and the Value of Information metric has been implemented in MATLAB. While more rigorous performance testing is provided in the next chapters, in this section, a brief illustration of the algorithm and discussion of its behavior are provided. The algorithm illustration is a two-dimensional test problem with two models, a low and a high fidelity. The generation of the test problem will be discussed in Section 5.1; here the focus is on the outcome. For this test problem, the cost for a model 1 analysis is $2.50 and the cost for one run of model 2 is assumed to be $800. The variances for models 1 and 2 are .05 and .0008, respectively.

Figure 4.2 shows the initial seeding of the VGO algorithm this 2-D optimization problem. The individual images within the figure are as follows. The upper left shows the truth. While in practice the truth is almost never known, for the purposes of the test cases provided in this thesis, it provides a useful basis for assessing the algorithm's performance. In the upper right, the current surrogate surface fit to the current set of samples is shown. For this example, the initial surrogate model is seeded with 30 LHS design sites from the low fidelity model; the blue dots indicate the presence of low fidelity samples, which is all that is present at the initial seeding. In subsequent iterations, red dots indicate the presence of high fidelity samples. The lower left and right images in the figure capture the calculated VoI for the low and high fidelity model, respectively. For each step in the optimization, the site and analysis combination with the highest VoI is sampled in the next iteration and the surrogate surface is refit.

Figure 4.7 shows the 20^{th} iteration of the VGO algorithm for this test problem. It can be seen that high fidelity analyses have been added in the most promising regions of the design space, and that some additional low fidelity analyses have been added as well. This natural balance of global search and local refinement is one of the attractive features of using VoI for sequential sampling. The sequential sampling process continues until the maximum VoI achieved for both analysis models is less than zero.

The final iteration of the VGO algorithm for this test problem is shown in Figure 4.11. It can be seen in the current fit that both global coverage and local accuracy have been achieved. The VoI plots have both dropped below zero; they settle to their prescribed costs as the predicted surface accuracy increases and the current best is known with more confidence.

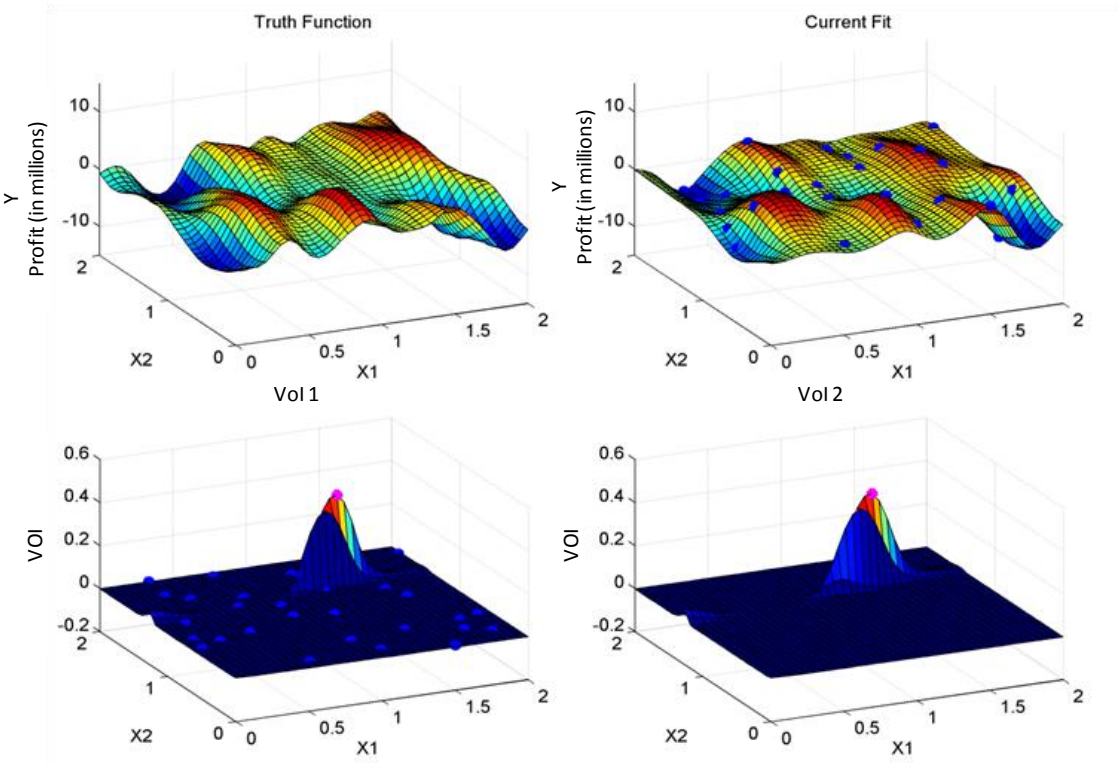

FIGURE 4.2: INITIAL SEEDING OF GAUSSIAN PROCESS MODEL

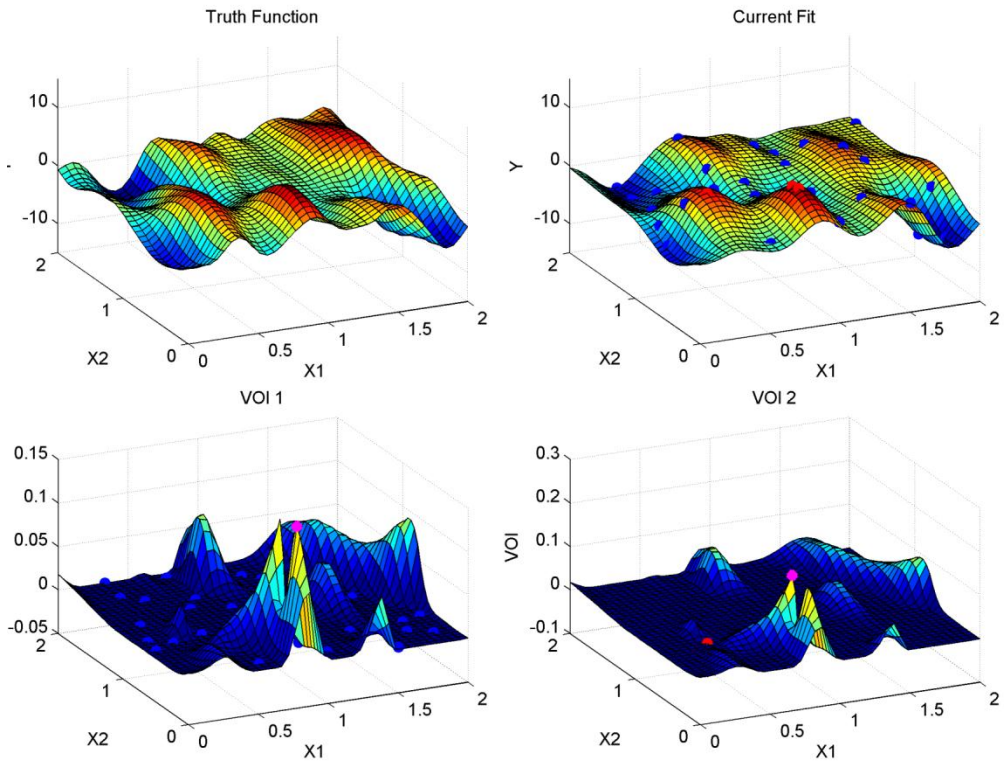

FIGURE 4.3: 4TH ITERATION OF VGO ALGORITHM

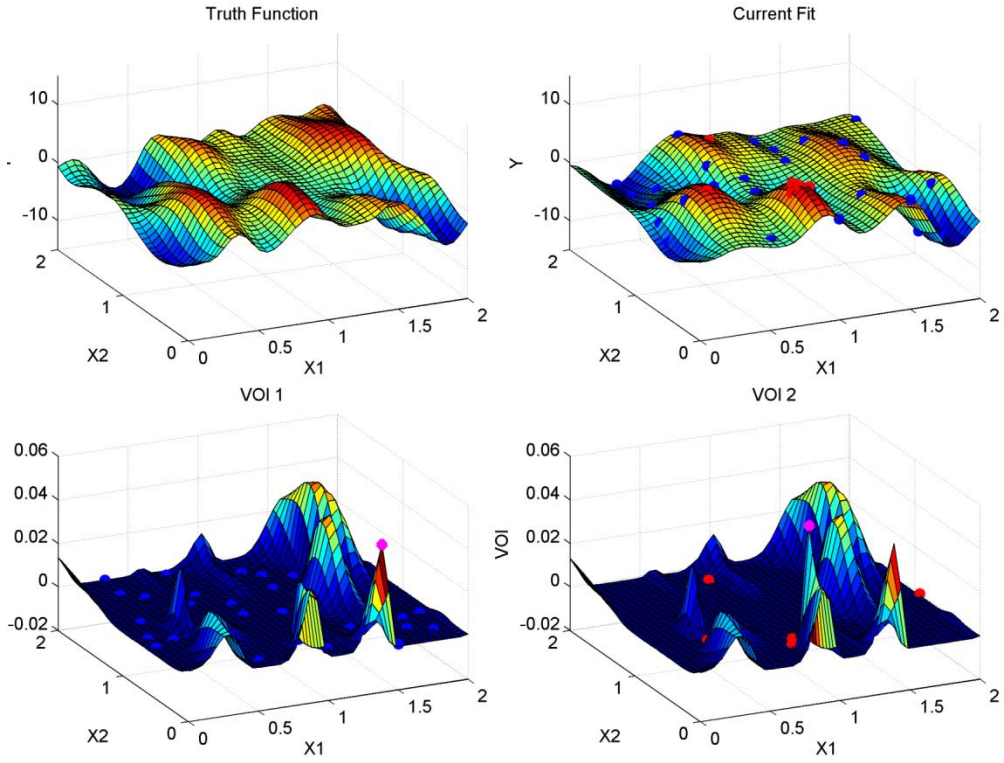

FIGURE 4.4: 8TH ITERATION OF VGO ALGORITHM

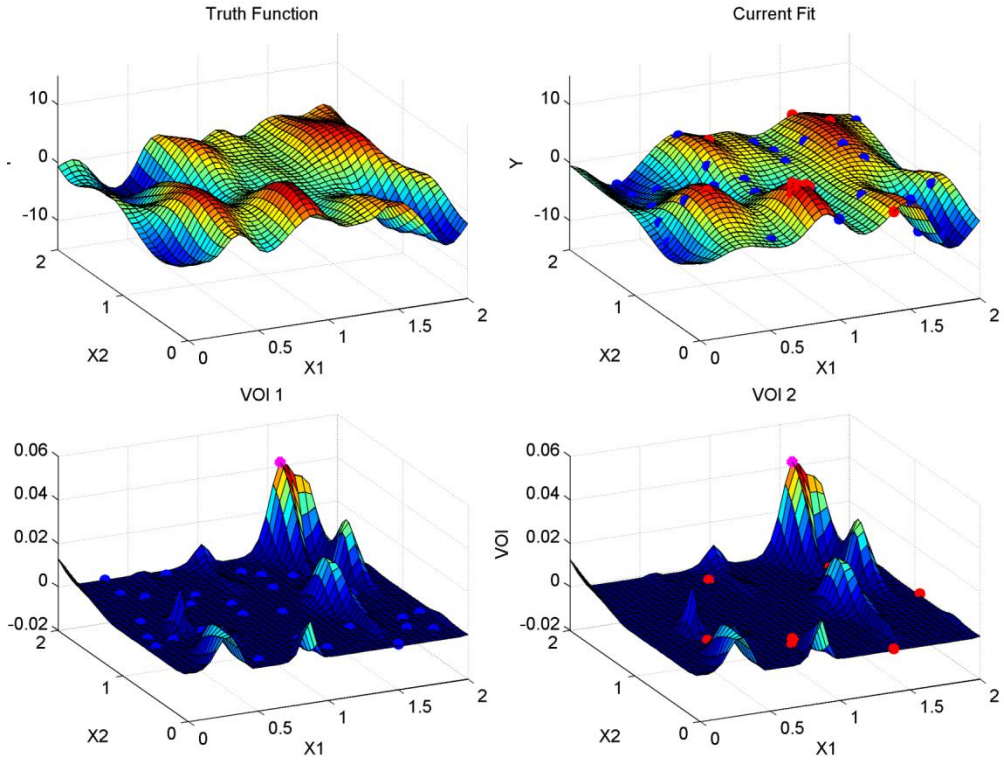

FIGURE 4.5: 12TH ITERATION OF VGO ALGORITHM

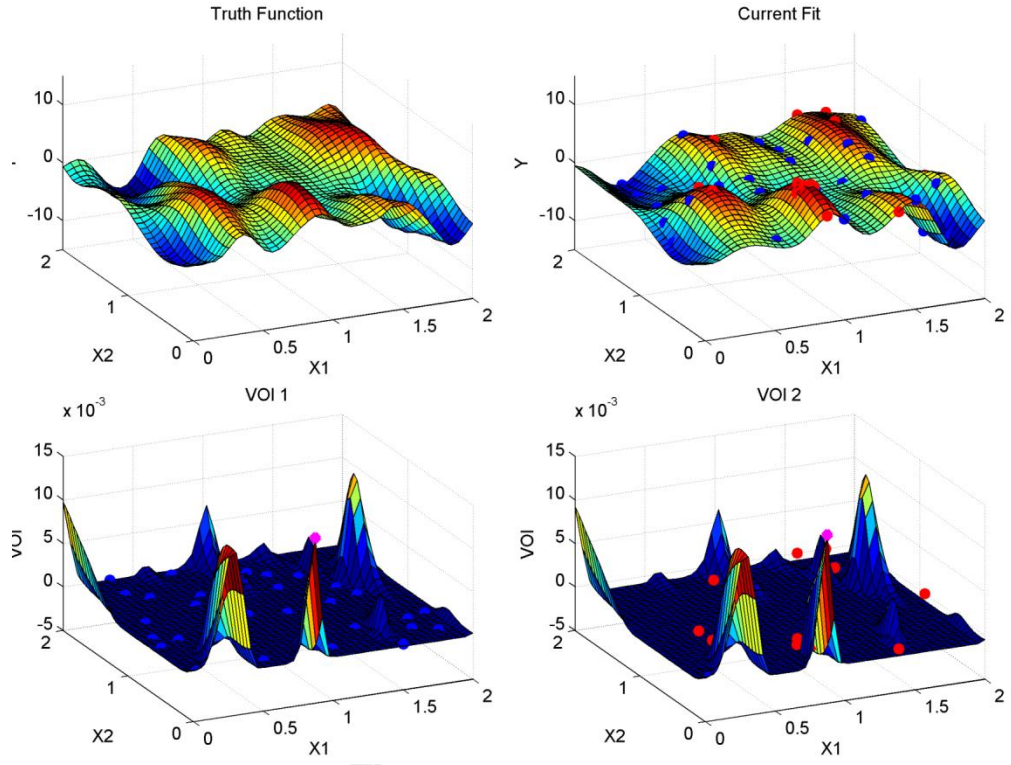

FIGURE 4.6: 16TH ITERATION OF VGO ALGORITHM

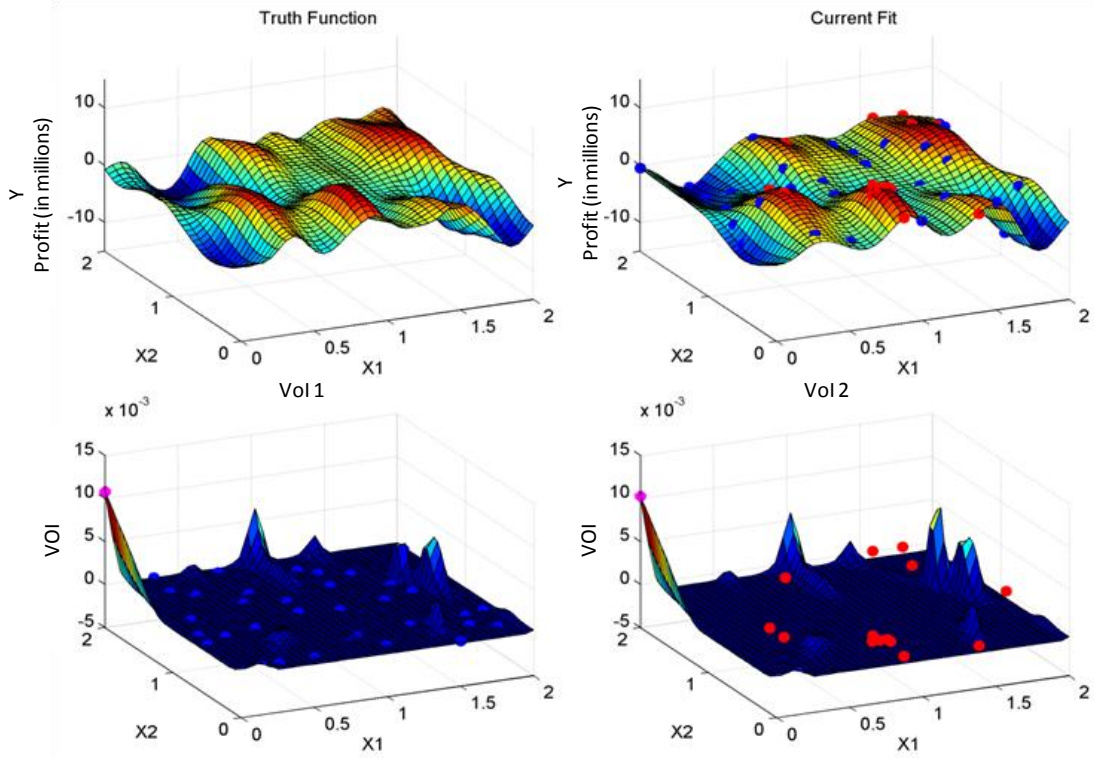

FIGURE 4.7: 20TH ITERATION OF VGO ALGORITHM

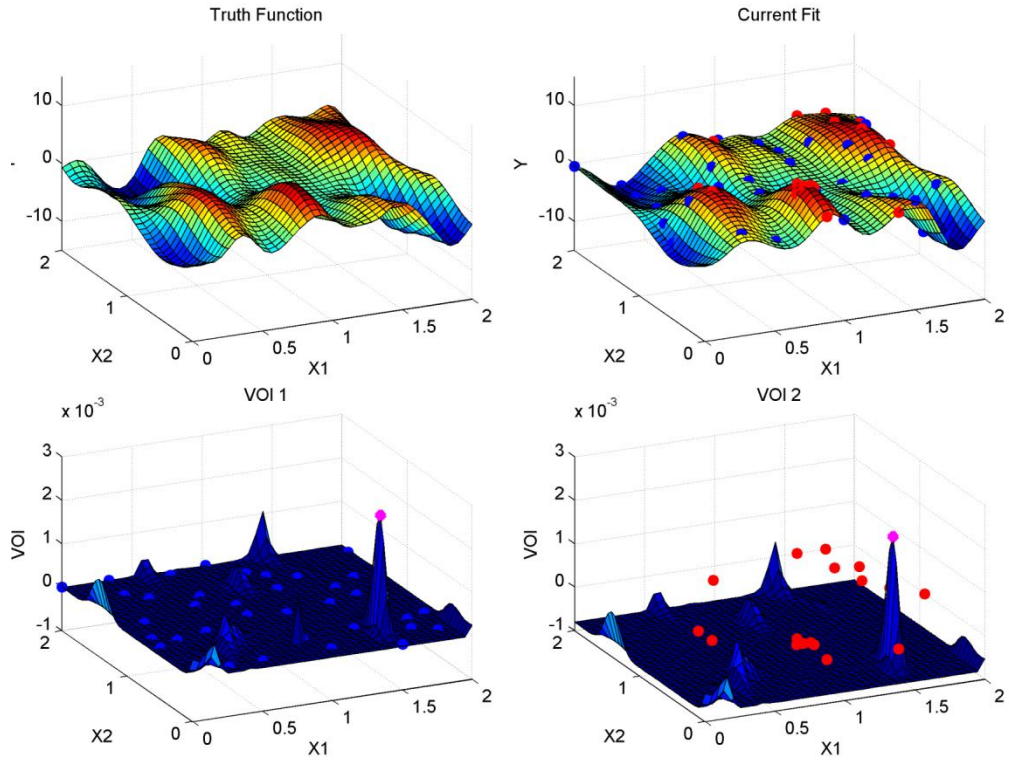

FIGURE 4.8: 24TH ITERATION OF VGO ALGORITHM

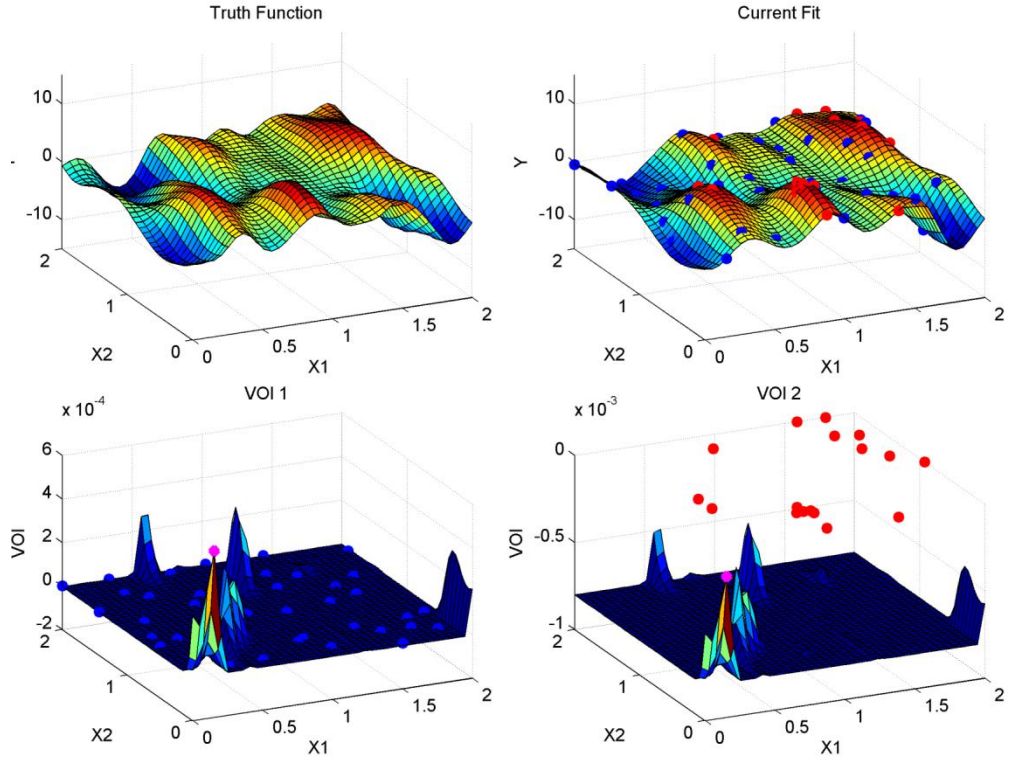

FIGURE 4.9: 28TH ITERATION OF VGO ALGORITHM

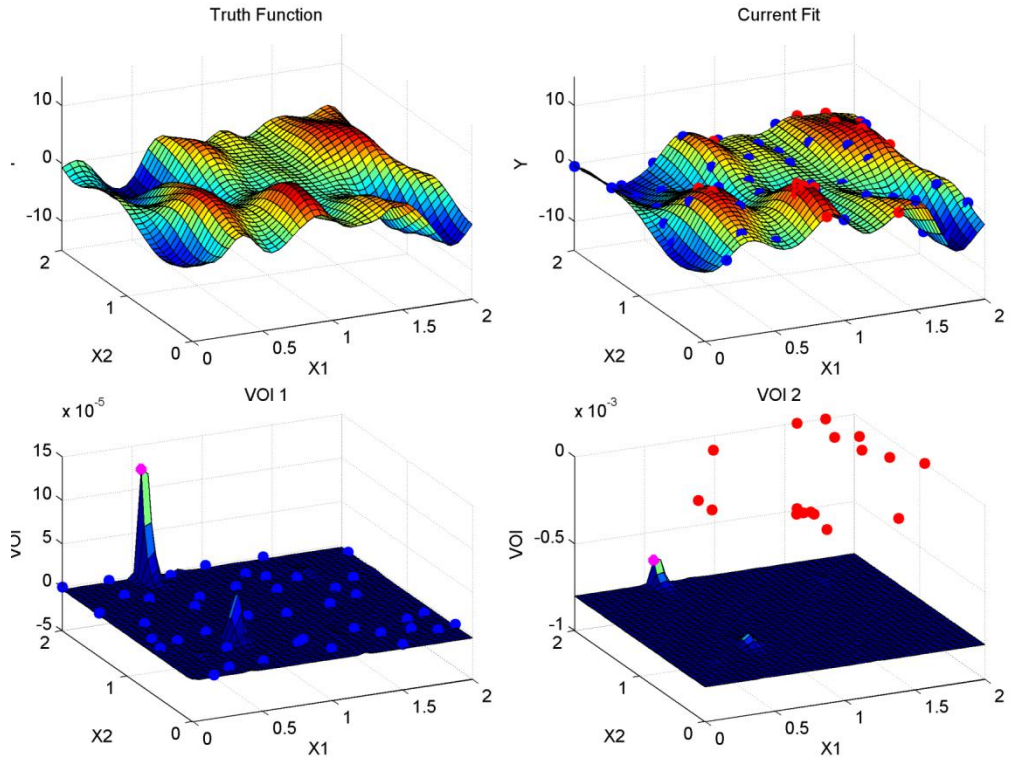

FIGURE 4.10: 32ND ITERATION OF VGO ALGORITHM

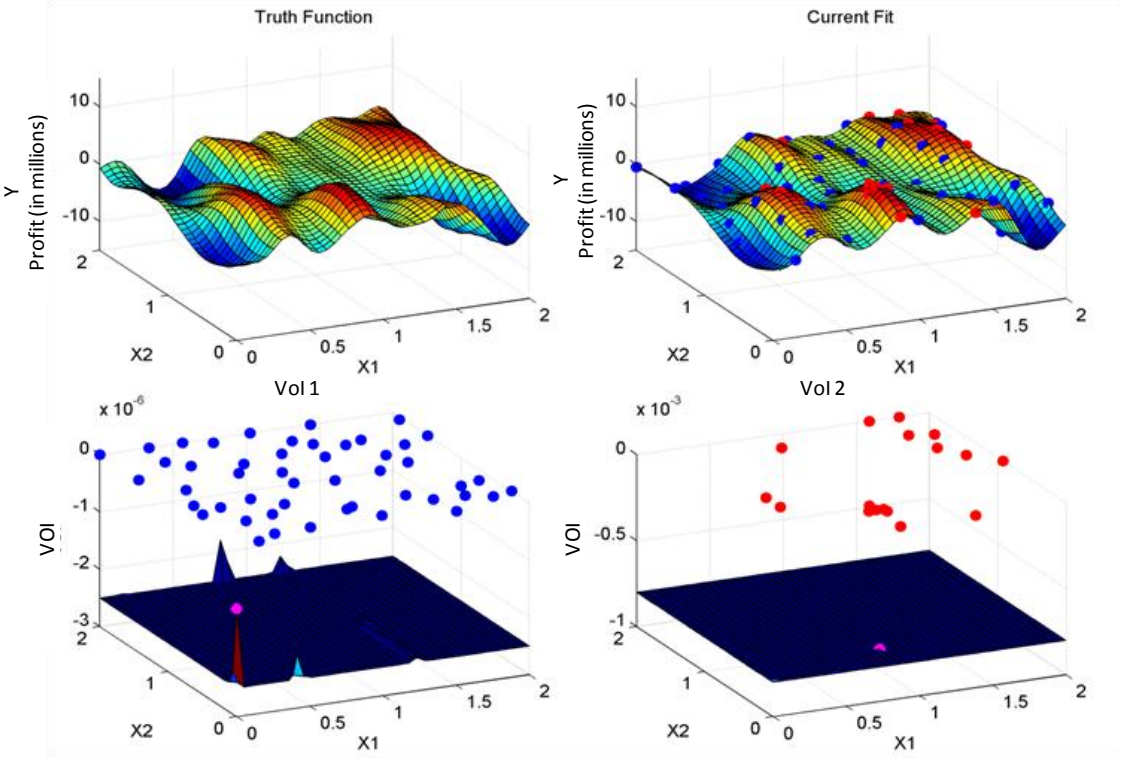

FIGURE 4.11: FINAL (37TH) ITERATION OF VGO ALGORITHM

4.7 THESIS ROADMAP

In this chapter, the theoretical foundations for the VGO algorithm were developed in detail. Emphasis was placed specifically on the mathematical derivation for the multi-accuracy, Gaussian process-based surrogate model that plays an important role in the functionality of the VGO algorithm. This derivation is similar to the classic kriging derivation, but different assumptions about the accuracy of the models with respect to the truth leads to a different covariance matrix and a greater number of hyperparameters to be determined using MLE. The second emphasis of this chapter is the application of VoI to this domain and the calculation of the individual terms in the VoI integral. This portion of the algorithm relies heavily on Bayesian statistics to formulate beliefs about the prior and posterior utility once a new sample site is added to the design space. Once the surrogate model and VoI were derived, the final sections of the chapter were focused on implementation details of the algorithm, including initialization, maximizing VoI, and final maximization of the truth prediction. The chapter concluded with a pictorial illustration of the VGO algorithm for a two-dimensional, two model test problem.

While the test problem illustrated in this chapter highlighted some of the promising attributes of VGO, rigorous experimentation is needed to characterize its performance. This will be the focus of Chapter 5.

CHAPTER 5: PERFORMANCE CHARACTERIZATION OF VALUE-BASED GLOBAL OPTIMIZATION

In the previous chapter, the theoretical foundations for the VGO algorithm were provided. In this chapter, several experiments are used to characterize the VGO algorithm. The goal is to show when it is favorable to use VGO and how best to use VGO. The experiments are presented in three sections. In Section 5.1, experiments are presented for comparing the performance of VGO with Efficient Global Optimization (EGO). In Section 5.2, the emphasis is on the scalability of VGO. The algorithm is applied to two well-known optimization test functions in three and six dimensions. Finally, in Section 5.3, experimental results are used to aid in a qualitative discussion about model usage; specifically, the usefulness of adding a model to the available analyses is discussed. The chapter concludes in Section 5.4 with a return to the thesis roadmap.

5.1 COMPARISON WITH EFFICIENT GLOBAL OPTIMIZATION (EGO)

In this section, the goal is to provide a rigorous performance comparison of VGO with EGO. The section begins with a review of EGO, which was first introduced in Chapter 2. This description of EGO is followed by a discussion of the ways in which VGO and EGO differ and a conceptual discussion about when it is best to use one algorithm or the other. Finally, an experimental comparison of the two algorithms and statistical analysis of the results is presented. The algorithm solutions are compared in terms of solution quality as well as costs incurred during the optimization process. For robustness, the experiment is run for different initial fixed samples.

5.1.1 Overview of Efficient Global Optimization

Efficient Global Optimization (EGO) [31] is a global optimization algorithm that leverages a kriging surrogate modeling technique in conjunction with the Expected Improvement (EI) sequential sampling strategy. The EI for a candidate design site *x* with an unknown observation y is the expected value of the potential improvement over the current best. The mathematical definition of EI is as follows:

$$\mathrm{E}[I(x)] = -(\hat{y}_{\max} - \hat{y})\left(1 - \Phi\left(\frac{\hat{y}_{\max} - \hat{y}}{\sqrt{\varphi}}\right)\right) + \sqrt{\varphi}\,\phi\left(\frac{\hat{y}_{\max} - \hat{y}}{\sqrt{\varphi}}\right) \tag{5.1}$$

In the above expression, Φ denotes a normal cumulative distribution function and ϕ is a normal probability distribution function. From Chapter 4, $\sqrt{\varphi}$ is the square root of the MSE function associated with the kriging model. The remaining terms are consistent with the definitions provided in Chapter 4. Again, *y* is assumed to be normally distributed. Conceptually, EI takes the expected value of the portion of the normal distribution on *y* that improves upon $\hat{y}_{\max}$. If the prior mean on *y* is sufficiently smaller than $\hat{y}_{\max}$, then integral of the tail of the distribution where the posterior mean would exceed $\hat{y}_{\max}$ is very small, resulting in a very small expected improvement. On the other hand, if the prior mean on *y* is near $\hat{y}_{\max}$, then there will exist a more substantial portion of the prior distribution that exceeds $\hat{y}_{\max}$, resulting in a larger integral and thus a larger expected improvement.

In EGO, sequential sampling ends when

$$\max(\mathrm{E}[I(x)]) < a \tag{5.2}$$

where *a* is a user-defined constant. It will be shown in the experimental results later in this section that the effectiveness of the EGO algorithm depends heavily on the choice of *a*. If *a* is too small, then the sampling and optimization process continues at the expense of additional analyses with little or no payout in terms of the objective function improving. On the other hand, if *a* is too large, then optimization can stop prematurely before a good design artifact is determined.

5.1.2 Comparison of Algorithms

For comparison purposes, the pseudo-code for EGO is provided below.

Efficient Global Optimization (EGO)

```
Initialize:
  set S := GenerateLHS(numSamples)
  set Y := AnalyzeModel(S)
Iterate:
  while forever
    set ŷ(x):=GenerateKriging(S,Y)
    set ŷ_max:= max(ŷ(s_1),…, ŷ(s_m))
    set [s_max, maxEI]:=MaximizeEI(ŷ(x), ŷ_max)
    if maxEI< a
      break while loop
```

```
    end if
    set S:=S ∪{s_max}
    set Y:= Y ∪{AnalyzeModel(s_max)}
  end while
Terminate:
  set globalMax:=Maximize(ŷ(x))
```

Similarly to VGO, the algorithm begins with a fixed sample of design sites to seed the initial kriging fit. Like VGO, the choice of this fixed sampling strategy is by no means limited to LHS, but that is the approach used in this thesis. The analysis model is evaluated at these design sites, providing the initial set of samples.

The surrogate modeling approach used in EGO is the classic kriging modeling technique and is not adapted for either multiple models or for model inadequacy. This is one primary disadvantage of EGO—EGO is only applicable for one analysis model. Not only is the choice of surrogates not suited to multi-accuracy modeling, but the EI metric does not provide a mathematical mechanism to account for analysis quality or cost; EI is based entirely on the prior mean and variance. As a result, if EI were to be calculated for more than one model, the result would be identical across models and would provide no mechanism for selecting the best analysis. Because no model uncertainty is assumed and only one model can be used, it is assumed in this thesis that EGO uses only the highest fidelity model when comparison experiments are run. This is probably the most significant drawback of EGO when compared to VGO.

Finally, the EGO stopping criterion is quite different from that of VGO, not only because it is based on EI instead of VoI, but because it relies so heavily on an appropriate selection of *a* by the user. This is another area where VGO improves on EGO; the stopping criterion is much more intuitive and less sensitive to a single user-defined parameter.

VGO does have some drawbacks when compared to EGO. The most significant drawback is that VoI is more expensive to calculate and subsequently optimize than EI. While both sampling metrics require a multi-start or global optimization technique in order to find the maximum, EI is a simpler calculation and lends itself easily to determining analytical gradient information. Therefore, if working with only inexpensive analysis models, it would be faster and simpler to run EGO with a few different choices of *a* and simply pick the best solution. It is assumed in this thesis that in engineering design, the analysis model costs are much more significant than the costs of surrogate fitting and optimization of VoI or EI. If this assumption is clearly violated, then some of the attractiveness of VGO wanes due to the complexity of the VoI maximization, which must be performed for every available model.

5.1.3 Performance Evaluation: VGO vs. EGO

In the previous section, a conceptual comparison of VGO and EGO was provided. In this section, the focus is on experimental data to validate that discussion. To test the performance of the VGO algorithm, it has been applied to a suite of randomly generated test problems. This same suite is then subsequently solved using EGO. In this section, we detail how these test problems were generated, the results of the algorithms, and the analysis costs incurred.

5.1.3.1 Generation of the Test Suite

While there is no shortage of known global optimization test problems, their primary objective is not in alignment with that of VGO. Normal global optimization test problems tend to be multi-modal with fairly sharp global maxima; this is perfect for testing a classic global optimization approach and even global optimization suites involving only one analysis model that do not account for cost. In VGO, the focus is not purely on mathematical optimality, but rather on good solution quality at reasonable cost. Additionally, to make full use of VGO, multiple test functions of different accuracies are required, which is not a standard feature of most known test functions. In Section 5.2, known test problems will be used with some adaptations, but in this section, the test suite used is original and designed to allow for multiple models and cost accounting.

The methodology developed to create the suite allows for the creation of any number of unique individual test problems or instances with any number of analysis models associated with a particular instance. To generate a particular test problem, a truth model is randomly generated as a realization of a Gaussian process with randomly selected variance and roughness parameters. This is done by generating a correlated set of random samples. The sample set must be sufficiently dense to avoid aliasing; that is, if the statistical parameters dictate a very rough surface, but samples are only sparsely generated, then the surface can look artificially smooth. As the number of samples increases, however, the size of the covariance matrix increases, resulting in more significant computational expense. For small test problems in one to three dimensions, this is generally not a problem, but as the dimensionality of the test problems increases, the size of this matrix must be taken into consideration.

After the correlated samples are generated, a classic kriging model is fit to those samples, and that function represents the *truth*. Although in practice the truth can never be known, for the purposes of performing a controlled computational experiment and assessing the algorithm's accuracy, a truth model is constructed.

Multiple analysis models are then generated using additional realizations of Gaussian processes to represent model inadequacies. By changing the variance of the Gaussian process, we can generate less accurate models (high variance) and more accurate models (low variance). The method is the same as that of generating the truth: generate a correlated sample, and then fit a kriging model to the sample. This model inadequacy term is then added to the truth to create a function for a particular analysis model. By generating the test suite this way, we ensure that the modeling assumption made in Equation 4.4 is in fact true. We know that the truth can be represented by a regression term and Gaussian process realization because we generated it that way, and similarly, we know that the model error can be characterized by an additional zero-mean Gaussian process, because that is how the analysis models were created. It is these analysis models from which observations are drawn in order to fit the Gaussian process surrogate surface and to run VGO.

This test suite is in some ways 'optimal' with respect to VGO because the truth is known, and we know that the modeling assumptions hold. In practice, the low and high fidelity models might correspond to a finite difference model and a differential equation-based model for the same system. Alternatively, the analysis models might be a set of finite element models of the system with different mesh resolutions, or even different considerations regarding linear and non-linear effects in the underlying model.

Additionally, the truth would be unknown. It is the responsibility of the user, in this case, to assign an appropriate variance, σ_i^2 to each model in order to characterize the analysis models' expected accuracy with respect to reality. These considerations will be discussed in more detail during the presentation of the practical example in Chapter 6. In this example, the variance used to generate the model inadequacy for each model is also used to calculate the VoI, which means that no additional model characterization is necessary. The only parameters that need to be supplied for this example are the costs of each of the models.

For the results presented in this section, a suite of 20 individual two-dimensional instances each having two analysis models is generated. The suite consists of a truth surface and two analysis models—one low fidelity model, and one high fidelity model. It is assumed that the truth models represent a profit function in millions of dollars; typical surfaces can range from about $1 million to as much as $10 million. As far as the specific statistical parameters, the truth models are generated using randomly generated variance (σ^2) values ranging from 0.01 to 10 and θ values ranging from 0.01 to 30. As such, the roughness of the surfaces is allowed to vary randomly when the suite is generated, so the surfaces range from very smooth to very peaked with many local optima. The result is similar to having different frequency content; when there is high correlation, the frequency content is lower, and when the correlation is lower, the frequency content is higher. For simplicity of comparison, the accuracies (σ_i^2) and costs of the two analysis models remain the same for each of the 20 instances. The cost for a model 1 analysis is $2.50 and the cost for one run of model 2 is $800. The model inadequacies characterized as variances are .05 and .0008, respectively.

5.1.3.2 Comparison Results

For this experiment, the 20 instance suite is run with the VGO algorithm and with the EGO algorithm. The same initial set of 30 LHS design sites are used for both VGO and EGO, but EGO is seeded with high fidelity design sites, while VGO is seeded with low fidelity design sites. For the EGO algorithm, multiple stopping criteria are used because the quality of the results can be highly dependent on the user's choice of a. Figure 5.1 shows the difference between VGO and EGO expected utility for various a values. To compute the *overall expected utility*, the *cost of analyses* is subtracted from the *artifact utility*. The artifact utility is the value of the *truth* at the input value returned by the optimizer. That is, the global optimization algorithm returns an x and a predicted $\hat{y}_{\max}$. Since the test suite affords the luxury of knowing the truth and the location of the true optimum, to evaluate the achieved artifact utility, the truth is evaluated at x. This represents the utility of the actual artifact that would be manufactured with specifications x. The cost of analysis is then determined by the number of function evaluations from each model (for EGO there is only the high fidelity model) multiplied by the assigned cost per analysis for that model. The difference between these two quantities gives the overall utility.

The goal of VGO is to achieve the best possible overall utility. To achieve a good overall utility, it is necessary to arrive at a good artifact utility while only incurring reasonable analysis costs. If the goal were purely to find the mathematical optimum, then VGO would not be required; it is because the costs are so vital engineering decision making that VGO is designed the way it is, and that these test suites were designed to make fair comparisons with respect to overall utility.

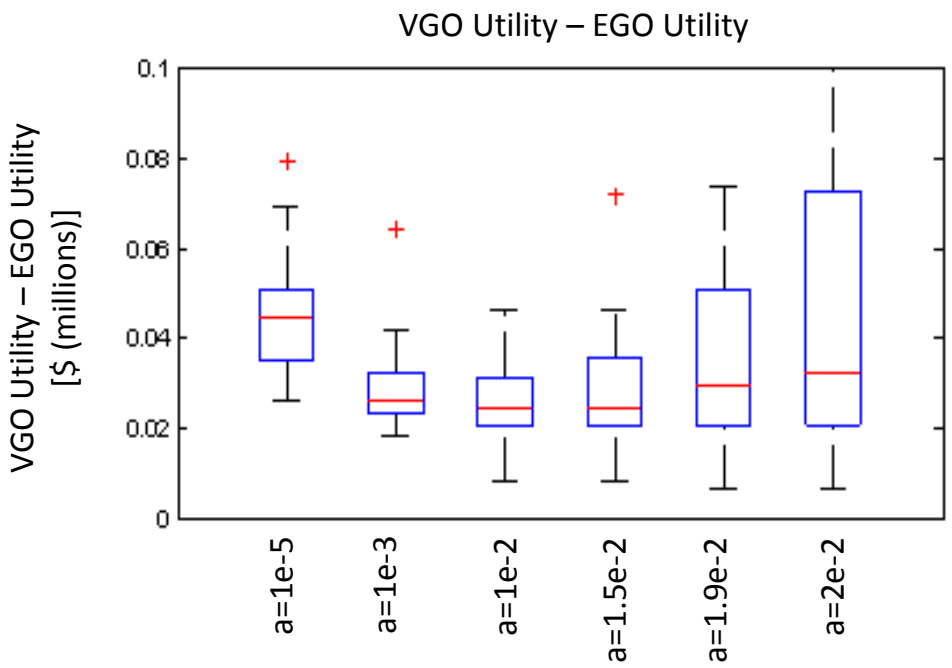

FIGURE 5.1: STATISTICS FOR VGO UTILITY MINUS EGO UTILITY WITH DIFFERENT STOPPING CRITERIA

The data captured in Figure 5.1 is a boxplot capturing the delta between VGO overall utility and EGO overall utility, assuming that the truth is known with certainty, for various *a* values over the suite of 20 instances. The boxplot shows the median in red, and the box shows the 25th and 75th percentiles. The whiskers indicate the extremes of the distribution, and additional points plotted separately are considered to be outliers. Overall, the expected utility is always higher on average for VGO than EGO. The different values of *a* illustrate an important trend. Moving from left to right in the figure, for small values of *a*, good quality solutions are found but at high cost. Then, as *a* increases, an optimal region can be seen where good solutions are found at reasonable cost—there is without a doubt, an optimal choice of *a*, but there is no way to determine this optimal value analytically. In this optimal region, VGO utility does not exceed EGO utility by very much. Then, as *a* continues to increase, solution quality is no longer

reliable, though the cost is low. This is why the variance on the solutions increases as *a* gets large; sometimes, the optimizer gets 'lucky' and finds a good solution quickly and hence at low cost, and other times stops very early before finding a good solution.

While the expected utility results for VGO show promise in comparison to EGO, this particular suite is such that the cost of analysis is not a very large portion of the overall utility; that is, artifact utility dominates the overall utility. The same results can be interpreted differently by focusing purely on the cost of analyses, as shown in Figure 5.2.

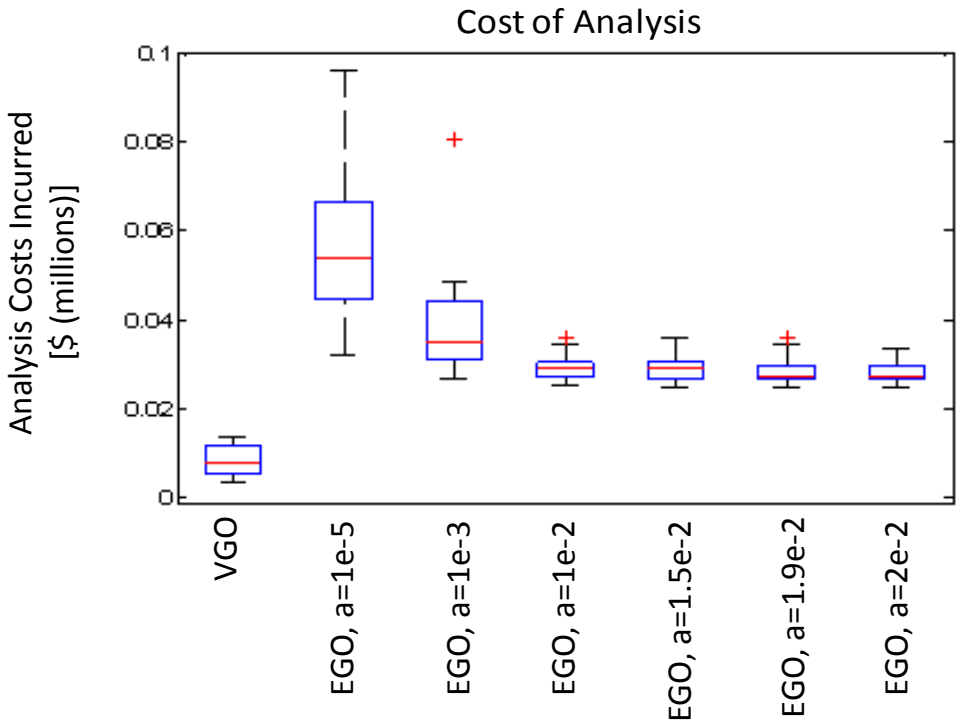

FIGURE 5.2: DIFFERENCE IN ANALYSIS COSTS USING VGO VERSUS EGO WITH DIFFERENT STOPPING CRITERIA

Here, a box plot containing VGO total cost of analysis followed by the EGO results of analysis costs for the various *a's* are presented. VGO is the clear winner in terms of lower analysis cost, independent of the stopping criteria for EGO. However, it is not

meaningful to consider cost in isolation without considering solution quality—the cheapest analyses would always win if artifact utility is not taken into consideration. The value of *a* for EGO which gave the most comparable results to VGO in terms of the quality of the artifact utility is *a*=1e-3. Comparing only that particular set of EGO results to VGO, the total costs incurred were $776,000 for EGO and $168,020 for VGO. Therefore, for only very small differences in artifact utility, the EGO algorithm incurred analysis costs 4.6 times those incurred by VGO. This is a very promising result. It indicates that leveraging multiple models at different costs and accuracies can help designers to conserve resources. If also indicates that VoI, in conjunction with multiple models, allows for cost effective selection of valuable analyses during optimization.

Finally, to compare strictly the quality of the optimization results, it is necessary to examine the artifact utilities achieved by both algorithms in isolation. Figure 5.3 depicts a box plot of the distributions of the differences between VGO artifact utility and EGO artifact utility for different stopping criteria. Positive values indicate that VGO found a superior design artifact, while negative values indicate that EGO found a superior design artifact for particular test problems. Overall, VGO performs comparably to EGO for the smallest stopping criteria; that is, VGO is on par with the tightest tolerances tested for EGO. As the stopping criteria becomes larger and therefore less stringent, the VGO solution quality tends to be measurably better than that of EGO.

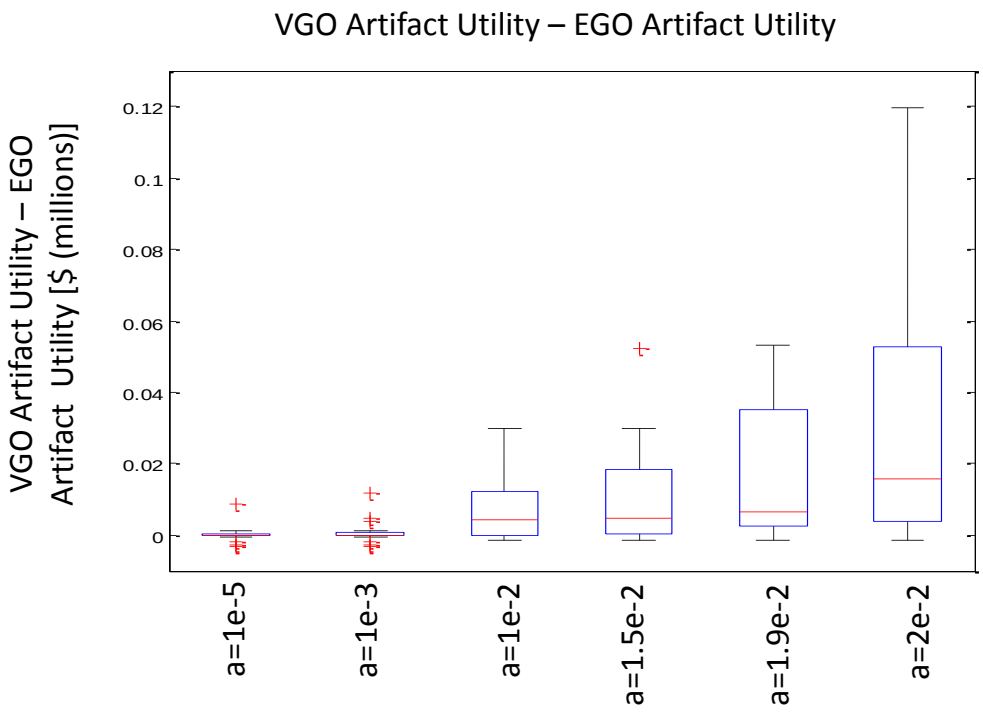

FIGURE 5.3: STATISTICS FOR VGO ARTIFACT UTILITY MINUS EGO ARTIFACT UTILITY WITH DIFFERENT STOPPING CRITERIA

5.1.3.3 Sign Test Results

In the previous section, a single suite of 20 two-dimensional, two-model problems with an initial sample of 30 LHS was used to compare the outcomes of VGO and EGO with respect to overall utility, artifact utility, and analysis cost incurred. While the previous section focused on a more visual version of the results and a discussion of the qualitative properties of the figures, the focus in this section is on showing statistical significance.

The first experiment in this section is the exact same experiment from the previous section, but the results are presented differently. In these experiments, a sign test is used to test the hypothesis '*The VGO utility exceeds the EGO utility with median=m'*. This is known as the null hypothesis. The sign test allows us to either accept or reject the null hypothesis with a particular one-sided p-value.

The first set of results is captured in Table 5.1. The EGO algorithm is run with different values of *a*, which is shown in the left-most column of the table. The top row of the table shows the median value *m* for the hypothesis being tested. The 'W' columns indicated the number of wins out of a possible 20 trials; a win is achieved if the VGO utility less the EGO utility is a positive number. On average, as the median value is increased, the number of wins decreases. For the results to be statistically significant, at least 15 out of 20 wins must be achieved to accept the null hypothesis; otherwise, the hypothesis is rejected. Cells indicating a rejected hypothesis are shaded in red.

TABLE 5.1. SIGN TEST RESULTS, 30 LHS SAMPLES

VGO vs EGO, 30 LHS Initial Sample										
	Median= 0		Median= 0.01		Median= 0.02		Median= 0.022		Median= 0.025	
	W	P-value	W	P-value	W	P-value	W	P-value	W	P-value
a=1e-5	20	9.54E-07	20	9.54E-07	20	9.54E-07	20	9.54E-07	19	2.00E-05
a=1e-3	20	9.54E-07	20	9.54E-07	18	2.01E-04	17	1.29E-03	13	1.32E-01
a=1e-2	20	9.54E-07	20	9.54E-07	17	1.29E-03	15	2.07E-02	11	4.12E-01
a=1.5e-2	20	9.54E-07	20	9.54E-07	18	2.01E-04	15	2.07E-02	12	2.52E-01
a=1.9e-2	20	9.54E-07	20	9.54E-07	17	1.29E-03	16	5.91E-03	14	5.77E-02
a=2e-2	20	9.54E-07	20	9.54E-07	17	1.29E-03	16	5.91E-03	14	5.77E-02

Obviously, with *m=0*, VGO is the clear winner independent of the value of *a*. The purpose of testing the different medians is merely to see by how much VGO wins. Since the objective is actually in millions, when VGO wins by *m=0.01*¸ that corresponds to a net increase in profit of $10,000.

To test the robustness of VGO with respect to the initial sample, the same experiment was run on the same test suite of 20 instances but with 20 LHS initial samples and 10 LHS initial samples. These results are captured in Table 5.2 and Table 5.3, respectively.

TABLE 5.2. SIGN TEST RESULTS, 20 LHS SAMPLES

VGO vs EGO, 20 LHS Initial Sample										
	Median= 0		Median= 0.01		Median= 0.02		Median= 0.022		Median= 0.025	
	W	P-value	W	P-value	W	P-value	W	P-value	W	P-value
a=1e-5	20	9.54E-07	19	2.00E-05	17	0.0013	16	0.0059	15	2.07E-02
a=1e-3	19	2.00E-05	17	1.30E-03	6	0.0577	6	0.0577	6	5.77E-02
a=1e-2	19	2.00E-05	19	2.00E-05	13	0.1316	12	0.2517	11	0.4119
a=1.5e-2	19	2.00E-05	19	2.00E-05	13	0.1316	12	0.2517	11	0.4119
a=1.9e-2	19	2.00E-05	19	2.00E-05	13	0.1316	12	0.2517	11	0.4119
a=2e-2	19	2.00E-05	19	2.00E-05	13	0.1316	12	0.2517	11	0.4119

TABLE 5.3. SIGN TEST RESULTS, 10 LHS SAMPLES

VGO vs EGO, 10 LHS Initial Sample										
	Median= 0		Median= 0.01		Median= 0.02		Median= 0.022		Median= 0.025	
	W	P-value	W	P-value	W	P-value	W	P-value	W	P-value
a=1e-5	20	9.54E-07	20	9.54E-07	16	0.0059	15	2.07E-02	15	2.07E-02
a=1e-3	20	9.54E-07	17	1.30E-03	8	0.2517	8	0.2517	6	5.77E-02
a=1e-2	19	2.00E-05	16	0.0059	13	0.1316	12	0.2517	12	0.2517
a=1.5e-2	19	2.00E-05	16	0.0059	12	0.2517	12	0.2517	12	0.2517
a=1.9e-2	20	9.54E-07	18	2.01E-04	13	0.1316	12	0.2517	12	0.2517
a=2e-2	20	9.54E-07	17	1.30E-03	12	0.2517	12	0.2517	12	0.2517

Clearly, from the above tables, VGO is still the clear winner overall with statistical significance, but on the whole, the amount by which VGO wins is smaller for these particular sets of samples. Some of that discrepancy is merely due to the stochastic nature of the algorithm. It can also happen that for a small initial sample, the VoI calculation is working with less complete information. In the 30 sample case, VGO is initialized with low fidelity samples. It can happen that with only 10 or 20 samples to go on, the high fidelity model is used sooner before a more complete exploration of the design space occurs, resulting in a higher total number of high fidelity function

evaluations. This hypothesis is validated by the fact that the analysis cost for VGO is on average higher for the 20 and 10 LHS case than it is for the 30 LHS case.

An alternate view of the results of this experiment is provided in Table 5.4. This table was produced by leveraging the knowledge that 15 wins amount to a statistically significant validation of the null hypothesis. By calculating the difference between VGO utility and EGO utility and sorting the values from largest to smallest, the 15^{th} value gives the maximum median value *m* for which the null hypothesis would be accepted. It is particularly clear in this table that VGO achieves the greatest improvement over EGO with 30 initial samples, and the smallest improvement over EGO with only 10 initial samples. However, VGO is still the overall winner regardless of the initial sample size.

TABLE 5.4. A SUMMARY OF MEDIAN VALUES BY WHICH VGO UTILITY EXCEEDS EGO UTILITY

Median Values by Trial			
	30 LHS	**20 LHS**	**10 LHS**
a=1e-5	0.0325	0.0264	0.0252
a=1e-3	0.0248	0.0129	0.0117
a=1e-2	0.0222	0.0158	0.0107
a=1.5e-2	0.0223	0.0158	0.0107
a=1.9e-2	0.0236	0.0158	0.0128
a=2e-2	0.0236	0.0158	0.0107

In summary, the VGO algorithm outperforms EGO in terms of overall utility (artifact utility less analysis costs) with statistically significant results as determined using the sign test.

5.2 SCALABILITY

In the previous section, it was shown that VGO works well as compared to EGO, but the test problems were limited to two dimensions. In this section, the goal is to show that

VGO can be successfully applied to higher dimensional problems. No comparisons are made in this section; these experiments are merely an illustration of the VGO algorithm's capabilities in higher dimensional design spaces.

For these experiments, known global optimization functions from Dixon and Szego are used [11]. Specifically, the Hartmann-3 in three dimensions and Hartmann-6 in six dimensions are used. Obviously, per the discussion in Section 5.1, the test functions must be adapted somewhat to accommodate multiple accuracy models. There are some slight differences in the approach taken for each function, but the overall concept is that the original published function is used as the truth, and Gaussian processes are again generated to represent model inadequacies. The sum of the truth and these inadequacies are used as the available analysis models to solve the optimization problem.

5.2.1 Hartmann-3 Results

The Hartmann-3 test function is a common global optimization test problem originally defined by Dixon and Szego [11]. The Hartmann-3 function is defined as

$$y(x_1, x_2, x_3) = -\sum_{i=1}^{4} \alpha_i \exp[-\sum_{j=1}^{3} A_{ij}\left(x_j - P_{ij}\right)^2] \tag{5.3}$$

The three dimensional search space is defined as

$$0 \leq x_i \leq 1 \; \forall \; i = 1,2,3 \tag{5.4}$$

The parameters are defined as follows:

$$\boldsymbol{\alpha} = [1.0\ , 1.2, 3.0, 3.2]' \tag{5.5}$$

$$A_{ij} = \begin{bmatrix} 3 & 10 & 30 \\ 0.1 & 10 & 35 \\ 3 & 10 & 30 \\ 0.1 & 10 & 35 \end{bmatrix} \tag{5.6}$$

$$P_{ij} = \begin{bmatrix} 0.36890 & 0.1170 & 0.26730 \\ 0.46990 & 0.43870 & 0.74700 \\ 0.10910 & 0.87320 & 0.55470 \\ 0.03815 & 0.57430 & 0.88280 \end{bmatrix} \tag{5.7}$$

The test problem is normally formulated as a minimization, but in this thesis the negative is used so that it becomes a maximization problem. For the maximization case, Hartmann-3 exhibits four local maxima.

The solution to Hartmann-3 is shown in .

TABLE 5.5. HARTMANN-3 TRUE SOLUTION

Hartmann-3	
Global Maximum	[0.114614, 0.555649, 0.852547]
Maximum Value	3.86278

For this experiment, the Hartmann-3 test function is used to represent the truth. A low and high fidelity model are then generated using the same correlated Gaussian process generation technique described in 5.1 when generating the test suite. These Gaussian processes represent the model inadequacy term and is added to the truth to create an analysis model. Two model inadequacies were generated with variances of 0.1 and 0.001 for low and high fidelity, respectively. Since the solution is ~3.8, these variances translate to 2-standard deviations of 0.63 and 0.063, respectively, or 16.3% and 1.63%, respectively. The assigned model costs are $.000025 and $.0004. In terms of the costs, they are inexpensive enough (considering the maximum profit is only $3.86) to ensure

that it is worthwhile to perform some additional analyses. Also, the high fidelity model is an order of magnitude more accurate than the low fidelity model, so the cost is at least an order of magnitude higher. Selecting appropriate costs in practice will be discussed in greater detail in Chapter 6.

The experiment is run with an initial LHS sample size of 40 low fidelity samples. Upon completion, the VGO algorithm has sampled the low fidelity model a total of 52 times (an additional 12 beyond the 40 seed samples) and the high fidelity model a total of 48 times. The results of the experiment are captured in . The predicted maximum is the final set of x-coordinates determined by the optimizer, and the predicted maximum value is the estimate of the surrogate model at the predicted maximum. The artifact utility is the actual Hartmann-3 function value at the predicted maximum. The cost is computed based on the number of function evaluations and their associated costs, and the difference between the artifact utility and the cost incurred gives the actual utility.

TABLE 5.6. HARTMANN-3 EXPERIMENTAL RESULTS FROM VGO

Hartmann-3 Experimental Result	
Predicted Maximum	[0.3466, 0.5541, 0.8662]
Predicted Maximum Value	3.9864
Artifact Utility	3.8141
Cost	0.0205
Final Utility	3.7936

It appears that the VGO algorithm found a local maximum instead of the global maximum, but that is less important than the overall solution quality in terms of cost and what is achievable in terms of the model accuracy. Realistically, the best possible solution VGO can find is the optimum of the highest fidelity model. If the optimum for the highest fidelity model is not the same as that of the truth, that discrepancy can only be

accounted for in the uncertainty and how much analysis cost should be invested in that model. For this example, the artifact utility found was 3.81 and the actual artifact utility maximum was 3.86, a difference of .05. Previously, it was established that two standard deviations for the highest fidelity model is .063 with respect to the truth, and this solution falls within those bounds.

The goal of this experiment was to show that VGO can find a good solution at a reasonable cost for a problem of higher dimensionality. In this experiment, VGO successfully found a good solution within the expected error bounds of the global optimum considering the model inadequacy of the best available model. In the next experiment, VGO will be applied to a 6-dimensional problem.

5.2.2 Hartmann-6 Results

The goal for this experiment is to show that VGO can be successfully applied to an even higher dimensional problem; for this problem, the design space is assumed to be in six dimensions. The Hartmann-6 test function is defined as follows:

$$y(x_1, \dots, x_6) = -\sum_{i=1}^{4} \alpha_i \exp[-\sum_{j=1}^{6} B_{ij}\left(x_j - Q_{ij}\right)^2] \tag{5.8}$$

With

$$0 \leq x_i \leq 1 \; \forall \; (i = 1, \dots, 6) \tag{5.9}$$

And the parameters are defined as follows:

$$\boldsymbol{\alpha} = [1.0\ , 1.2, 3.0, 3.2]' \tag{5.10}$$

$$B_{ij} = \begin{bmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{bmatrix} \tag{5.11}$$

$$Q_{ij} = \begin{bmatrix} 0.1312 & 0.1696 & 0.5569 & 0.0124 & 0.8283 & 0.5886 \\ 0.2329 & 0.4135 & 0.8307 & 0.3736 & 0.1004 & 0.9991 \\ 0.2348 & 0.1451 & 0.3522 & 0.2883 & 0.3047 & 0.6650 \\ 0.4047 & 0.8828 & 0.8732 & 0.5743 & 0.1091 & 0.0381 \end{bmatrix} \tag{5.12}$$

Much like the Hartmann-3 test case, the known Hartmann-6 test function is used as a truth function to assess the quality of the results achieved. In the Hartmann-3 test, a three dimensional Gaussian process realization to represent model inadequacy was generated for each of the two desired analysis models and then added to the truth to give the final analysis model functions. In this case, the same concept is applied, but a six dimensional problem would result in a very large covariance matrix to generate the correlated sample. To make the problem more computationally feasible, two three dimensional correlated samples are generated and added as separate error terms; that is, one is assumed to be a function of $x_i \ \forall \ i = 1,2,3$ and the other is assumed to be a function of $x_i \ \forall \ i = 4,5,6$. This is an approximation, but a fairly subtle one, and simply saves time in the problem generation stage.

The Hartmann-6 function is also intended for minimization, so for this thesis, the negative is taken to create a maximization problem. It has six local maxima, and the global maximum is in .

TABLE 5.7. HARTMANN-6 TRUE SOLUTION

Hartmann-6	
Global Maximum	[0.2017, 0.15, 0.4769, 0.2753, 0.3117, 0.6573]
Maximum Value	3.32237

To generate the low and high fidelity analysis models, variances of 0.1 and 0.001 and costs of $.000025 and $.0004 are used for low and high fidelity models, respectively. These are the same generation parameters as used in the Hartmann-3 example.

To run the Hartmann-6 optimization, 120 low fidelity LHS samples were used. After the VGO algorithm was completed, the low fidelity model was sampled 7 additional times, and the high fidelity model was sampled 6 times. The optimization results are captured in Table 5.8.

TABLE 5.8: HARTMANN-6 EXPERIMENTAL RESULTS FROM VGO

Hartmann-6 Experimental Result	
Predicted Maximum	[0.1848, 0.1522, 0.4675, 0.2694, 0.3032, 0.6637]
Predicted Maximum Value	3.2838
Artifact Utility	3.312
Cost	0.005575
Final Utility	3.306425
Max Hi Fidelity	3.3288

For this experiment, the predicted maximum came out to be reasonably close to the true maximum. The true artifact utility at the solution was 3.312, as compared to 3.32237 for the true Hartmann-6 function, and 3.3288 which was the maximum of the highest fidelity model. Since the same distribution parameters hold, the artifact utility is within about .01 of the true solution, which is well within the two standard deviations of .063 for the high fidelity model with respect to the truth.

Overall, the algorithm performed effectively and efficiently with relatively few additional function evaluations on a high dimensional test problem. The results of the Hartmann-3 and Hartmann-6 test cases illustrate that VGO is scalable and very functional in higher dimensional design spaces.

5.3 SELECTION OF MEANINGFUL COLLECTIONS OF MODELS

The purpose of the experiments presented in this section is to provide some basis for a qualitative discussion about selecting meaningful models. In this section the focus is not so much on the individual experimental outcomes but on the heuristics that can be extracted from them. Specifically, the goal is guide the selection of models with meaningful cost and accuracy combinations that will get used if VGO is run. Intuitively, it may be tempting to use any and all available models when running VGO; however, for any analysis model available during the optimization, VoI is maximized, which does incur some computational expense. Hence, it is useful to use some discretion when selecting analysis models for use during VGO.

The first heuristic is very logical and can be justified without any experimentation. Sometimes, there will be cases where one model *dominates* another in the Pareto sense. For example, if you a designer has two models available, model A and model B, and model B is cheaper with the same or better accuracy as model A, then there is no value in ever using model A. Similarly, if model B is more accurate than A at the same cost or cheaper, then there will be no value in using model A. This is illustrated in Figure 5.4. It should be noted that a model is more accurate as variance decreases in order for the figure to be logical.

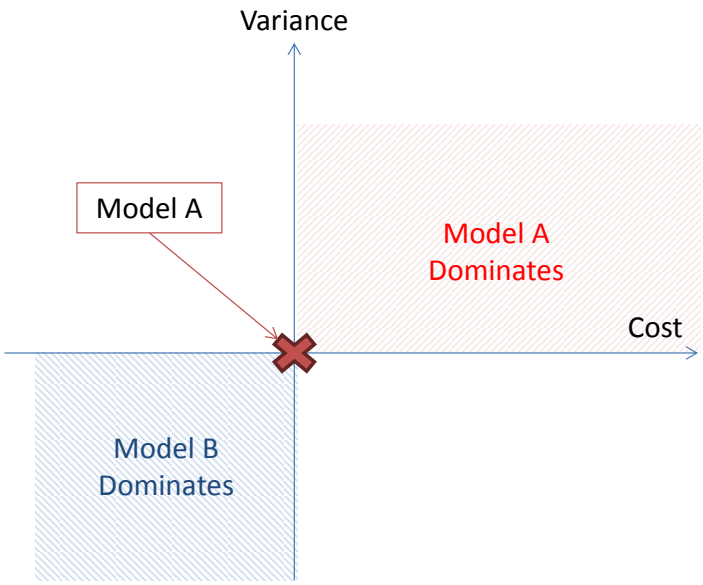

FIGURE 5.4: ANALYSIS MODEL PARETO DOMINANCE

If a second model, Model B, is to be added to the available analyses, and it does not dominate Model A (in which case Model A should not be used), then the question becomes, where within the two remaining quadrants is adding Model B valuable? For some experimental validation, it is assumed that a model is valuable if it is called for analysis at least one time during the VGO optimization. Clearly, the cost-variance combinations on the axes are still dominated, but at some points both models are clearly valuable.

To begin a study on the value of particular analysis models, work from Thompson *et al.* provides a useful starting point [63]. Before considering the case when multiple models are valuable, it is first necessary to consider when only one model is valuable in the context of decision making. In some cases, the cost-variance combination of a model

results in it being completely inefficient and a waste of resources; it is better to simply make a decision. This boundary is captured by Thompson in Figure 5.5.

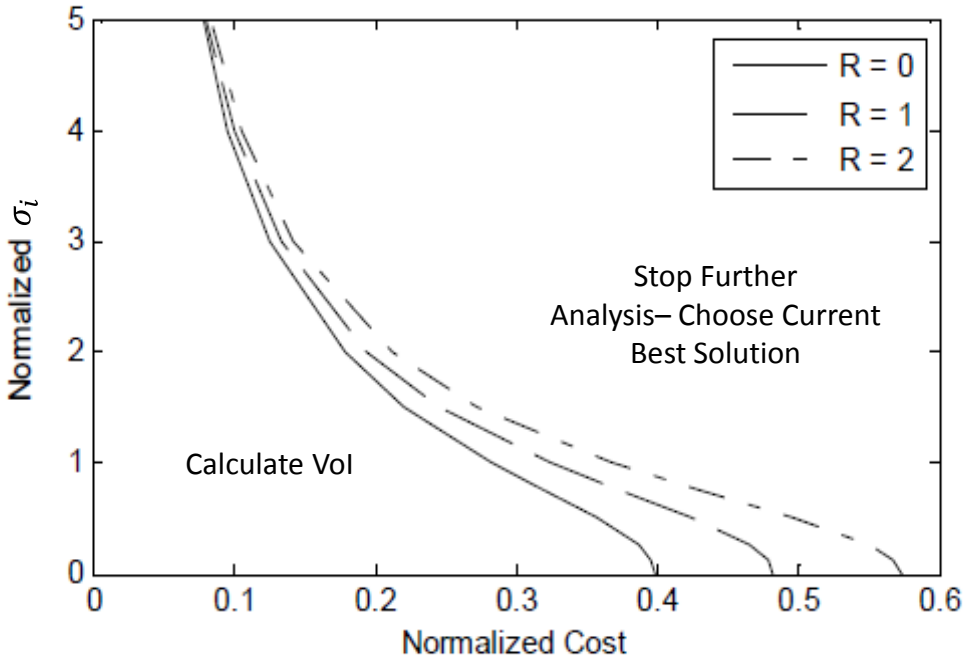

FIGURE 5.5: SCREENING TEST FOR COST-VARIANCE COMBINATIONS [63]

To fully understand this plot, a few terms need to be defined. In this figure, the axes are normalized cost and a normalized standard deviation. Different analysis models have different properties, but everything can be reduced to a scaling problem. For example, a surface that appears very flat over a large range of possible outcomes will exhibit more surface features over a smaller range of possible outcomes. Similarly, a utility function may predict outcomes in the billions or on the order of only a dollar. The analysis costs must be scaled accordingly. This scaling problem can be solved by normalizing the analysis model cost and variance with the variance of the truth or of the highest fidelity model. In the case of this figure, it is assumed that there are two possible decision

outcomes and two available analysis models. The decision maker can either make a choice up front or choose to analyze one of the models before either analyzing the other model or making a decision. Thus, one model is normalized with respect to the standard deviation other to achieve the normalized ratios in the plot. In this thesis, when dealing with test problems where the truth is available, these normalized quantities are computed with respect to the standard deviation of the truth.

The plot also shows different values of R, which is the constant representing risk aversion. In this thesis, we are concerned only with the risk neutral case, though other risk preferences could surely be applied. Therefore, it is the solid line with which we are most concerned. Interestingly, at some point, no matter how accurate a particular analysis model is, it will never be used if it is too costly with respect to the range of the objective function. This is the situation to the right of the solid line, where the plot is labeled 'Select product with higher expected utility.' This means that analysis should be skipped altogether and a decision should be made based on the available information. The other side of the line reads 'See boundary plot'—these are not addressed in this thesis, but the idea is that a more rigorous test should be used to determine if an analysis is valuable. In this thesis, we would calculate the VoI before performing the analysis.

To ensure that the ratios calculated by Thompson are indeed applicable to VGO, a brief experiment was conducted. For a suite of 20 analyses, the maximum VoI calculated for any available analysis at any step in the optimization was retrieved. This quantity, when scaled with respect to the standard deviation of the truth, gives the maximum analysis cost that would result in the model being used at that particular iteration. In other words, this quantity is an absolute upper bound on when a model would have ever been used at a

particular cost during the optimization process. By scaling the maximum VoI, which translates to the maximum possible cost, and the model accuracy with respect to the standard deviation of the truth, the distribution in Figure 5.6 is achieved.

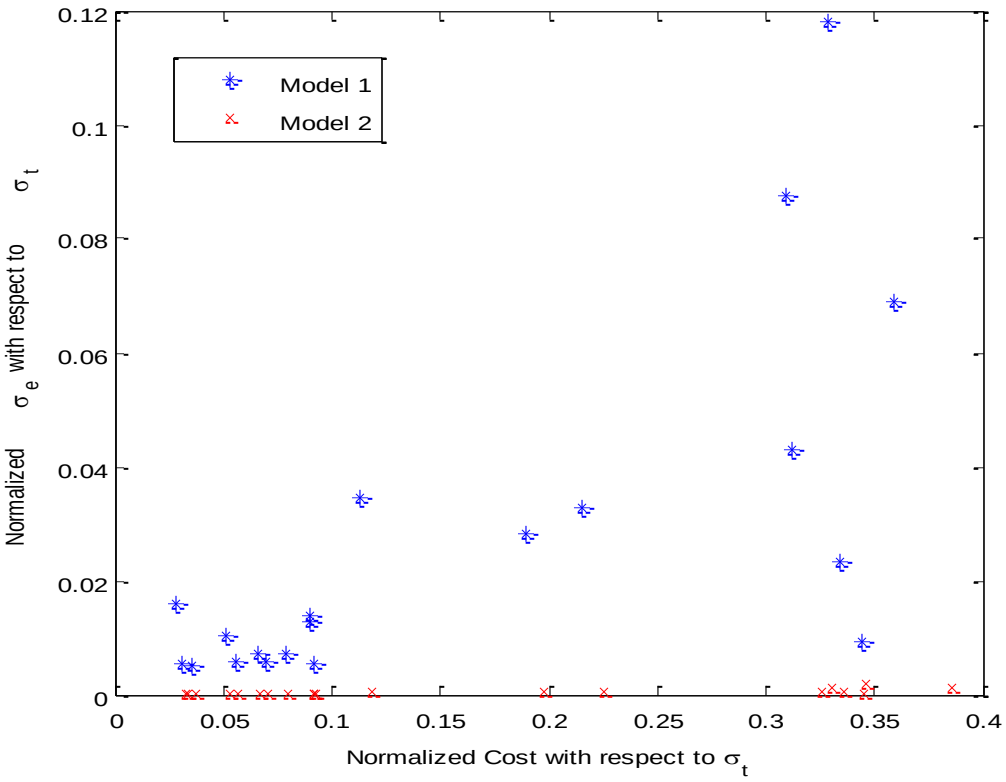

FIGURE 5.6: INITIAL SCREENING FOR MODEL COST-ACCURACY COMBINATIONS USED IN VGO

As can be seen by the higher model standard deviations (σ_ε), Model 1 is the low fidelity model and Model 2 is the high fidelity model. This data indicates that the cutoff normalized cost of 0.4 computed by Thompson is likely valid for determining if a model is too costly in the context of VGO.

To address the original question of where within the quadrant a *second* model is valuable, a large suite of experiments was run. For this experiment, a set cost-variance

combination was selected for model 1, and the model 2 cost-variance combinations were allowed to vary. For each cost-variance combination, a suite of 20 test problems was run. If the second model was called even once, the optimization was terminated and the second model was considered to be valuable. If the second model is never used, it is considered to not be valuable. Using this approach, a probability of a model being valuable can be calculated as the number of valuable uses out of a possible 20 trials. The results are shown in Figure 5.7. It should be noted that the sample sizes are not large enough to achieve statistical significance, and there is a stochastic property to these optimizations such that there is a reasonable degree of uncertainty in the probability estimations. This study is not intended to rigorously determine the probability of a model being valuable; it is simply an illustration to guide the discussion.

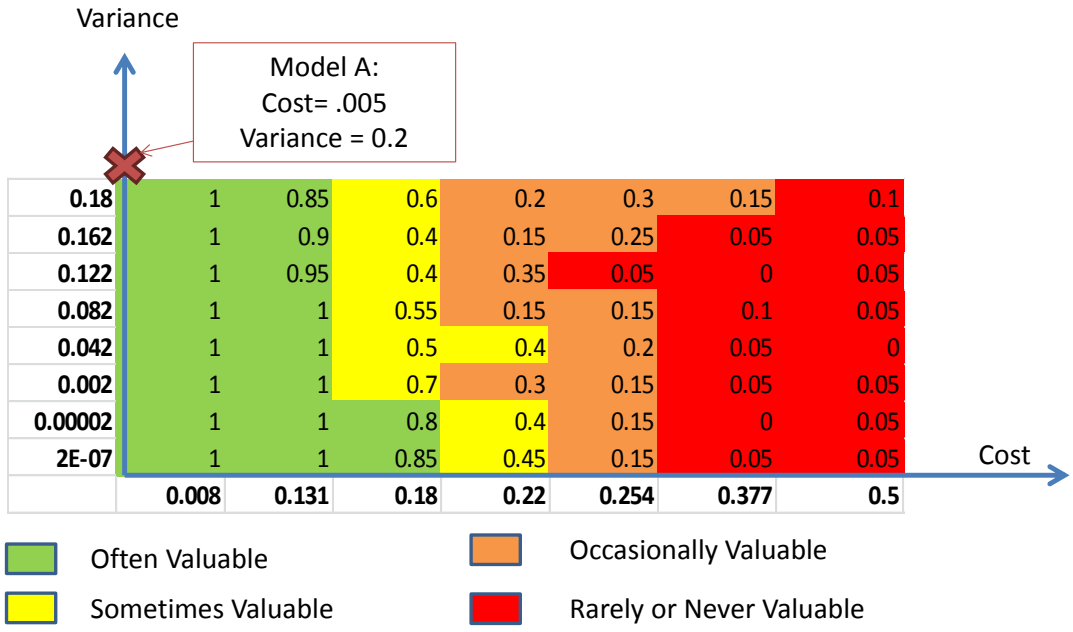

FIGURE 5.7: PROBABILITY OF MODEL B BEING VALUABLE

The figure presented is intended to mimic the fourth quadrant (bottom right) of Figure 5.4. The values along the cost axis represent normalized costs, and the values on the variance axis are normalized variances. Each cost-variance combination is considered in conjunction with the given cost-variance combination for Model A. The probability in each entry of the matrix represents the number of times Model B was deemed valuable out of 20 trials. As per the caveat in the previous paragraph, this sampling is insufficient to show any statistical significance. As such, the different outcomes are grouped into four categories: Often valuable, for estimated probabilities of 0.8 and higher, Sometimes valuable, for estimated probabilities of 0.4 to 0.8, Occasionally valuable, for estimated probabilities of 0.15 to 0.6, and Rarely or Never valuable, for estimated probabilities of 0 to 0.15. These breakdowns are based purely on the outcomes shown to give some intuition about the potential for a model to be valuable. For a binomial distribution with p=0.5 and 20 samples, there is a variance of 5. This means that there is a very high degree of uncertainty associated with the mid-range and small probabilities.

In spite of this uncertainty, certain trends can be discussed. One interesting trend is that in spite of very high accuracy, there is a threshold for cost above which a model is rarely valuable. This can be seen in the bottom row of the matrix; the accuracy is high, but the likelihood of a model being valuable drops off significantly as the cost increases, even though the normalized cost is substantially less than the upper bound of 0.4. For the high cost analyses, it was hypothesized that there would be a much clearer trend; at a similar variance to Model A, it is seems that Model B would be much less valuable than a very accurate Model B. This trend is occluded by small sample sizes and the drop off in value resulting from increasing cost on the whole. This is quite different from the case

where the models have similar costs but Model B is more accurate; this case lands quite clearly in the 'Often Valuable' range. It seems that more rigorous study is needed to determine when two similarly accurate models with different costs are both valuable.

While the experimental results in this section are more qualitative in nature than those presented in Sections 5.1 and 5.2, they provide a starting point for understanding what it means for a model to be valuable, and what it means for multiple models to be valuable. Some basic understanding of whether or not it is useful to include a second analysis model based on normalized cost and variance combinations is provided.

5.4 THESIS ROADMAP

This chapter concludes the theoretical experiments and characterization of VGO. In this chapter, suite of test problems was developed and used to compare the effectiveness and efficiency of VGO with respect to EGO. VGO was shown to produce a higher overall utility than EGO with statistical significance. Then, the scalability of VGO was tested by performing experiments on tailored versions of the Hartmann-3 and Hartmann-6 global optimization test problems. Results indicated that VGO could perform within the expected accuracy bounds for high dimensional problems and incur relatively low analysis costs. Finally, the chapter concluded with a discussion of model value in terms of normalized cost-accuracy combinations. Some cost thresholds were provided and validated for single analysis models; no matter how accurate a model is, if analysis cost is considered, then there is a price at which performing an analysis is never valuable. The notion of model dominance in the Pareto sense was discussed, as well as some basic trends for determining when a pair of models is valuable as opposed to only a single model. In the next chapter, VGO is applied to a practical engineering example.

CHAPTER 6: APPLICATION OF VGO: HYDRAULIC HYBRID CAR

In this chapter, the VGO algorithm is applied to a hydraulic hybrid car design problem. The goal in this chapter is to show that VGO can be successfully applied to practical engineering problems and can achieve an acceptable level of accuracy at reasonable costs. The problem setup is described in Section 6.1 and includes a description of the design variables, simulation outputs, and construction of the utility function. Some detail about the modeling of the physical system is given in Section 6.2. The methodology for creating simulations at varying accuracies is given in Section 6.3. This is followed by a discussion of the demand modeling used to form the utility function in Section 6.4 and some details on the setup and implementation in Section 6.5, including assignment of the model costs and accuracies. Some characteristic results and discussion of their significance are provided in Section 6.6. The chapter concludes with a final look at the thesis roadmap.

6.1 BACKGROUND FOR HYDRAULIC HYBRID PROBLEM

Given increasing global energy demands, many automobile manufacturing companies are developing cars that run on a mixture of energy sources or solely on an energy source other than gasoline. Fully electric cars, hybrid gasoline/electric cars, diesel automobiles, and even natural gas vehicles have made their way to the commercial marketplace. They have enjoyed mixed success; gasoline powered vehicles are being made smaller and more efficient and are more competitive in terms of fuel economy than they used to be. The hybrid-electric car, such as the Prius, has enjoyed the most success of the alternative fueled vehicles.

In this problem, the decision maker is a lead product developer for an automobile manufacturer. Rather than trying to duplicate or improve upon the already popular Prius technology, this designer is investigating the profitability of a *hydraulic* hybrid car. While some hydraulic hybrid delivery trucks are currently available, there are no passenger cars currently in this market niche. The basic idea is that the car would have a diesel engine running on diesel that would run at maximum efficiency at all times. Sometimes the engine will drive the car directly; at times when this energy is not needed, the energy will be used to charge up a hydraulic accumulator, the pressure from which can also help to power the drive train. Much like the Prius charges its electric battery while stopped, similarly, this vehicle would build pressure in the accumulator at low speeds and while stopped.

In this problem, it is assumed that three attributes will affect the overall profitability of the vehicle: the fuel economy, the top speed, and the maximum acceleration. These attributes, along with the price at which the vehicle is sold, will affect the *demand* for the vehicle. It is subsequently assumed that demand is a driver for *revenue*, and that this revenue less the cost to produce the vehicle results in profit. It is this profit that is used as a utility function for this problem. Finally, to keep the size of the problem manageable, it is assumed that the design variables for the vehicle are restricted to the engine size, the front pump size, and the rear pump size. These relationships are captured in an influence diagram in Figure 6.1. In the diagram, the pink rectangles indicate design variables, and the green rectangles indicate intermediate outputs determined through modeling and simulation. Market perception is an uncertain variable, indicated by the beige oval.

Finally, utility is the orange hexagon, and is a function of all of the other variables captured in the influence diagram.

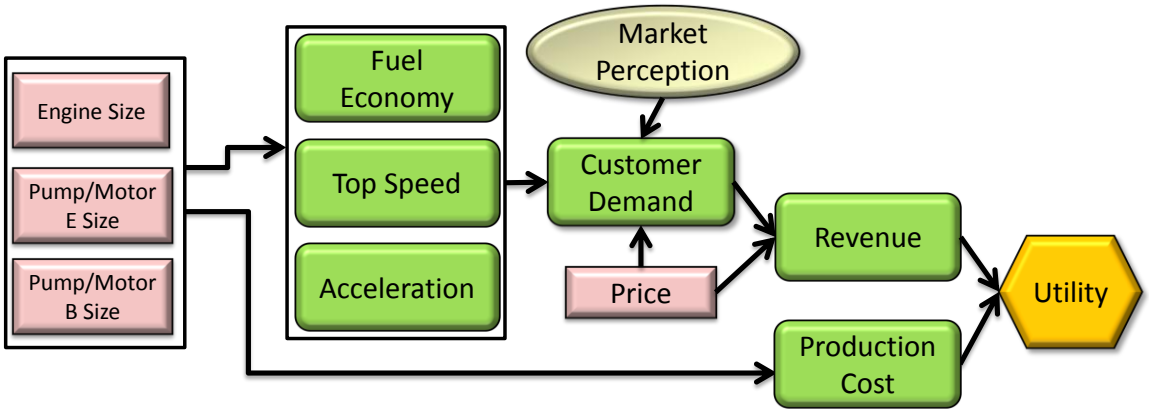

FIGURE 6.1: INFLUENCE DIAGRAM FOR HYDRAULIC HYBRID PROBLEM

The role of the decision maker in this problem is to determine the most profitable combination of design variables to present to the CEO of the company before the vehicle goes to detailed design and manufacturing.

At this stage in the design process, many assumptions have already been made. It has already been assumed that the vehicle will be a hydraulic hybrid, and some architecture has already been assigned in order to simulate the vehicle behavior. Still, the level of uncertainty in the design is very high. It is still assumed to be in the early design stages, and it has not yet been determined if it is even profitable to produce this type of vehicle. Still, to make the best possible decision, it is important to ensure that we have optimized this vehicle. That way, if we compare it to other alternatives, we are comparing the best possible instantiation of this vehicle to other similarly optimized alternatives.

6.2 HYDRAULIC HYBRID MODEL

To simulate the behavior of the hydraulic hybrid vehicle under consideration, a MATLAB model was created. The model is structured in a steady state fashion, even though it is actually a dynamic simulation. To assess the performance of the hydraulic hybrid vehicle, the Urban Dynamometer Driving Schedule (UDDS) is used. The purpose of UDDS is to simulate city driving for light duty or passenger vehicles, and it serves as a benchmark for assessing fuel economy in city conditions. Normally, passenger vehicles have two different fuel ratings, city and highway, so the UDDS would be used to simulate city conditions, as opposed to the highway conditions.

The UDDS is divided into one second time increments with a starting and ending vehicle velocity. After reading in the UDDS, the MATLAB hybrid model back-calculates the state of the vehicle for each time step given the vehicle velocity. This negates the need for using an ODE solver and makes the simulation reasonably inexpensive. As described in the previous section, the model has three input or design variables: diesel engine size (Watts), Pump/Motor B size (m^3/rev), and Pump/Motor E size (m^3/rev). It is assumed that the engine is constantly running at the most efficient power output. The power from the engine is then simply channeled differently depending on the velocity demands of the vehicle. The pumps function both as regular pumps, sending hydraulic oil into an accumulator (charging) when the engine power is not needed, and as motors, allowing oil from the accumulator to power the drive train along with the diesel engine. The model also includes regeneration, whereby the motors pump oil into the accumulator during braking. The overall system architecture is very similar to the one shown in Figure 6.2, as published by Li *et al.* [40].

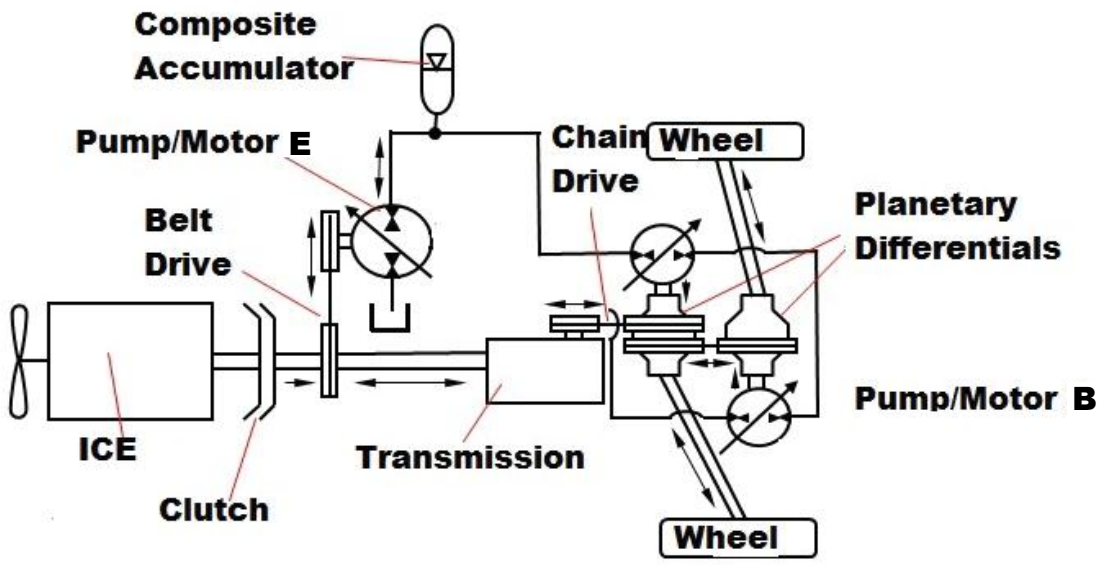

FIGURE 6.2: SCHEMATIC OF HYDRAULIC HYBRID [40]

The MATLAB model of the hybrid vehicle is designed to calculate fuel economy in miles per gallon based on a particular architecture's performance of the UDDS. However, because of the way velocity is constrained, it is necessary to also assess a particular architecture's ability to even perform the drive cycle based on speed and acceleration demands. Sometimes, the component sizes under consideration are too small to achieve the desired velocities, but the vehicle velocity is already constrained. Rather than the model simply failing, what happens is that the demands placed on the pumps during the motoring phase are too high; if the state of the pump is examined, it can be seen that the pump is allowing more oil to pass than it physically can. That is, the pump's swash plate is open more than the full amount. Another way the model can draw power that does not truly exist is to draw power from the accumulator after it is already empty. If either of these conditions is detected, then a penalty is incurred.

The reason for using a penalty function, and specifically, a penalty function that scales with the degree of severity of the violation, is to ensure a meaningful objective function to guide the optimization process. We never want to actually select a vehicle that cannot perform the drive cycle, but if we simply assign a zero or negative constant profit to those vehicles, the design space will exhibit steep, sharp valleys that are flat on the bottom. If the penalty does not reflect improvement and degree of severity, it is hard for the optimizer to find more promising solutions without getting stuck in the deep valleys. In addition, because a kriging-like surrogate model is used during VGO, having a design space that is too rough is hard to model accurately. Thus, intelligent decisions about the structuring of the utility function must be made.

If a particular design artifact is penalized in anyway, then the penalty is used as the utility or profit value for the particular artifact. The penalty is scaled appropriately with the profit function, and this set up allows for the demand modeling portion of the analysis to be skipped if a particular artifact is unpromising. The details of the demand model are given in Section 6.4.

6.3 MODELS AT DIFFERING ACCURACIES

In the previous section, the basic elements of the MATLAB model for the hydraulic hybrid model were described. What is needed for VGO, however, is (at least) two models at different accuracies. Rather than trying to tweak the physics of the MATLAB model, the approach taken in this example is to consider uncertainty in the high fidelity model. For the low fidelity model, the hybrid model described in the previous section is used with the three design variables and all other parameters completely specified, and

the model is deterministic. For the high fidelity model, 18 uncertain inputs are considered.

To construct the high fidelity model, the same base model as the low fidelity model is used in conjunction with consideration of uncertainty. For this model, a triangular distribution is constructed for all 18 uncertain inputs. Most of the uncertain inputs are uncertain due to environmental factors, natural variation, or machine tolerance. For example, some of these inputs include the air density, drag coefficient, and the energy density of diesel fuel. Other uncertain inputs, such as mass, are a function of the size of the components being used, but for the sake of model simplicity is approximated as a reasonably estimated constant. To get a better calculation of expected utility, 100 Monte Carlo samples are run over the uncertain inputs before running the drive train model. This model is therefore, more accurate and more expensive to evaluate than its low fidelity counterpart.

6.4 DEMAND MODELING

For both the low and high fidelity models, a demand model is used to translate from the fuel economy attribute to expected profit for the vehicle. The demand model is implemented in Excel, and is called by MATLAB *only* for cases where there is no penalty assessed. No matter how a demand model is constructed, there is always a high amount of uncertainty. Until a product actually goes to market and consumer behavior can be observed directly, it is difficult to predict the way a product will be received. However, by making some assumptions, it is possible to at least come up with an estimate of demand so that profit can be predicted.

There are two aspects to the demand model: one is to determine the potential market size, and the other is to determine the price at which a product will sell. The market size is estimated using a simple statistic. In 2011, 2% of automobile buyers bought hybrid vehicles, which amounts to 286,620 consumers. This number represents the maximum number of potential consumers regardless of vehicle price. The pricing aspect is bit more complicated. In this thesis, a polling approach is used. For a sample of potential consumers, a list of fuel economies is provided, and the consumers are asked to name the maximum price that they would pay for a vehicle with that fuel economy. The same approach can be used for multiple attributes, but for simplicity, this example is restricted to a single attribute.

To determine the *market share* or the percentage of the market size that will actually go through with purchasing the vehicle, and the pricing strategy, a Bayesian approach is taken. The basic idea is that there is an assumed prior on the market share, which varies between zero and one, that can be characterized as a beta distribution. Then, based on the survey results, a binomial distribution is used to characterize a response 'Yes' or 'No' to the question 'Would you buy this vehicle at this price?'. This response is then used to update the beta distribution to better estimate the expected market share for a particular price. By collecting and interpolating data over a range of fuel economies and a range of prices, a maximization can be performed for the selection of the price that yields the maximum profit, where profit is the market share, times the market size, times the price at which the product is sold, less the costs to produce the product.

Assuming the survey information is constant, the Excel workbook for demand is run as a macro with inputs of fuel economy and the production cost, which is estimated based on

the engine and pump sizes for the particular artifact under consideration, and the outputs are the expected profit and the optimal vehicle price. It is the expected profit which is adopted as the utility function over which the optimization is performed.

6.5 PROBLEM SETUP

So far, the low and high fidelity simulation models to calculate the expected fuel economy have been described. In the previous section, an overview of the demand modeling used to predict the profitability of a particular vehicle was given. Many of the modeling details have been omitted, as they are not vital to the success of this experiment. More important considerations with respect to VGO are the determination of the model inadequacies and model costs in practice. In the previous chapter, when most of the experiments were theoretical in nature, assigning the model inadequacy was often trivial because it was used to generate the model error up front. Here, we are faced with two models and an unknown truth, so it is necessary to now assign model inadequacies. This section is dedicated to discussion the determination of the model inadequacies and subsequently the costs, and how these parameters can affect the experimental results.

6.5.1 Determining Model Inadequacies

In general, a good strategy for determining model inadequacy is to start with the highest fidelity model, and work down to the lowest fidelity. Without knowing the truth or having some experimental validation to assess the accuracy of the highest fidelity model, the best characterization that can be made about the highest fidelity model's accuracy is a good faith estimate. Unfortunately, this is an unscientific answer; if there is experimental data available, then statistical analysis could be used to characterize the accuracy of the highest fidelity model. In the absence of this information, however, the best a decision

maker can do is to estimate. For the hybrid problem, the utility function is taken to be profit in millions, with many predicted outcomes being on the order of billions. In this case, the high fidelity model variance was assigned to be \$700. What this actually means is that $\sigma_2^2 = \$700$, so the standard deviation $\sigma_2 = \$26.45$, and $2\sigma_2 = \$52.92$. However, since the profit is actually in millions, this means that $2\sigma_2 \cong \$53$ million. For an early stage of this design process with anticipated profits in the billions, this is a reasonably accurate prediction.

Determining the lower fidelity model inadequacies can be done a bit more scientifically, in that we can characterize the less accurate models with respect to the most accurate model. For the low fidelity hybrid model, a fast estimate was made by calculating the expected profit for the same design variables with each model. The difference in expected profit (~\$555 M) was set to be equal to $2\sigma_1$. Taking the scaling of the objective into consideration, the resulting variance is $\sigma_1^2 = \$77000$. Clearly, more rigorous characterization could be performed, but as long as the order of magnitude of relative accuracy is captured for the different models, the performance of VGO is not heavily affected by the precise model inadequacy values.

In special cases, particular simulation model might be biased such that the expected error Gaussian process realization would not have a mean of zero. For example, finite element analyses are known to converge from above; therefore, it might be known that this simulation data lies strictly above the truth. In this special case, expert knowledge or a good-faith estimate should be used to determine both upper and lower bounds for the simulation error. That is, a simulation outcome may differ with respect to the truth with a range from 0 to some upper bound, *n*, rather than from –*n* to +*n*. In this case, rather than

using *n* as the assumed model inadequacy (which would imply error of magnitude *n* both above and below the mean), the simulation data should be adjusted in the negative direction by constant *n/2* to make the prediction mean of the error function equal to zero. Then, this model can be assigned a standard deviation of *n/2* instead of *n*. This eliminates the need for overestimating the error bounds, if we can intelligently capture the existence of the bias.

6.5.2 Determining Model Costs

There are many potential ways to assess the cost of a particular analysis model. It is logical that the model cost should scale with computation time; if computation time were not a consideration, then VGO probably would not be used in the first place. Generally, it is possible to estimate an average run time for a model once it is constructed. To calculate cost, this time could be translated to time on cloud or cluster to get from time to cost. However, these costs do not necessarily tell the whole story and might undercut the actual costs of running the model. One cost that should also be considered is the engineer's hourly rate for running the model. Even if a model is fairly automated, an engineer is probably required to keep an eye on the simulation. If a simulation requires the engineer to be in the loop, then an even higher rate should be assigned to the model cost. Another cost to consider is the cost of model construction; while one could argue that these are sunk costs, these costs can be depreciated over the simulations run during the optimization problem. It is also necessary to consider the time sensitivity of the decision; if profit is lost if a decision is not made within a certain time frame, a cost penalty should be assessed accordingly. These are some of the costs that could

theoretically be taken into account when determining the analysis cost for a particular model.

In practice, it is necessary to realize that very cheap models when compared to very profitable objective functions will result in many function evaluations. On the other hand, if a model is too costly, as shown in Section 5.3, it will never be used. For the hybrid problem, time was not accounted for directly, so it was necessary to ensure that the costs assigned led to a solution in a 'reasonable' amount of time. In running the hybrid experiments, a few different costs were used. At first the costs were assigned to be $1000 for a low fidelity analysis and $10 million for a high fidelity analysis. These costs may seem high, but a decision is ensured quickly, and this experiment served as a quick screening result to assess what a good solution might be. After this very quick analysis, less expensive costs were used. One trial was run at $100 and $5 M, and finally more realistic costs of $10 for a low fidelity analysis and $1 M for a high fidelity analysis were used. Intuitively, the relative cost should bear some resemblance to the relative accuracy, but as was shown in the experiment in Section 5.3, it is not always easy to tell on sight when a particular cost-accuracy combination is valuable.

6.6 RESULTS

The hydraulic hybrid optimization problem was solved three times using the VGO algorithm. For each trial, an initial LHS sample of 100 low fidelity analyses was used. The primary difference between the trials is the assumed cost of the models. The first trial used the highest costs, and the final trial employed the lowest costs. The results from the first trial are captured in Table 6.1. The table entries are as follows: the assumed model costs, low and high fidelity; the assumed variances, low and high fidelity;

the number of initial low fidelity samples; the total number of low fidelity analyses performed during the optimization, including the initial sample; the total number of high fidelity analyses performed during the optimization; the resulting design variables, engine size (W), Pump E size (m^3/rev), and Pump B size (m^3/rev); the predicted profit in millions of dollars from the final surrogate model fit; the predicted profit in millions of the high fidelity analysis for the resulting design variables; the cost of analysis incurred in millions of dollars; the final overall utility, or high fidelity profit less the analysis costs, again in millions of dollars.

TABLE 6.1. TRIAL 1 HYDRAULIC HYBRID RESULTS

Hybrid Experiment 1	
Costs (in millions):	[.01; 10]
Variances (in millions):	[77000; 700]
Initial Samples:	100
Low Fidelity Analyses:	110
High Fidelity Analyses:	14
Predicted Maximum (engine size, Pump E, Pump B):	[28230, 2.7473e-5, 2.8132e-5]
Predicted Profit from Surrogate (in millions):	2132.4
High Fidelity Profit at Predicted Maximum (in millions):	2107.9
Analysis Cost (in millions):	141.1
Utility (in millions):	1966.8

In these experiments, there is no known truth, nor is the true maximum profit known. In order to characterize the solution quality, the high fidelity profit, which is assumed to be the artifact utility, can be compared to the best solution from an optimization of the high fidelity model. The result found by VGO is $2,107,900,000. The predicted profit is within the expected accuracy bounds. In this trial, the high fidelity model is expensive to evaluate; the error of the artifact utility is exceeded by the analysis costs incurred for this

trial. By setting the analysis cost high, it is ensured that the algorithm does not take too many steps. This trial is not intended to produce a high quality result, but rather a benchmark. Better results are achieved in the next trials.

The results for the second hydraulic hybrid optimization are shown in Table 6.2.

TABLE 6.2. TRIAL 2 HYDRAULIC HYBRID RESULTS

Hybrid Experiment 2	
Costs (in millions):	[.001, 5]
Variances (in millions):	[77000; 700]
Initial Samples:	100
Low Fidelity Analyses:	105
High Fidelity Analyses:	9
Predicted Maximum (engine size, Pump E, Pump B):	[27319, 2.8596e-5, 2.849e-5]
Predicted Profit from Surrogate (in millions):	2095.6
High Fidelity Profit at Predicted Maximum (in millions):	2108.3
Analysis Cost (in millions):	25.105
Utility (in millions):	2083.195

In this trial, the model costs are lower, and interestingly, few analyses are used. The analysis cost incurred in this trial is only 17% of what it was in the previous trial. In terms of solution quality, the artifact utility is $2,108,300,000, which is slightly better than the result achieved previously, but the overall utility for this trial, taking analysis costs into account, is much higher.

The final trial results are shown in Table 6.3.

TABLE 6.3. TRIAL 3 HYDRAULIC HYBRID RESULTS

Hybrid Experiment 3	
Costs (in millions):	[.0001, 1]
Variances (in millions):	[77000; 700]
Initial Samples:	100
Low Fidelity Analyses:	102

High Fidelity Analyses:	9
Predicted Maximum (engine size, Pump E, Pump B):	[27803, 2.9655e-5, 2.543e-5]
Predicted Profit from Surrogate (in millions):	2047.2
High Fidelity Profit at Predicted Maximum (in millions):	2048.9
Analysis Cost (in millions):	0.0102
Utility (in millions):	2048.8898

There is an inherent stochastic nature to the algorithm, so the fact that the artifact utility is not as high for this trial as in the previous trial is likely an issue of bad luck. The costs for analysis were probably the most realistic for this trial, but the algorithm still stopped after relatively few iterations. For all three trials, the overall utility achieved was similar.

The goal in this chapter was to show that VGO could be successfully applied to practical engineering problems. The trials shown illustrate that this is in fact the case. The accuracy of the optimization is reasonably good considering the model inadequacies given. By combining global search and local refinement, these solutions are achieved with relatively few function evaluations. As a comparison, if a single gradient optimization were run (which is impractical, a multi-start would be needed in this dimensionality), 25 steps would not be unreasonable. With four function evaluations per step, one for the function value and three for the gradient, that would be 100 function evaluations. With VGO, the whole optimization with global coverage is completed with approximately the same number of function evaluations.

As far as the manager's decision is concerned, the best design artifact had a fuel economy of about 28.5 mpg, with a suggested vehicle price of $35,280. The demand model may be inaccurate, but tuning the demand model would not affect the predicted fuel economy, nor would it affect the applicability of VGO. With those results, the manager may want

to consider investing in compact, high fuel economy, classic gasoline cars, where higher gas mileage might be achieved with a lower associated price tag.

6.7 THESIS ROADMAP

This focus of this chapter was the applicability of VGO to a practical, engineering design example. Results showed that VGO could in fact, be successfully applied and achieve good solutions at very reasonable costs. The accuracy of the solutions achieved were in line with the assumed accuracy of the models. This chapter also contained discussion about determining model accuracy and costs in practice. The next and final chapter provides a summary and critical review of the work presented in this thesis.

CHAPTER 7: CONCLUSION

This thesis concludes with a summary of the contributions and a critique of the research presented. The purpose of this thesis is to provide the theoretical foundations for a new global optimization suite known as Value-Based Global Optimization. A review of the algorithm and its characterization as presented up to this point is provided in Section 7.1. The research questions and hypotheses are revisited in Section 7.2, with a complete discussion of the hypothesis validity given the theoretical contributions of VGO and the characterizations provided in Chapters 5 and 6. The contributions of the thesis are then summarized in Section 7.3. Finally, the thesis is concluded with suggestions for future work in Section 7.4 and some closing remarks in Section 7.5.

7.1 A SUMMARY OF THIS THESIS

During the engineering design process, it is often necessary for a decision maker to rely on simulation data from various system models. These models enable a designer to make more informed decisions when uncertainty is high and physical prototypes are too costly. While computational models and experiments can greatly reduce the cost of analysis over a physical experiment, even simulation models can be too costly to perform exhaustive searches over the range of possible alternatives and to perform optimizations. Often, designers are forced to choose between exhaustive global search with inexpensive, less accurate models and accurate assessment of only a few design alternatives with more costly models. Global optimization suites that make use of different computational cost reduction techniques can aid designers in achieving good solutions at a reasonable cost. The current state of the art for design problems and cost reduction techniques are

reviewed in Chapter 2, and global optimization algorithms targeted toward this type of problem are reviewed in Section 2.5. It is shown that many of the current cost reduction techniques are not specifically targeted toward optimization, and that many of the current global optimization algorithms are not ideal for engineering design problems, where the objective is to find good solutions at reasonable cost.

In Chapter 3, the VGO global optimization algorithm is presented. VGO makes use of surrogate modeling, multi-accuracy modeling, and sequential sampling to enable users to find accurate solutions in an efficient manner. It is argued that using models of different accuracies allows for more efficient global optimization by channeling the computational resources toward promising solutions; global exploration can be performed using the less accurate, inexpensive models and local refinement can be performed with the more accurate, more expensive models. In this way, global search and local refinement can both be achieved at reasonable cost without sacrificing solution quality.

There are two primary aspects of VGO that distinguish it from similar algorithms. The first is a novel surrogate modeling technique that allows for data from any number of models at different accuracies with no restrictions on the correlation between the different models. The second contribution is the use of Value of Information as a sequential sampling strategy. While VoI is a known entity from decision theory, it has not been previously applied in this context. Value of Information allows the designer to choose both the next design site and analysis in a way that is cost-effective. This sampling criterion naturally balances global search and local refinement and indicates if an analysis is too costly to justify. There is a very natural stopping criterion—stop sampling when it

is no longer *valuable* to do so, i.e. when the potential benefit of analysis is outweighed by the costs incurred.

While Chapter 3 is focused on the conceptual approach for the VGO algorithm, the theoretical foundations for this algorithm are then delineated in Chapter 4. In Chapter 4, the mathematical formulation for the surrogate modeling technique is derived. The surrogate modeling approach derived in this thesis is based largely on kriging modeling. However, rather than resulting in an interpolator, the mathematical formulation is tailored to result in a weighted fit based on the accuracy of the samples to which the surrogate is fit. In addition to the surrogate model, the theory for VoI is also described, along with some implementation details for the algorithm.

In Chapter 5, several experimental results are presented with the intention of characterizing the performance of the VGO algorithm. In particular, VGO is rigorously compared to Efficient Global Optimization (EGO), which is considered to be the current state of the art. It was shown that VGO outperformed EGO on suites of test problems with statistical significance when accuracy and cost were both taken into consideration. That is, VGO found equally good or better solutions at a lower analysis cost. In addition to the EGO comparison, VGO is tested for scalability to higher dimensional search spaces using known optimization test functions. It is shown that VGO is capable of finding good solutions within the accuracy bounds of the available models while using relatively few costly function evaluations. Finally, Chapter 5 concludes with a qualitative discussion of model value. In particular, boundaries on the acceptable costliness individual models are established. In addition, the value of adding a second model to the available models given the cost and accuracy of the first model is discussed qualitatively.

Finally, in Chapter 6, VGO is used to solve an engineering design problem on a hydraulic hybrid vehicle. The applicability of VGO to practical problems is illustrated, and more of the subtleties of using VGO in practice are discussed. Two different simulation models are used to run the algorithm and solve the problem, one where the model is assumed to be deterministic (low fidelity) and one where uncertainty is taken into account (high fidelity). Three different optimization trials are run, and the results show that VGO finds solutions within the expected accuracy bounds with relatively few function evaluations. While the total number of function evaluations is low considering the problem's dimensionality, the number of high fidelity function evaluations is very low. The end result is that the analysis costs are lower than could be achieved using only a single model or by using a gradient search method.

7.2 REVISITING THE RESEARCH QUESTIONS AND HYPOTHESES

From Chapter 1, the primary research question for this thesis is:

How can designers perform design optimizations at a reasonable cost without sacrificing solution quality?

The hypothesis is that a Value-Based Global Optimization (VGO) algorithm will allow designers to achieve good solutions (design artifacts) at better costs than can be achieved with comparable existing algorithms. This hypothesis is validated conceptually in Chapter 3, where it is reasoned that using models at different accuracies in conjunction with a value-based search criterion allows for designers to assess the quality of good solutions with high accuracy while still reducing uncertainty in the global search space by using lesser quality models. This hypothesis is then validated experimentally in Chapter

5, where it is shown with statistical significance that VGO outperforms EGO on a suite of global optimization test problems. The results of the experiments show that even in the cases where VGO and EGO achieve similar quality solutions, VGO does so with much lower analysis costs.

The first secondary research question pertains specifically to adapting surrogate modeling techniques to multi-accuracy data:

How can data from multiple models of varying levels of accuracy be used advantageously during the design optimization process?

The hypothesis is that a Gaussian process-based surrogate model, similar to a kriging model, can be derived mathematically to accommodate multi-accuracy observations from any number of different models. This approach is conceptually validated in Section 3.4, and the mathematical derivation is provided in Section 4.1. The result is a surrogate modeling technique that retains many of the desirable properties of the standard kriging modeling formulation while taking into account the accuracies of the individual samples to which the model is fit. Illustrations of the surrogate model are shown in Section 4.6.

The second secondary research question is related to the sequential sampling technique employed during VGO:

How can the most valuable design site and analysis be dynamically selected at each step in the optimization process?

The hypothesis is that maximizing the Value of Information (VoI) provides a metric for choosing the next design site and associated analysis model at each step in the optimization process. The reasoning for using a value-based sequential sampling method is provided in Chapter 3, and the theoretical foundations of VoI are described in Chapter 4. It can be seen from the sample results in Section 4.6 and any of the experimental results shown in Chapters 5 and 6 that VoI does indeed provide a meaningful sampling criterion for choosing both design sites and analyses during the optimization process. Moreover, VoI is naturally cost-effective because model accuracy and cost are taken into account explicitly, which makes it an ideal, intuitive metric for engineering design. By performing sequential sampling using VoI, low analysis costs are achieved for high quality solutions as compared to algorithms using the Expected Improvement (EI) metric for sequential sampling.

7.3 CONTRIBUTIONS

The primary overarching contribution of this thesis is the development and characterization of Value-Based Global Optimization, a global optimization algorithm intended for simulation-based engineering design problems. This algorithm combines a new surrogate modeling technique with a novel sequential sampling technique. These new techniques allow VGO to channel the available computational resources in a way that is efficient and effective in an optimization context. It is based on the premise that engineers are not concerned with mathematical optimality, but rather on finding good solutions at reasonable cost in a reasonable amount of time. By allowing for multiple computational models to be used during the optimization process and focusing on the value of each individual analysis, solution quality comparable to other algorithms can be

achieved at only a fraction of the analysis cost. In this thesis, the VGO algorithm is developed at a conceptual and theoretical level. It is also tested and compared to similar algorithms in order to characterize its performance.

One of the main components of the VGO algorithm is the Gaussian process-based surrogate modeling technique for multi-accuracy data, which is a significant research contribution on its own. While multi-accuracy modeling and surrogate modeling are both relatively mature fields, combining the two is still fairly uncommon. In addition, very few existing algorithms allow for data from more than two different models without any restrictions about the higher fidelity design sites being a subset of the lower fidelity design sites. When comparing the surrogate modeling technique from VGO to existing surrogate modeling techniques for multi-accuracy data, the one used in VGO allows for the fewest assumptions about model correlations and the smallest number of hyperparameters that must be estimated by maximizing likelihood. That is, by assuming that the model error terms are independent Gaussian process realizations, there is no resulting correlation between the various models even if they exhibit similar behaviors. The resulting surrogate model retains many desirable properties of kriging models, rooted in its statistical nature and its simple calculation for determining the expected mean squared error. However, rather than being an interpolator, the resulting surrogate surface is weighted by the accuracy of the samples to which it is fit.

The surrogate surface modeling technique presented in this thesis is a significant research contribution not only because it is different from any other available surrogate modeling technique, but because it is targeted toward optimization in an engineering context. This surrogate modeling technique is not only fit to multi-accuracy data, but is a variable

accuracy modeling technique unto itself. This surrogate allows for high accuracy regions in promising areas and low accuracy regions where the solutions are obviously poor. This is precisely what is desired for performing accurate optimizations at reasonable costs—the surface is accurate enough to guide the optimizer toward better solutions but reserves the best accuracy for areas that could lead to the final solution.

A second independent research contribution stemming from VGO is the application of Value of Information from decision theory to sequential sampling during global optimization. While VoI is an established metric, it has never before been applied as a sequential sampling metric for surrogate modeling and optimization. In this thesis, VoI is introduced in this new domain and the individual quantities needed to calculate VoI are all defined for use in the context of VGO. Several derivations are required to compute VoI correctly in the context of VGO, including the derivation of the simulation mean and variance, as well as the posterior on the truth estimate. All of the necessary calculations to use VoI as a sequential sampling metric are provided as a contribution of this thesis. In addition, a novel characteristic of VoI is an intuitive stopping criterion that is not dependent on a user-defined constant. This makes the VGO algorithm more robust to user input in terms of the resulting solution quality.

The remaining research contributions presented in this thesis are in some way related to the characterization of VGO and illustrating its capabilities. In Chapter 5, a rigorous performance comparison of VGO and EGO is presented. There are two research contributions in this experiment. The first contribution is the test suite used to perform the experiments. The suite is specifically designed to emulate engineering design problems involving multiple analysis models, an unknown truth objective function, and

evaluation costs. The generation technique for the test suite provides a novel approach for generating infinite possible test cases with different surface features and for providing any number of analysis models to simulate the unknown truth. The second contribution in this performance characterization is the results and analysis of the results. VGO is shown to outperform EGO, the current state of the art, with statistical significance by using a sign test to interpret the results.

In Section 5.2 and in Chapter 6, VGO is shown to be applicable to high dimensional problems and practical engineering problems. This is the first time VGO has been applied to such test problems. These test cases function as a first step toward illustrating all of the capabilities of VGO. While no direct comparisons to other algorithms are made in these example problems, the number of function evaluations required to run the optimizations is small enough that meaningful discussions can be provided about the cost of analyses incurred. In Section 5.2, the solutions of the test problems are known, so it is not difficult to assess the accuracy of the solutions obtained. In Chapter 6, the truth is not known, but repeat trials show that good solutions can be achieved at a reasonable cost on practical examples.

The final contribution of this thesis is a first step toward defining a meaningful set of analysis models, as shown in Section 5.3. While this discussion is very qualitative in nature, this is not a topic that is typically broached in the existing literature. The benefit of using multiple models with different cost-accuracy combinations is generally accepted in the literature, and different model construction methods are discussed. However, what is not often discussed is how many models should be used and what combination of cost-accuracy combinations is most efficient. While VoI will naturally 'filter' out models that

are not useful, it can save some additional calculation time of maximizing VoI if an efficient set of models is selected before VGO is run. As variable accuracy modeling continues to gain momentum in the optimization domain, it will be necessary to continue making research contributions in this area.

7.4 LIMITATIONS AND FUTURE WORK

In the previous section, the many research contributions of VGO and this thesis were enumerated. The algorithm has been shown to outperform the previous state of the art in global optimization techniques for engineering design optimization problems. However, there are certain limitations of VGO that warrant further discussion.

A very common limitation of variable accuracy modeling techniques in general, which is also a limitation of VGO, is the assumed parameterization of the available models. Specifically, it is assumed that the inputs and outputs of the various models are in alignment. In order to fit surrogate surfaces and run optimizations using multi-accuracy data, the models must map the same inputs to the same outputs. This is not always practical in real engineering examples; often, models of different fidelities are at different abstractions. It is likely that the inaccurate models have fewer, more general inputs while the more accurate models exhibit more detailed inputs and parameters. It is also assumed that the output space is the same; in VGO, it assumed that all model outputs can be mapped to utility or a profit function. There are not currently any algorithms for multi-accuracy modeling that can accommodate models of different abstractions that rely on different sets of inputs and outputs. This is an area where further research is required.

One limitation of VGO has to do with the assumptions made in developing the algorithm. The derivation of the surrogate surface presented in this thesis assumes no correlation between model error and the truth. While this assumption simplifies the derivation and the number of hyperparameters which must be estimated, it would be useful to further assess the validity of this assumption and its effect on performance. Clearly, there could be cases where the error of a particular model is correlated with the truth, particularly if the model is a low frequency approximation of a high frequency phenomenon. These cases need to be investigated explicitly.

Another limitation of VGO has to do with the cost of the algorithm itself. While the costs of the individual analyses are modeled explicitly during VGO, the cost of the surrogate model fitting and the cost of optimizing the Value of Information are ignored. There are opportunities here both for reducing the costs associated with the algorithm and modeling them explicitly. For example, the cost of maximizing VoI could be included with the individual analysis cost so that more intelligent cost-benefit tradeoffs could be made. If the costs of the algorithm can be modeled and accounted for, better decisions can be made in the context of optimization.

Beyond modeling the algorithm costs incurred, there is still room for improvement in the implementation of the algorithm. While the version of VGO used in this thesis contains many efficiency tactics, there is still room for improving the speed and cost of several calculations. In particular, if the cost associated with maximizing VoI could be reduced, the algorithm would be much more effective. Currently, VoI is maximized using a multi-start gradient optimization with a relatively large number of starting points. Because the VoI calculation involves calculating the posterior on the truth, which is effectively a new

surrogate surface less the likelihood maximization, these maximizations are fairly expensive. This is one area where expected improvement is more attractive; the calculation of EI is very inexpensive. While VGO is conceptually superior to EGO in cases where multiple analysis models are available, it is less attractive if the costs of the analyses are low. In this thesis it is assumed that the cost of analysis is much more significant than the cost of the algorithm; if this assumption is violated then there are cases where EGO is more efficient in terms of total computation time.

There is still room for improvement in describing a method for determining the costs and accuracies of the individual models in VGO. Recall that these are user-supplied inputs so that the models can be correctly characterized by their accuracy and their value assessed. Particularly in the hybrid example shown in Chapter 6, meaningful determination of these parameters was not trivial and had an effect on the performance of the algorithm. Assuming higher model costs causes the algorithm to terminate sooner, while less expensive models result in a longer runtime. If the costs and accuracies are not determined in a meaningful way, the desired results may not be achieved.

Along these same lines, further research is needed on what constitutes an effective set of analysis models. While some discussion was provided in this thesis about the value of models in terms of cost-accuracy combinations, much more rigorous experimentation is needed to show results with statistical significance. This is a research area that has largely been ignored in the past. The use of variable accuracy modeling is gaining in popularity, and there has been discussion on basic ways of constructing such models. However, the cost-accuracy characterization has not been considered explicitly in the literature, and a desirable number of models has not yet been established.

In this thesis, the VGO algorithm was subjected to some performance testing applied to problems with as many as six dimensions. There is still room for additional performance characterization, and even higher dimensional problems should be tested in the future. Problems up to 10 dimensions with more analysis models and a greater variety of cost-accuracy combinations are needed to continue to characterize the performance of VGO. Larger test suites with randomly generated model parameters could aid in a more rigorous comparison of VGO and EGO. Beyond theoretical examples, VGO should be run on a greater variety of practical engineering examples.

A final area of suggested improvement which will enable many of the other recommended tests and improvements to be implemented more readily is an open-source implementation of VGO that is publicly available. If VGO is readily accessible, more practitioners can run tests and make improvements to the algorithm. There will naturally be more practical examples and resources available if the VGO code is freely available.

7.5 CLOSING REMARKS

VGO is a global optimization algorithm that builds on novel surrogate modeling and sequential sampling techniques in order to solve engineering design problems with high accuracy at reasonable cost. Several currently available optimization techniques employ surrogate modeling in conjunction with sequential sampling, but VGO takes these techniques a step further by allowing models at different accuracies and by explicitly accounting for analysis costs incurred. Results show that VGO can achieve high quality solutions with low analysis cost by strategically selecting the analyses dynamically during the optimization process. By using a value-based sequential sampling metric, cost effective trade-offs can be made with minimal user input.

The contributions of VGO will affect the state of the art for global optimization algorithms in a design context. By shifting the focus from true global optimality to a more practical perspective of good solutions at reasonable cost, optimization algorithms can be tailored more appropriately to meet designers' needs. Looking beyond the optimization aspects of the algorithm, the surrogate modeling technique presented in this thesis is likely to have an impact on the use of surrogates in conjunction with multi-accuracy data. By providing a mathematical formulation that is similar enough to classic kriging, it may encourage more users to consider expanding from optimizations and analyses using only one underlying simulation to multiple simulations. This shift will save a lot of time and resources in the context of many future design decisions.

APPENDIX A: DERIVATION OF SIMULATION PRIOR VARIANCE

$$\sigma_{si}^2 = \mathrm{E}[((y - \hat{y}) + (z_i - \hat{z_i}))^2]$$

$$y(x) = f^{\mathrm{T}}\beta + z$$

$$\hat{y}(x) = c^{\mathrm{T}}Y_s = c^{\mathrm{T}}(F^T\beta + \mathrm{Z} + \mathrm{Z_m})$$

$$y_i = y + z_i$$

$$\hat{z_i} = \hat{y}_i - \hat{y} = r_i{}^{\mathrm{T}}\gamma_i^* - \hat{y}$$

$$\gamma_i^* = R_i^{-1}\big(Y_{Si} - \hat{y}(s_i)\big)$$

$$\sigma_{si}^2 = \mathrm{E}\left[\left(\underbrace{(f^{\mathrm{T}}\beta + z - c^{\mathrm{T}}(F^T\beta + \mathrm{Z} + \mathrm{Z_m}))}_{A} + \underbrace{\left(z_i - r_i{}^{\mathrm{T}}R_i^{-1}(I - \tilde{C})Y_s\right)}_{B}\right)^2\right]$$

$$\tilde{C} = \begin{bmatrix} c^{\mathrm{T}}(s_{i1}) \\ \vdots \\ c^{\mathrm{T}}(s_{in_q}) \end{bmatrix}$$

$$\sigma_{si}^2 = \mathrm{E}[A^2 + 2AB + B^2]$$

Recall that $F^{\mathrm{T}}c - f = 0$

$$A = -c^{\mathrm{T}}(Z + \mathrm{Z_m}) + z$$

$$A^2 = c^{\mathrm{T}}\big(Z^2 + 2\mathrm{Z}\mathrm{Z_m} + \mathrm{Z_m}{}^2\big)c - 2c^{\mathrm{T}}(Z + \mathrm{Z_m})z + z^2$$

$$\mathrm{E}[A^2] = \mathrm{E}[c^{\mathrm{T}}ZZ^{\mathrm{T}}c + 2c^{\mathrm{T}}\mathrm{Z}\mathrm{Z_m}c + c^{\mathrm{T}}\mathrm{Z_m}\mathrm{Z_m}c - 2c^{\mathrm{T}}(Z + \mathrm{Z_m})z + z^2]$$

$$\mathrm{E}[A^2] = \sigma^2 c^{\mathrm{T}} R c + \sigma^2 c^{\mathrm{T}} R_i c + \sigma^2$$

$$2AB = -2(c^{\mathrm{T}}(Z + \mathrm{Z_m}) + z)\left(z_i - r_i{}^{\mathrm{T}} R_i^{-1}(I - \tilde{C})Y_s\right)$$

$$\mathrm{E}[2AB] = \mathrm{E}\left[2(-c^{\mathrm{T}}(Z + \mathrm{Z_m}) + z)\left(z_i - r_i{}^{\mathrm{T}} R_i^{-1}(I - \tilde{C})Y_s\right)\right]$$

Let $b_i = r_i^{\mathrm{T}} R_i^{-1}(I - \tilde{C})$

$$\mathrm{E}[2AB] = \mathrm{E}\left[2(-c^{\mathrm{T}}(Z + \mathrm{Z_m}) + z)\left(z_i - b_i{}^{\mathrm{T}} Y_s\right)\right]$$

$$\mathrm{E}[2AB] = \mathrm{E}\left[-2c^{\mathrm{T}} Z z_i - 2c^{\mathrm{T}} \mathrm{Z_m} z_i + 2z z_i + 2c^{\mathrm{T}} Z Y_S{}^{\mathrm{T}} b_i + 2c^{\mathrm{T}} \mathrm{Z_m} Y_S{}^{\mathrm{T}} b_i - 2z Y_S{}^{\mathrm{T}} b_i\right]$$

$$\mathrm{E}[2AB] = -2c^{\mathrm{T}}\sigma^2 r - 2c^{\mathrm{T}}\sigma_i^2 r_i + 2c^{\mathrm{T}}\sigma^2 R b_i + 2c^{\mathrm{T}}\sigma_{\mathrm{i}}^2 R_i b_i - 2\sigma^2 r^{\mathrm{T}} b_i$$

$$B^2 = z_i{}^2 - 2z_i r_i{}^{\mathrm{T}} R_i^{-1}\left(I - \tilde{C}\right)Y_s + \left(r_i{}^{\mathrm{T}} R_i^{-1}\left(I - \tilde{C}\right)Y_s\right)^2$$

$$B^2 = z_i{}^2 - 2z_i Y_S{}^{\mathrm{T}} b_i + b_i{}^{\mathrm{T}} Y_S Y_S{}^{\mathrm{T}} b_i$$

$$\mathrm{E}[B^2] = \mathrm{E}[z_i{}^2 - 2z_i Y_S{}^{\mathrm{T}} b_i + b_i{}^{\mathrm{T}} Y_S Y_S{}^{\mathrm{T}} b_i]$$

$$\mathrm{E}[B^2] = \sigma_i^2 - 2\sigma_i^2 r_i{}^{\mathrm{T}} b_i + b_i{}^{\mathrm{T}}(\sigma^2 R + \sigma_i^2 R_i) b_i$$

$$\sigma_{si}^2 = \sigma^2 c^{\mathrm{T}} R c + \sigma^2 c^{\mathrm{T}} R_i c + \sigma^2 - 2c^{\mathrm{T}}\sigma^2 r - 2c^{\mathrm{T}}\sigma_i^2 r_i + 2c^{\mathrm{T}}\sigma^2 R b_i + 2c^{\mathrm{T}}\sigma_{\mathrm{i}}^2 R_i b_i$$
$$- 2\sigma^2 r^{\mathrm{T}} b_i + \sigma_i^2 - 2\sigma_i^2 r_i{}^{\mathrm{T}} b_i + b_i{}^{\mathrm{T}}(\sigma^2 R + \sigma_i^2 R_i) b_i$$

Let $K = \sigma^2 R + \sigma_i^2 R_i$

$$\sigma_{si}^2 = c^{\mathrm{T}} K c + \sigma^2 - 2c^{\mathrm{T}}\sigma^2 r - 2c^{\mathrm{T}}\sigma_i^2 r_i + 2c^{\mathrm{T}} K b_i - 2\sigma^2 r^{\mathrm{T}} b_i + \sigma_i^2 - 2\sigma_i^2 r_i{}^{\mathrm{T}} b_i + b_i{}^{\mathrm{T}} K b_i$$

$$\sigma_{si}^2 = \sigma^2 + \sigma_i^2 + (c + b_i)^{\mathrm{T}} K (c + b_i) - 2\sigma^2 (c + b_i)^{\mathrm{T}} r - 2\sigma_i^2 (c + b_i)^{\mathrm{T}} r_i$$

APPENDIX B: DERIVATION OF POSTERIOR MEAN OF TRUTH

$$a = \underline{r}'\underline{R}^{-1}\underline{Y} - (\underline{F}'\underline{R}^{-1}\underline{r} - \underline{f})'(\underline{F}'\underline{R}^{-1}\underline{F})^{-1}\underline{F}'\underline{R}^{-1}\underline{Y}$$

$$\left[\underline{R}^{-1}\right]_{(m+1)\times(m+1)}$$

$$\underline{R} = \begin{bmatrix} R & \overline{r} \\ \overline{r}' & \overline{r}_{m+1} \end{bmatrix}$$

$$\left[\underline{Y}\right]_{(m+1)\times 1}$$

$$\left[\underline{r}\right]_{(m+1)\times 1}$$

$$\left[\underline{F}\right]_{(m+1)\times p}$$

$$\left[\underline{f}\right]_{p\times 1}$$

$$\left[\overline{r}\right]_{m\times 1} = B = \sigma^2 r(x) + \sigma_i^2 r^i(x)$$

$$\overline{r}_{m+1} = D \,(\text{scalar})$$

$$\delta = \overline{r}_{m+1} - \overline{r}^{-1}R^{-1}\overline{r} = \overline{r}_{m+1} - \tilde{\overline{r}}'\tilde{\overline{r}}$$

First term:

$$\begin{bmatrix} r' & r_{m+1} \end{bmatrix} \begin{bmatrix} R^{-1} + \dfrac{R^{-1}\bar{r}\bar{r}R^{-1}}{\delta} & -\dfrac{R^{-1}\bar{r}}{\delta} \\ \dfrac{-\bar{r}'R^{-1}}{\delta} & \dfrac{1}{\delta} \end{bmatrix} \begin{bmatrix} Y \\ y_{m+1} \end{bmatrix}$$

$$= \begin{bmatrix} r'R^{-1} + \dfrac{r'R^{-1}\bar{r}\bar{r}R^{-1}}{\delta} - \dfrac{r_{m+1}\bar{r}'R^{-1}}{\delta} & -\dfrac{r'R^{-1}\bar{r}}{\delta} + \dfrac{r_{m+1}}{\delta} \end{bmatrix} \begin{bmatrix} Y \\ y_{m+1} \end{bmatrix}$$

$$= \underbrace{r'R^{-1}Y + \frac{r'R^{-1}\bar{r}\bar{r}R^{-1}Y}{\delta} - \frac{r_{m+1}\bar{r}'R^{-1}Y}{\delta}}_{a1} + \underbrace{\frac{\left(r_{m+1} - r'R^{-1}\bar{r}\right)y_{m+1}}{\delta}}_{a2}$$

$$S1 = \underline{F}'\underline{R}^{-1}\underline{r} - \underline{f}$$

$$= \begin{bmatrix} F' & 1 \end{bmatrix} \begin{bmatrix} R^{-1} + \dfrac{R^{-1}\bar{r}\bar{r}R^{-1}}{\delta} & -\dfrac{R^{-1}\bar{r}}{\delta} \\ \dfrac{-\bar{r}'R^{-1}}{\delta} & \dfrac{1}{\delta} \end{bmatrix} \begin{bmatrix} r \\ r_{m+1} \end{bmatrix} - 1$$

$$= \begin{bmatrix} F'R^{-1} + \dfrac{F'R^{-1}\bar{r}\bar{r}R^{-1}}{\delta} - \dfrac{\bar{r}'R^{-1}}{\delta} & -\dfrac{F'R^{-1}\bar{r}}{\delta} + \dfrac{1}{\delta} \end{bmatrix} \begin{bmatrix} r \\ r_{m+1} \end{bmatrix} - 1$$

$$= F'R^{-1}r + \frac{F'R^{-1}\bar{r}\bar{r}R^{-1}r}{\delta} - \frac{\bar{r}'R^{-1}r}{\delta} + \frac{\left(1 - F'R^{-1}\bar{r}\right)r_{m+1}}{\delta} - 1$$

$$S2 = \underline{F}'\underline{R}^{-1}\underline{F}$$

$$= \begin{bmatrix} F' & 1 \end{bmatrix} \begin{bmatrix} R^{-1} + \dfrac{R^{-1}\bar{r}\bar{r}R^{-1}}{\delta} & -\dfrac{R^{-1}\bar{r}}{\delta} \\ \dfrac{-\bar{r}'R^{-1}}{\delta} & \dfrac{1}{\delta} \end{bmatrix} \begin{bmatrix} F \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} F'R^{-1} + \dfrac{F'R^{-1}\bar{r}\bar{r}R^{-1}}{\delta} - \dfrac{\bar{r}'R^{-1}}{\delta} & -\dfrac{F'R^{-1}\bar{r}}{\delta} + \dfrac{1}{\delta} \end{bmatrix} \begin{bmatrix} F \\ 1 \end{bmatrix}$$

$$= F'R^{-1}F + \frac{F'R^{-1}\bar{r}\bar{r}R^{-1}F}{\delta} - \frac{\bar{r}'R^{-1}F}{\delta} + \frac{\left(1 - F'R^{-1}\bar{r}\right)}{\delta}$$

$$\underline{F}'\underline{R}^{-1}\underline{Y}$$

$$= \begin{bmatrix} F' & 1 \end{bmatrix} \begin{bmatrix} R^{-1} + \dfrac{R^{-1}\bar{r}\bar{r}R^{-1}}{\delta} & -\dfrac{R^{-1}\bar{r}}{\delta} \\ \dfrac{-\bar{r}'R^{-1}}{\delta} & \dfrac{1}{\delta} \end{bmatrix} \begin{bmatrix} Y \\ y_{m+1} \end{bmatrix}$$

$$= \begin{bmatrix} F'R^{-1} + \dfrac{F'R^{-1}\bar{r}\bar{r}R^{-1}}{\delta} - \dfrac{\bar{r}'R^{-1}}{\delta} & -\dfrac{F'R^{-1}\bar{r}}{\delta} + \dfrac{1}{\delta} \end{bmatrix} \begin{bmatrix} Y \\ y_{m+1} \end{bmatrix}$$

$$= F'R^{-1}Y + \frac{F'R^{-1}\bar{r}\bar{r}R^{-1}Y}{\delta} - \frac{\bar{r}'R^{-1}Y}{\delta} + \frac{\left(1 - F'R^{-1}\bar{r}\right)y_{m+1}}{\delta}$$

$$a = \underline{r}'\underline{R}^{-1}\underline{Y} - (\underline{F}'\underline{R}^{-1}\underline{r} - \underline{f})'(\underline{F}'\underline{R}^{-1}\underline{F})^{-1}\underline{F}'\underline{R}^{-1}\underline{Y}$$

$$a = \underline{r}'\underline{R}^{-1}\underline{Y} - S1'S2^{-1}(\underline{F}'\underline{R}^{-1}\underline{Y})$$

$$a = \underbrace{r'R^{-1}Y + \frac{r'R^{-1}\overline{rr}R^{-1}Y}{\delta} - \frac{r_{m+1}\overline{r}'R^{-1}Y}{\delta}}_{a1} + \underbrace{\frac{\left(r_{m+1} - r'R^{-1}\overline{r}\right)y_{m+1}}{\delta}}_{a2}...$$

$$-S1'S2^{-1}\left(F'R^{-1}Y + \frac{F'R^{-1}\overline{rr}R^{-1}Y}{\delta} - \frac{\overline{r}'R^{-1}Y}{\delta} + \frac{\left(1 - F'R^{-1}\overline{r}\right)y_{m+1}}{\delta}\right)$$

$$a1 = r'R^{-1}Y + \frac{r'R^{-1}\overline{rr}R^{-1}Y}{\delta} - \frac{r_{m+1}\overline{r}'R^{-1}Y}{\delta} - S1'S2^{-1}\left(F'R^{-1}Y + \frac{F'R^{-1}\overline{rr}R^{-1}Y}{\delta} - \frac{\overline{r}'R^{-1}Y}{\delta}\right)$$

$$a2 = \frac{\left(r_{m+1} - r'R^{-1}\overline{r}\right)y_{m+1}}{\delta} - S1'S2^{-1}\left(\frac{\left(1 - F'R^{-1}\overline{r}\right)y_{m+1}}{\delta}\right)$$

REFERENCES

[1] Akao, Y., 2004, *Quality Function Deployment: Integrating Customer Requirements into Product Design*, Productivity Press, New York.

[2] Alexandrov, N. M., Lewis, R. M., 2001, "An Overview of First-Order Model Management for Engineering Optimization," *Optimization and Engineering*, **2**, pp. 413-430.

[3] Alexandrov, N. M., Lewis, R. M., Gumbert, C.R., Green, L.L., and Newman, P.A., 1999, "Optimization with Variable-Fidelity Models Applied to Wing Design," NASA, Langley Research Center, Hampton, VA

[4] Alexandrov, N. M., Lewis, R. M., Gumbert, C.R., Green, L.L., and Newman, P.A., 2001, "Approximation and Model Management in Aerodynamic Optimization with Variable-Fidelity Models," *Journal of Aircraft*, **38**(6), pp. 1093-1101.

[5] Ankenman, B., Nelson, B. L., and Staum, J., 2008, "Stochastic Kriging for Simulation Metamodeling," Winter Simulation Conference, pp. 362-370.

[6] Bakr, M. H., Bandler, J.W., 2000, "Review of the Space Mapping Approach to Engineering Optimization and Modeling," *Optimization and Engineering*, **1**, pp. 241-276.

[7] Belegundu, A. D., and Chandrupatla, T. R., 1999, *Optimization Concepts and Applications in Engineering*, Prentice-Hall, Upper Saddle River, NJ.

[8] Cellier, F. E., 1991, *Continuous System Modeling*, Springer-Verlag, New York.

[9] Conigliaro, R. A., Kerzhner, A.A, Paredis, C.J.J., 2007, "Model-Based Optimization of a Hyraulic Backhoe Using Multi-Attribute Utility Theory," *SAE International Journal of Materials & Manufacturing*, **2**, pp. 298-309.

[10] Cressie, N. A. C., 1993, *Statistics for Spatial Data*, John Wiley & Sons, Inc., New York.

[11] Dixon, L. C. W., and Szego, G. P., 1978, "The Global Optimisation Problem: An Introduction," *Towards Global Optimisation*, **2**, pp. 1-15.

[12] Dyn, N., Levin, D., and Rippa, S., 1986, "Numerical Procedures for Surface Fitting of Scattered Data by Radial Functions," *SIAM Journal on Scientific and Statistical Computing*, **7**, pp. 639.

[13] Eldred, M. S., Giunta, A. A., Wojtkiewicz, S. F., Trucano, T. G., 2002, "Formulations for Surrogate-Based Optimization under Uncertainty," in *9th AIAA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Atlanta, GA.

[14] Friedman, J. H., and File, P. D. F., 1991, "Rejoinder: Multivariate Adaptive Regression Splines," *Ann. Statist*, **19**(1), pp. 123-141.

[15] Gano, S. E., Perez, V. M., Renaud, J. E., Batill, S. M., 2004, "Multilevel Variable Fidelity Optimization of a Morphing Unmanned Aerial Vehicle," in *AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics & Materials Conference*, Palm Springs, CA.

[16] Gano, S. E. R., J. E.; Martin, J. D.; Simpson, T. W., 2005, "Update Strategies for Kriging Models for Use in Variable Fidelity Optimization," in *1st AIAA Multidisciplinary Design Optimization Specialist Conference*, Austin, TX.

[17] Giannakoglou, K. C., 2002, "Design of Optimal Aerodynamic Shapes Using Stochastic Optimization Methods and Computational Intelligence," *Progress in Aerospace Sciences*, **38**, pp. 43-76.

[18] Giese, H., Levendovszky, T., Vangheluwe, H., 2007, "Summary of the Workshop on Multi-Paradigm Modeling: Concepts and Tools," *Models in Software Engineering*, Springer Berlin, Heidelberg, **4364/2007**, pp. 252-262.

[19] Ginsbourger, D. L. R., Rodolphe; Carraro, Laurent, 2008, "A Multi-Points Criterion for Deterministic Parallel Global Optimization Based on Gaussian Processes," *hal-00260579*.

[20] Gurnani, A., Donndelinger, J., Kemper, L., 2005, "Feasibility Assessment in Preliminary Design Using Pareto Sets," in *ASME 2005 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, Long Beach, CA.

[21] Gurnani, A., Ferguson, S., Lewis, K., Donndelinger, J., 2006, "A Constraint-Based Approach to Feasibility Assessment in Preliminary Design," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, **20**, pp. 351-367.

[22] Handcock, M. S., 1991, "On Cascading Latin Hypercube Designs and Additive Models for Experiments," *Communications in Statistics - Theory and Methods*, **20**(2), pp. 417-439.

[23] Handcock, M. S. S., Michael L., 1993, "A Bayesian Analysis of Kriging," *Technometrics*, **35**(4), pp. 403-410.

[24] Hays, R., Singer, M., 1989, *Simulation Fidelity in Training System Design: Bridging the Gap between Reality and Training*, Springer-Verlag, New York.

[25] Hazelrigg, G. A., 1996, *Systems Engineering: An Approach to Information-Based Design*, Prentice-Hall, Upper Saddle River, NJ.

[26] Hazelrigg, G. A., 1998, "A Framework for Decision-Based Engineering Design," *ASME Journal of Mechanical Design*, **120**, pp. 653-658.

[27] Howard, R., 1966, "Information Value Theory," *IEEE Transactions on Systems Science and Cybernetics*, **SSC-2**(1), pp. 779-783.

[28] Huang, D., Allen, T., Notz, W., and Zeng, N., 2006, "Global Optimization of Stochastic Black-Box Systems Via Sequential Kriging Meta-Models," *Journal of Global Optimization*, **34**(3), pp. 441-466.

[29] Huang, D., Allen, T. T., Notz, W.I., Miller, R. A., 2006, "Sequential Kriging Optimization Using Multiple-Fidelity Evaluations," *Structural and Multidisciplinary Optimization*, **32**(5), pp. 369-382.

[30] Jones, D. R., 2001, "A Taxonomy of Global Optimization Methods Based on Response Surfaces," *Journal of Global Optimization*(21), pp. 345-383.

[31] Jones, D. R. S., Matthias; Welch, William J., 1998, "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, **13**, pp. 455-492.

[32] Joseph, V. R., and Hung, Y., 2008, "Orthogonal-Maximin Latin Hypercube Designs," *Statistica Sinica*, **18**(1), pp. 171-186.

[33] Keeney, R. L., and Raiffa, H., 1993 (1976), *Decisions with Multiple Objectives*, Cambridge University Press, Cambridge, UK.

[34] Kennedy, M., and O'Hagan, A., 2001, "Bayesian Calibration of Computer Models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**(3), pp. 425-464.

[35] Kennedy, M. C., and O'Hagan, A., 2000, "Predicting the Output from a Complex Computer Code When Fast Approximations Are Available," *Biometrika*, **87**(1), pp. 1-13.

[36] Kushner, H. J., 1964, "A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise," *Journal of Basic Engineering*, **86**(1), pp. 97-106.

[37] Lawrence, D. B., 1999, *The Economic Value of Information*, Springer Verlag.

[38] Le Moigne, A., Qin, N., 2004, "Variable-Fidelity Aerodynamic Optimization for Turbulent Flows Using a Discrete Adjoint Formulation," *AIAA Journal*, **42**(7), pp. 1281-1292.

[39] Lewis, K., Chen, W., and Schmidt, L. C., 2006, *Decision Making in Engineering Design*, American Society of Mechanical Engineers, New York.

[40] Li, P. Y., and Mensing, F., 2010, "Optimization and Control of a Hydro-Mechanical Transmission Based Hybrid Hydraulic Passenger Vehicle," in *7th International Fluid Power Conference*, Aachen, Germany.

[41] Lophaven, S. N., Nielsen, H. B., and Sondergaard, J., 2002, "Dace: A Matlab Kriging Toolbox," Technical Report IMM-TR-2002-12, Technical University of Denmark

[42] Malak, R. J., and Paredis, C. J. J., 2008, "Modeling Design Concepts under Risk and Uncertainty Using Parameterized Efficient Sets," *SAE World Congress*, Detriot, MI.

[43] McKay, M. D., Beckman, R. J., and Conover, W. J., 1979, "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, **22**(2), pp. 239-245.

[44] Moore, R. A., 2009, *Variable Fidelity Modeling as Applied to Trajectory Optimization for a Hydraulic Backhoe*, Thesis, Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA.

[45] Moore, R. A., and Paredis, C. J. J., 2009, "Variable Fidelity Modeling as Applied to Trajectory Optimization for a Hydraulic Backhoe," in *2009 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, San Diego, CA, USA.

[46] Ong, Y. S., Nair, P. B., Keane, A. J., Wong, K. W., 2004, "Surrogate-Assisted Evolutionary Optimization Frameworks for High-Fidelity Engineering Design Problems," *Knowledge Incorporation in Evolutionary Compuation*, Jin, Y. eds., Springer Verlag, pp. 307-322.

[47] Pahl, G., and Beitz, W., 1996, *Engineering Design: A Systematic Approach*, Springer-Verlag, London.

[48] Paredis, C. J. J., 1996, *An Agent-Based Approach to the Design of Rapidly Deployable Fault Tolerant Manipulators*, Thesis, Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh.

[49] Perez, V. M., Renaud, John E., Gano, Shawn E., 2000, "Constructing Variable Fidelity Response Surface Approximations in the Usable Feasible Region," in *AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Long Beach, CA.

[50] Pugh, S., 1991, *Total Design: Integrated Methods for Successful Product Engineering*, Addison-Wesley, Reading, MA.

[51] Qian, Z., Seepersad, C. C., Joseph, V. R., Allen, J. K., and Wu, C. F. J., 2006, "Building Surrogate Models Based on Detailed and Approximate Simulations," *Journal of Mechanical Design*, **128**(4), pp. 668-678.

[52] Rodriguez, J. F., and Renaud, J.E., 1998, "Convergence of Trust Region Augmented Lagrangian Methods Using Variable Fidelity Approximation Data," *Structural Optimization*, **15**, pp. 141-156.

[53] Rodriguez, J. F., Perez, V.M., Padmanabhan, D., and Renaud, J.E., 2001, "Sequential Approximate Optimization Using Variable Fidelity Response Surface Approximations," *Structural and Multidisciplinary Optimization*, **22**, pp. 23-34.

[54] Sacks, J. S., Susannah B., Welch, William J., 1989, "Designs for Computer Experiments," *Technometrics*, **31**(1), pp. 41-47.

[55] Sacks, J. W., William J.; Mictchell, Toby J.; Wynn, Henry P., 1989, "Design and Analysis of Computer Experiments," *Statistical Science*, **4**(4), pp. 409-435.

[56] Sage, A. P., and Armstrong Jr., J. E., 2000, *Introduction to Systems Engineering*, Wiley and Sons.

[57] Sage, A. P., Armstrong, J, 2000, *Introduction to Systems Engineering*, John Wiley & Sons, Inc., New York.

[58] Sasena, M. J., Papalambros, P., and Goovaerts, P., 2002, "Exploration of Metamodeling Sampling Criteria for Constrained Global Optimization," *Engineering Optimization*, **34**(3), pp. 263-278.

[59] Schmit, L. A., and Farshi, B., 1974, "Some Approximation Concepts for Structural Synthesis," *AIAA Journal*, **12**(5), pp. 692-699.

[60] Shan, S. W., G. Gary, 2010, "Metamodeling for High Dimensional Simulation-Based Design Problems," *Journal of Mechanical Design*, **132**.

[61] Simpson, T., Poplinski, J., Koch, P., and Allen, J., 2001, "Metamodels for Computer-Based Engineering Design: Survey and Recommendations," *Engineering with Computers*, **17**(2), pp. 129-150.

[62] Tang, B., 1993, "Orthogonal Array-Based Latin Hypercubes," *Journal of the American Statistical Association*, **88**(424), pp. 1392-1397.

[63] Thompson, S. C., and Paredis, C. J. J., 2009, "A Process-Centric Problem Formulation for Decision-Based Design," in *International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, San Diego, CA, USA.

[64] Thurston, D. L., 1991, "A Formal Method for Subjective Design Evaluation with Multiple Attributes," *Research in Engineering Design*, **3**(2), pp. 105-122.

[65] von Neumann, J., and Morgenstern, O., 1980, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ.

[66] Wassenaar, H. J., and Chen, W., 2003, "An Approach to Decision-Based Design with Discrete Choice Analysis for Demand Modeling," *Journal of Mechanical Design*, **125**(3), pp. 490-497.

[67] Welch, W. J. B., Robert J.; Sacks, Jerome; Wynn, Henry P.; Mitchell, Toby J.; Morris, Max D., 192, "Screening, Predicting, and Computer Experiments," *Technometrics*, **34**(1), pp. 15-25.

[68] Xia, L., and Gao, Z.-h., 2006, "Application Fo Variable-Fidelity Models to Aerodynamic Optimization," *Applied Mathematics and Mechanics*, **27**(8), pp. 1089-1095.

[69] Xiong, F., Xiong, Y., Chen, W., and Yang, S., 2009, "Optimizing Latin Hypercube Design for Sequential Sampling of Computer Experiments," *Engineering Optimization*, **41**(8), pp. 793-810.

[70] Zang, T. A., Green, L. L., 1999, "Multidisciplinary Design Optimization Techniques: Implications and Opportunities for Fluid Dynamics Research," in *30th AIAA Fluid Dynamics Conference*, Norfolk, VA.

[71] Zhou, Z., Nair, P.B., Keane, A. J., Lum, K. Y., 2007, "Combining Global and Local Surrogate Models to Accelerate Evolutionary Optimization," *IEEE Transactions on Systems, Man, and Cybernetics--Part C: Applications and Reviews*, **37**(1), pp. 66-76.