Black box simulation optimization

Angun, M.E.

Publication date:
2004

BLACK BOX SIMULATION OPTIMIZATION: GENERALIZED RESPONSE SURFACE METHODOLOGY

BLACK BOX SIMULATION OPTIMIZATION: GENERALIZED RESPONSE SURFACE METHODOLOGY

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Tilburg,
op gezag van de rector magnificus, prof. dr. F.A. van der Duyn Schouten,
in het openbaar te verdedigen ten overstaan van
een door het college voor promoties aangewezen commissie
in de aula van de Universiteit
op dinsdag 29 juni 2004 te 14.15 uur door

Mevlüde Ebru Angün

geboren op 5 juni 1972 te Ankara, Turkije.

Promotor:	Prof. dr. J.P.C. Kleijnen
Promotor:	Prof. dr. ir. D. den Hertog
Copromotor:	dr. G. Gürkan

Acknowledgment

This thesis consists of three papers that were written during my Ph.D. at the Department of Information Systems and Management in Tilburg University. I take this opportunity to thank to the people who made my stay here a productive and enjoyable experience.

My deepest thanks go to my advisors Jack Kleijnen and Dick den Hertog, and to my co-advisor Gül Gürkan.

I am also very grateful to the members of my Ph.D. committee, namely, Enrique Del Castillo from Pennsylvania State University, Rommert Dekker from Erasmus University, Bertrand Melenberg from Tilburg University, and Maarten van der Vlerk from Groningen University.

Finally, I feel very much indebted to my parents and my sister for their support throughout.

İstanbul, May 2004

Contents

List of Tables

List of Figures

Chapter 1

General Introduction to Simulation Optimization

1.1 Introduction

The term "simulation optimization" has become widespread in both academical and practical studies. From the academical point of view, simulation optimization has become one of the new entries in the updated second edition of the "Encyclopedia of Operations Research and Management Science" (Gass and Harris (2000)). Furthermore, many surveys and panel discussions about the future of simulation optimization, and its methodologies and applications have been published; see Fu (2002), Andradóttir et al. (2000), Azadivar (1999), and Andradóttir (1998) (also all the Winter Simulation Conference proceedings, which are available online at the website: www.wintersim.org). From the practical point of view, optimization modules have been recently implemented in many commercial discrete-event simulation packages; Table 1.1 - taken from Fu (2002, p. 3) - shows simulation packages and optimization methods incorporated into these packages.

We can explain one of the many reasons for the interest in simulation optimization, as follows. For problems that arise in practical applications, explicit mathematical formulations may be too restrictive; that is where simulation is relevant. Therefore, for many practical cases one cannot obtain an analytical solution through methods that require explicit mathematical formulations. Indeed, simulation optimization has led to the numerical solution of large-scale, real-world decision-making problems; see, for example, Azadivar and Truong (2003), April, Glover, and Kelly (2003), and Martin and Schouwenaar (2003).

Table 1.1: An overview of simulation optimization in practice (source: Fu (2002, p. 3))

Optimization package (Simulation platform)	**Vendor (URL)**	**Primary search strategies**
AutoStat (AutoMod)	Auto Simulations, Inc. (www.autosim.com)	Genetic algorithms
OptQuest (Arena, CrystalBall)	Optimization Technologies, Inc. (www.opttek.com)	Scatter search, tabu search
SimRunner (ProModel)	PROMODEL Corp. (www.promodel.com)	Genetic algorithms
Optimizer (WITNESS)	Lanner Group, Inc. (www.lanner.com/corporate)	Simulated annealing, tabu search

In this thesis, we mainly consider stochastic simulation. Moreover, we focus on black box simulation optimization methods - which we will detail later in this chapter - because of their generality, and their ease of use and implementation. To illustrate their generality, we give the following references that apply black box simulation optimization methods to either deterministically or stochastically simulated systems in very diverse fields, such as engineering design (den Hertog and Stehouwer (2002)) and ergonomic design of workstations (Ben-Gal and Bukchin (2002)) - where both papers use deterministic simulation -, and air traffic control (Hutchison and Hill (2001)), production planning (Kleijnen (1993)), and design of manufacturing cells (Irizarry, Wilson, and Trevino (2001)) - where all three papers use stochastic simulation.

In the rest of this thesis, we do not consider screening, which can be considered as stage zero (pre-processing phase) of any black box simulation optimization methods. For screening, we give the following brief description and references. Screening is the process of searching for the few truly important input variables among the great many potentially important input variables that affect a system's performance; for a recent reference, see Dean and Lewis (2004). Trocine and Malone (2000) compares and contrasts screening methods in terms of efficiency, effectiveness, and robustness. As screening methods, Trocine and Malone (2000) considers classical factorial designs (Myers and Montgomery (2002)), two-stage group screening (Kleijnen (1987)), sequential bifurcation (Bettonvil and Kleijnen (1996), and Cheng (1997)), and iterated fractional factorial designs (Campolongo, Kleijnen, and Andres (2000)).

More specifically, we focus on the black box method called Response Surface Methodology (RSM). RSM originated in Box and Wilson (1951). Myers and Montgomery (2002), which is a classic textbook on RSM, gives the following general description, for a minimization problem. In the first stage of RSM, an experimental design is used to locally fit a first-order polynomial to the observed values of the random response; this fit uses linear regression. Then, a steepest descent search direction is estimated from the fitted first-order polynomial, and a number of steps are taken along this direction - until no additional decrease in objective is evident. This procedure is repeated until a first-order polynomial becomes an inadequate model, which is indicated when the gradient is not significantly different than a zero vector. In the second stage of RSM, a second-order polynomial is fitted locally, and this polynomial is minimized. Furthermore, canonical and ridge analyses are performed to determine the nature of the fitted objective function (i.e., convex, concave, or indefinite) and the nature of the estimated optimum (i.e., single optimum or multiple optima).

Some disciplines interpret RSM in a completely different way: RSM becomes

a one-shot approach that fits a single response surface, which is either a second-order polynomial or a kriging model, to the random response over the whole experimental area. Next, this nonlinear fitted model is optimized using a global optimization procedure; see Sacks et al. (1989), Jones, Schonlau, and Welch (1998), and Simpson et al. (2001). In this thesis, we do not consider the one-shot RSM, but we concentrate on the sequential version as recently described in Myers and Montgomery (2002).

RSM is broadly applicable in the sense that it can be easily integrated into both stochastic and deterministic simulations, since RSM does not necessarily exploit the stochastic structure of the simulated system. Not exploiting the stochastic structure has the advantage of being very flexible. However, a disadvantage is that RSM may be computationally more expensive than the other methods that do take the stochastic structure into account.

RSM has the two important features of black box approaches, namely generality and simplicity. Furthermore, unlike genetic algorithms that have a family of solution points at each iteration, RSM has a single solution point at each iteration along the search path. RSM has the following disadvantages compared to metaheuristics (i.e., genetic algorithms, tabu search, and simulated annealing): (i) RSM is assumed to be a continuous optimization method, since RSM is similar to gradient-based approaches. Hence, unlike metaheuristics, RSM is not suitable for discrete optimization; (ii) RSM may find a local optimum, as opposed to metaheuristics that search for a global one.

In this study, we aim at solving several methodological problems of RSM that have not been considered intensively yet. In particular, we deal with inefficiency and scale dependence of RSM's search direction, and inefficiency in step size selection along the search direction for both unconstrained and stochastically constrained optimization problems. Furthermore, we provide a well-established search procedure for computationally expensive problems. Notice that in classic RSM, one switches from the first stage to the second stage - where we described both stages in the previous paragraph - since the estimated gradient of the fitted objective - which is classic RSM's search direction - is not significantly different than a zero vector in a neighborhood of the true optimum. Then, the second stage completes the whole procedure. However, when there are (possibly stochastic) constraints, the estimated gradient of the fitted objective may be significantly different than a zero vector in a neighborhood of the true optimum. Hence, the search direction, which is a function of the estimated gradient, will be also significantly different than a zero vector. For such cases, our conjecture is that we may not need to switch to the second stage of classic RSM. Hence, we provide an asymptotic stopping rule for the whole procedure, without switching to the second

stage. It is worth mentioning that to the best of our knowledge, there has been no study done so far to compare RSM with other black box methods. We expect that once the methodological problems in RSM are solved, such comparisons will receive more attention.

Originally, RSM was derived for problems with a single, random response to be optimized; see Myers and Montgomery (2002). We call this version of RSM classic RSM. In general, however, optimization problems have multiple random responses; see, for example, Law and Kelton (2000). We will consider all but one of these responses as constraints, in addition to deterministic box constraints on the input variables. There have been several approaches to generalize RSM to cope with (possibly) stochastic constraints, such as the desirability function (Harrington (1965), and Derringer and Suich (1980)), the generalized distance (Khuri and Conlon (1981)), the dual response (Myers and Carter (1973), Vining and Myers (1990), Del Castillo and Montgomery (1993), and Fan and Del Castillo (1999)), and the prediction-interval constrained goal programming (Wei, Olson, and White (1990)). A brief description of these approaches will be given in Chapter 3. Notice that when there are multiple responses, it is possible to consider all of them as objectives - which gives a multi-criteria optimization problem. In this study, we do not deal with this approach. Yet, this is one of the recent advances in the classic RSM literature; see Yang and Tseng (2002), and Koksoy (2003).

As we mentioned previously, in the rest of this thesis we do not consider the second stage of RSM; we refer to Irizarry, Wilson, and Trevino (2001), and Draper and Pukelsheim (2000).

The remainder of this chapter is organized as follows. §1.2 distinguishes between black box and white box approaches to gradient estimation through simulation, with the emphasis on black box approaches. §1.3 gives an overview of black box simulation optimization methods. Finally, §1.4 summarizes the main contributions of this thesis per chapter.

1.2 Black box approaches to gradient estimation

In this section, we differentiate between black box and white box approaches to gradient estimation through simulation. We emphasize black box approaches, since RSM - the focus of this thesis - is a black box method. By definition, simulation is treated as a black box if the gradient estimates (and possibly higher order derivative estimates) through simulation are not available using either perturbation analysis (PA) - see Ho and Cao (1991), Glasserman (1991), and Fu and Hu (1997) - or likelihood ratio score

function (LR/SF) - see Rubinstein and Shapiro (1993).

Before presenting the common approaches to obtain black box gradient estimates through simulation, we introduce a general problem formulation for simulation optimization, as follows:

$$\underset{\underline{d} \in \Theta}{\text{minimize}} \quad E_{\underline{\omega}} \left[F \left(\underline{d}, \ \underline{\omega} \right) \right] \tag{1.1}$$

where $E_{\underline{\omega}}$ is the expectation operator with respect to the simulation's seed vector $\underline{\omega}$, $F\left(\underline{d}, \ \underline{\omega}\right)$ is a random response estimated through simulation, $\underline{d}$ is $k \times 1$ vector of input variables, and Θ is the (explicitly or implicitly defined) feasible search space.

The most straightforward way for obtaining gradient estimates uses finite differences. Forward finite differencing (FFD) needs $k+1$ simulation runs to obtain a single gradient estimate; i.e., the ith ($i = 1, \ ..., \ k$) component of the gradient estimate, say $\widehat{g}_i$, at $\underline{d}$ is

$$\widehat{g}_i \left(\underline{d} \right) = \frac{\widehat{F} \left(\underline{d} + c_i \underline{e}_i \right) - \widehat{F} \left(\underline{d} \right)}{c_i} \tag{1.2}$$

where $\widehat{F} \left(\underline{d} + c_i \underline{e}_i \right)$ and $\widehat{F} \left(\underline{d} \right)$ are estimates of F in (1.1) at the two input vectors $\underline{d} + c_i \underline{e}_i$ and $\underline{d}$ with $\underline{e}_i$ the unit vector in the ith direction and c_i a scalar.

Central finite differencing (CFD) conducts $2k$ simulation runs to obtain a single gradient estimate:

$$\widehat{g}_i \left(\underline{d} \right) = \frac{\widehat{F} \left(\underline{d} + c_i \underline{e}_i \right) - \widehat{F} \left(\underline{d} - c_i \underline{e}_i \right)}{2 c_i} \tag{1.3}$$

where $\widehat{F} \left(\underline{d} - c_i \underline{e}_i \right)$ is an estimate of F in (1.1) at the input vector $\underline{d} - c_i \underline{e}_i$. Obviously, CFD is computationally more expensive than FFD. On the other hand, the CFD estimators are less biased than the FFD estimators. In either case, however, a single gradient estimate requires $O\left(k\right)$ simulation replicates.

In the following chapters, we will use resolution-3 designs to obtain gradient estimates. By definition, this design type gives unbiased estimators of the gradients of the random responses, provided that first-order polynomial approximations are adequate for these responses; see Kleijnen (1998). To obtain a single gradient estimate, resolution-3 designs need only $k+1$ simulation runs, rounded upwards to a multiple of four. Therefore, resolution-3 designs are computationally more efficient than CFD; they require the same number of runs as FFD, but give smaller variances (standard errors) than FFD. For further comparisons of design of experiments schemes - including resolution-3 designs - with FFD and CFD, we refer to Brekelmans et al. (2004).

The method of simultaneous perturbation (SP) of Spall (2000) avoids these $O\left(k\right)$ simulation replicates, and estimates a single gradient by perturbing in all direc-

tions simultaneously. To obtain a single gradient estimate, SP needs only two simulation runs, independent of k, as follows:

$$\widehat{g}_i\left(\underline{d}\right) = \frac{\widehat{F}\left(\underline{d} + c_i\underline{\widehat{\Delta}}\right) - \widehat{F}\left(\underline{d} - c_i\underline{\widehat{\Delta}}\right)}{2\widehat{\Delta}_i} \tag{1.4}$$

where $\widehat{F}\left(\underline{d} + c_i\underline{\widehat{\Delta}}\right)$ and $\widehat{F}\left(\underline{d} - c_i\underline{\widehat{\Delta}}\right)$ are estimates of F in (1.1) at the two input vectors $\underline{d} + c_i\underline{\widehat{\Delta}}$ and $\underline{d} - c_i\underline{\widehat{\Delta}}$, and $\underline{\widehat{\Delta}} = \left(\widehat{\Delta}_1, ..., \widehat{\Delta}_i, ..., \widehat{\Delta}_k\right)^T$ represents a realization of a vector, say $\underline{\Delta}$, of independent, identically distributed random perturbations satisfying certain conditions given in Spall (1992). In practice, the simplest and most popular distribution for $\underline{\Delta}$ is the symmetric (scaled) Bernoulli distribution. The difference between the FFD/CFD estimators and the SP estimators is that the numerator, which involves the expensive simulation replicates, varies in the FFD/CFD estimates (see (1.2) and (1.3)), whereas the numerator is constant in the SP estimates, and it is the denominator involving the random perturbations that varies (see (1.4)). A difficulty encountered in implementing FFD, CFD, and SP is that the choice of the scalar c_i must balance between too much noise and too much bias; that is, in order for the bias to be small, it is necessary to let the scalar c_i be small. However, when c_i is small, the FFD, CFD, and SP estimators usually have large variances; see Spall (1998) for practical guidelines for choosing c_i in SP.

In Table 1.2 - again taken from Fu (2002, p. 29) - we compare white box approaches to gradient estimation such as infinitesimal perturbation analysis (IPA), which is the simplest form of PA, and LR/SF, as well as FFD, CFD, and SP. IPA and LR/SF require more knowledge about the simulated system than FFD, CFD, and SP. For example, LR/SF assumes the knowledge of a distribution that dominates the input distribution; see L'Ecuyer (1991). Under certain conditions, IPA and LR/SF provide gradient (and possibly higher order derivatives) estimators with desirable statistical properties such as unbiasedness and strong consistency through a single simulation run; for these conditions, see, for example, Glasserman (1991) for PA, and Rubinstein and Shapiro (1993) for LR/SF. LR/SF applies more generally than IPA; see L'Ecuyer (1991). However, Glynn (1989) proves that when both IPA and LR/SF yield unbiased and strongly consistent estimators, LR/SF's estimators have larger variances. For further comparison of finite differences, IPA, and LR/SF, we refer to L'Ecuyer (1991).

Table 1.2: Gradient estimation through simulation (source: Fu (2002, p. 29))

Approach	**Number of simulations**	**Key feature(s)**	**Disadvantage**
IPA	1	highly efficient, easy to implement	limited applicability
LR/SF	1	assumes the knowledge of a distribution that dominates the input distribution	possibly high variance
SP	2	widely applicable	generally noisy
FFD	$k+1$	widely applicable	generally noisy
CFD	$2k$	widely applicable	generally noisy

1.3 Black box simulation optimization methods

In this section, we present several black box simulation optimization methods that we consider as alternatives to RSM. We do not claim to give an exhaustive literature survey on black box simulation optimization. We focus on methods that we think to be most widely used; also see Andradóttir et al. (2000) and Boesel et al. (2001), which are the panel discussions at the Winter Simulation Conferences in 2000 and 2001, respectively, and also Table 1.1, which shows the most popular optimization approaches among practitioners.

In the following, we do not consider statistical ranking, selection, and multiple comparison methods; see, for example, Goldsman et al. (1999), or a more recent reference Boesel, Nelson, and Kim (2003). The primary difference between statistical ranking, selection, and multiple comparison methods and the methods described below is that the former methods evaluate exhaustively all members of a fixed and finite set of alternatives. However, the latter methods attempt to search efficiently through the feasible search space to find better solutions, because exhaustive search is impractical or impossible (i.e., the feasible search space can be unbounded or uncountable). Furthermore, we do not consider sample path optimization of Gürkan, Özge, and Robinson (1999, 1994), and random search methods (see, for example, Andradóttir (1998)). The sample path methods use IPA to estimate gradients through simulation (hence, it is a white box approach). To the best of our knowledge, random search methods have been applied solely to discrete optimization problems.

In the following subsections, we shall summarize black box methods, namely the stochastic approximation method, genetic algorithms, tabu search, simulated annealing, and ordinal optimization.

1.3.1 The stochastic approximation method

The stochastic approximation method attempts to mimic the gradient search method in deterministic optimization, while taking into account the stochastic setting. Given the problem in (1.1), the general form of stochastic approximation is:

$$\underline{d}_{n+1} = \Pi_{\Theta}\left(\underline{d}_n - a_n \widehat{g}\left(\underline{d}_n\right)\right) \tag{1.5}$$

where Θ is closed and convex, Π_{Θ} denotes some projection back into Θ when the iteration leads to a point outside Θ (e.g., the simplest projection would be to return to the previous point (Fu (2002, p. 25)), n denotes the iteration number, and $\{a_n\}$ is a

sequence of step size multipliers such that $\sum_{n=1}^{\infty} a_n = \infty$ and $\sum_{n=1}^{\infty} a_n^2 < \infty$. If the finite differences ((1.2) or (1.3)) are used to obtain $\widehat{g}$ in (1.5), then the procedure in (1.5) is called Kiefer-Wolfowitz's algorithm; see Kiefer and Wolfowitz (1952). If simultaneous perturbation in (1.4) is used to obtain $\widehat{g}$ in (1.5), then it is called simultaneous perturbation stochastic approximation (SPSA); see Spall (2003, 2000, 1999), and Kleinman, Spall, and Naiman (1999).

Under appropriate conditions, one can prove the convergence of the sequence $\{\underline{d}_n\}$ to the local minimum with probability one, as n goes to infinity; for these conditions in case of SPSA, see Spall (1992). In practice, however, the performance of stochastic approximation is very much sensitive to the sequence $\{a_n\}$; see, for example, Fu (2002, p. 28). Theoretically, a constant step size results in weak convergence - that is, convergence in distribution - which means that the iterates may oscillate around the local minimum. Yet, in practice, a constant step size results in much quicker convergence in the early stages of the method - unlike a step size decreasing at each step.

Some of the (recent) advances on stochastic approximation are as follows. Since stochastic approximation is a gradient search method, it generally finds a local optimum. Therefore, Maryak and Chin (2001) enhances SPSA to find the global optimum. Furthermore, Gerencsér, Hill, and Vágó (1999), and Whitney, Solomon, and Hill (2001) apply stochastic approximation to discrete optimization, although it has been usually used for continuous optimization.

1.3.2 Genetic algorithms

Unlike stochastic approximation and RSM, evolutionary search strategies such as genetic algorithms work with a family of solution points - namely the population - rather than a single solution point. More importantly, the solution points in the current population interact to form the next population - also called the next generation. Important factors that affect the success of genetic algorithms are the selection procedure and the types of genetic operators. Selection can be done either deterministically or probabilistically - and is based on the fitness of a solution point, where the fitness of a solution point corresponds to its objective function value. Two of the simplest (deterministic) selection procedures include keeping each generation at a constant number of the fittest solution points (survival of the fittest), or keeping only the offspring from reproduction (complete generational turnover). Genetic operators operate on a genetic representation (code) of a solution point, and are generally classified as crossover oper-

ators and mutation operators. The crossover operators take two solution points from the population that have relatively good fitnesses, and combine them to make two new solution points. The mutation operators take a single, well-performing solution point, and alter it slightly. Notice that the crossover operators distinguish genetic algorithms from other metaheuristics, such as simulated annealing and tabu search. For more details and references, see Michalewicz and Schoenauer (2001), and Fouskakis and Draper (2002).

1.3.3 Tabu search

Tabu search (Glover and Laguna (1997)) can be thought of as a variation on local search that incorporates two main strategies, namely adaptive memory and responsive exploration. The features of these strategies modify the neighborhood of a solution point as the search progresses, and thus determine the effectiveness of the algorithm. In particular, the modification from which the method derives its name may forbid certain points (classifying them tabu); i.e., these tabu points do not belong to the current neighborhood of a solution point. Thus, for example, short-term memory can prevent the search from revisiting recently visited points, whereas long-term memory can encourage moves that have historically led to improvements (intensification) and moves into previously unexplored regions of the search space (diversification). For details and more references, see Glover (2001), and Fouskakis and Draper (2002).

1.3.4 Simulated annealing

Simulated annealing (Kirkpatrick, Gelatt, and Vwecchi (1981)) may be thought of as a variation on local search, in which the main idea for a minimization problem is to accept all downhill improving moves, but sometimes accept also uphill moves, where the acceptance probability of uphill moves decreases to zero at an appropriate rate (this is the cooling schedule from which the method derives its name, in analogy with the physical annealing process, where the system seeks the lowest energy state). By accepting uphill moves, simulated annealing tries to avoid local minima. An attractive property of this algorithm is that - unlike genetic algorithms and tabu search - convergence can be rigorously proved in many settings; see, for example, Gutjahr and Pflug (1996), and Alrefaei and Andradóttir (1999). On the other hand, the procedure has been found to converge relatively slow to a good solution point, compared to genetic algorithms and tabu search. For more details and references, we refer to Anandalingam (2001), and Fouskakis and Draper (2002).

1.3.5 Ordinal optimization

In this section, we present the ordinal approach to simulation optimization - as opposed to the cardinal approach; see Ho, Sreenivas, and Vakili (1992), and Ho et al. (2000). Note that RSM is a cardinal approach. Furthermore, it will become clear in this section that ordinal optimization is not meant to replace traditional cardinal optimization, but supplements it.

Ordinal optimization differs from cardinal optimization in two important ways: (i) the aim of ordinal optimization is not to look for the best but to find a solution that is good enough (goal softening), which has to be statistically defined; (ii) ordinal optimization focuses on approximating the order among the outputs of the given input vectors rather than accurately estimating the values of these outputs (and possibly their gradients and higher-order derivatives) at these input vectors.

Before detailing these two properties, we explain a general weakness related to simulation optimization problems, following Ho, Sreenivas, and Vakili (1992). Suppose that in (1.1), Θ is a huge, arbitrary, but finite search space. The standard approach to estimate $E_{\underline{\omega}}\left[F\left(\underline{d},\ \underline{\omega}\right)\right]$ is

$$\overline{\widehat{F}}_N\left(\underline{d}\right) = \frac{1}{N}\sum_{j=1}^{N}\widehat{F}\left(\underline{d},\ \widehat{\underline{\omega}}_j\right) \tag{1.6}$$

where $\widehat{\underline{\omega}}_j$ is the jth sampled value of the random vector $\underline{\omega}$ and N is the number of simulation replicates (or the length of the simulation run). Now, the problem is that the confidence interval of (1.6) cannot improve faster than $1/\sqrt{N}$.

Ordinal optimization is based on two tenets:

- Order converges exponentially fast, whereas (1.6) converges at the rate $1/\sqrt{N}$; see Dai (1996), and Dai and Chen (1997). This is intuitively reasonable: think of the problem of holding identically looking packages in each hand and trying to determine which package weighs more versus estimating the difference in weight between the two packages.

- Goal softening eases the computational burden of finding the optimum. In ordinal optimization, one settles for the set of good enough input vectors with high probability (e.g., any of the top r of the input vectors, 95% of the time).

Ordinal optimization has some common features with statistical ranking, selection, and multiple comparison methods: for example, relative ordering in ordinal optimization is in the same spirit as multiple comparisons; goal softening is in the

same spirit as statistical ranking and selection. The primary difference is the scale; the former method deals with a very large search space, whereas Goldsman and Nelson (1994) suggests two to twenty input vectors for the latter methods. Furthermore, ordinal optimization does not address questions such as the distance between the best and the rest, which is a cardinal notion in multiple comparison methods, or whether or not the observed order is a maximum likelihood estimate of the actual order.

Some advantages of ordinal optimization are simplicity (the procedure is easy to state and implement), generality (applicable to a very large class of problems), and efficiency (it is possible to select good input vectors with high probability in the presence of significant estimation errors). In some cases, however, obtaining a good enough subset of input vectors may not be very satisfactory. Then, ordinal optimization can be considered as an approach complementary to cardinal optimization (for example, as a pre-processing step that narrows the search space). Moreover, once a statistically good solution is obtained through the ordinal approach, the final fine tuning to reach the true optimum can be accomplished via cardinal optimization.

Finally, there have been many applications of ordinal optimization to different problems such as communication networks (Patsis, Chen, and Larson (1997)), rare event simulation (Ho and Larson (1995)), resource allocation problems (Ho et al. (2000), and robot motion planning problem (Chen, Kumar, and Luo (1998)). Additional applications and an interactive online demo can be found at the website: hrl.harvard.edu/~ho/DEDS.

1.4 Summary of thesis

In this section, we summarize the contributions of each chapter of this thesis to the classic RSM literature. Each chapter has its own terminology and notation, and can be read independently of the other chapters. Chapter 2, Chapter 3, and Chapter 4 correspond to Kleijnen, den Hertog, and Angün (2004), Angün et al. (2003), and Angün and Kleijnen (2004), respectively. Furthermore, a summary of the main results of Chapter 2 and Chapter 3 can be found in Angün et al. (2002).

In Chapter 2, we consider a classic RSM problem with a single random response. Unlike the classical approach that maximizes the estimated random response, we maximize the lower bound on the estimated random response; that is, we have a more robust (pessimistic) approach. We focus on the first stage, in which RSM fits a first-order polynomial to the observed values of the random response through linear regression, per local area.

Myers and Montgomery (2002) states the following two problems in the classic RSM literature:

- The steepest ascent search direction is scale dependent.
- The step size along the steepest ascent path is selected intuitively.

The main contributions of that chapter are as follows. We derive a novel search direction, which we call adapted steepest ascent. We obtain that search direction by adjusting the estimated first-order factor effects through their estimated covariance matrix. Furthermore, we prove that adapted steepest ascent is scale independent. In most of our numerical experiments, that novel search direction is shown to perform better than classical steepest ascent. We also derive and explore possible solutions for the step size selection.

In Chapter 3, we focus on a problem with a single random objective function and multiple random constraints, in addition to deterministic box constraints on the input variables. Obviously, our problem type is more general than the one in classic RSM (i.e., a problem with a single random objective function). So far, our problem type has not received much attention in simulation optimization; see Fu (2002, p. 6). Also, Myers (1999) points out the importance of extending classic RSM to cope with multiple random responses.

We have already mentioned the approaches in RSM that deal with constrained optimization problems; see §1.1. In all those approaches, the constrained optimization problem is reformulated by combining the constraints and the original objective function into a new, single objective function by appropriate transformations. Then, the resulting unconstrained problem is solved, using an ordinary nonlinear programming algorithm. A major drawback of those approaches is the arbitrariness in the choice of the transformations. We overcome this drawback by considering the original stochastically constrained optimization problem, rather than transforming it into an unconstrained problem.

The contribution of that chapter is its extension of classic RSM to handle stochastic constraints, as well as deterministic box constraints. This is achieved through the generalization of the estimated steepest descent search direction using ideas from interior point methods, more specifically the affine scaling algorithm. Our proposed search direction is the scaled and projected estimated steepest descent - also called the estimated affine scaling search direction. Furthermore, we prove that the proposed search direction has two important features: it is indeed a descent direction, and it is scale independent.

We also develop a heuristic for the first stage of RSM, which uses the proposed search direction iteratively. The step size along the search direction is determined by "adapted" - with respect to the constraints - binary search. Additionally, that heuristic uses statistical hypothesis testing to determine the next iterate along the proposed search direction. Furthermore, although we introduce our heuristic for a stochastic setting, we also explain how to simplify it for a deterministic setting. Our numerical experiments show that our heuristic quickly reaches a desired neighborhood of the true optimum.

Chapter 4 also deals with problems that have a stochastic objective function and stochastic constraints. There, the focus is on the last stage of black box simulation optimization methods. More specifically, the main contribution of that chapter is a statistical, asymptotic stopping rule that tests for the first-order necessary optimality conditions at a feasible point. A related recent paper is by Bettonvil and Kleijnen (2004), which derives a novel procedure to test the first-order necessary optimality conditions for computationally expensive, black box simulation optimization problems, using bootstrapping.

In particular, the rule in Chapter 4 may end the whole RSM procedure - without necessitating to switch to the second stage - in presence of (possibly stochastic) constraints, since, in general, the estimated gradient of the fitted objective is significantly different than a zero vector in a neighborhood of the true optimum - by definition of classic RSM, the first stage ends when the estimated gradient is not significantly different than a zero vector. Notice that to the best of our knowledge, there has been no study in the classic RSM literature that mentions the first-order necessary optimality conditions, since that literature has not yet considered constraints extensively; an exception is Fan and Del Castillo (1999).

Moreover, we introduce two alternative tests, namely Roy's test and the classic F test combined with Bonferroni's inequality, for lack of fit - which is only relevant for RSM. We also show that Roy's test is a direct generalization of the classic F test to multiple responses.

Chapter 2

Response Surface Methodology's (RSM) Steepest Ascent and Step Size Revisited

2.1 Introduction

Response Surface Methodology (RSM) is a stagewise heuristic that searches for the input combination that maximizes the output (finding the minimum is equivalent to finding the maximum of minus the output; the maximization problem is without explicit constraints or side-conditions). Originally, Box and Wilson (1951) meant RSM for experiments with real, non-simulated systems; see the vast literature including Box (1999), Khuri (1996), Khuri and Cornell (1996), Myers (1999), and Myers and Montgomery (2002).

Later on, RSM was also applied to random simulation models; see the large literature including Donohue, Houck, and Myers (1993, 1995), Hood and Welch (1993), Irizarry, Wilson, and Trevino (2001), Kleijnen (1998), Law and Kelton (2000, pp. 646-655), Neddermeijer et al. (2000), and Safizadeh (2002).

RSM treats simulation models - either random or deterministic - and real systems as black boxes. Other black box methods are metaheuristics, including tabu search, simulated annealing, and genetic algorithms. The black box approach is discussed by Jones, Schonlau, and Welch (1998). Various simulation optimization methods are presented by Fu (2002).

In this paper, we do not explain all stages of RSM, but we refer the readers to the literature cited above. We focus on the early stages, in which RSM fits a first-order

polynomial in the inputs, per local area. This fitting uses Ordinary Least Squares (OLS), and estimates the steepest ascent (SA) path, as follows. Let d_j denote the value of the original, non-standardized input j with $j = 1, \ldots, k$. Hence k main or first-order effects (say) β_j are estimated. To enable this estimation, RSM may use a Resolution-3 design specifying which $n \approx k + 1$ input combinations are to be observed or simulated. These n input/output (I/O) combinations give the estimates $\widehat{\beta}_j$. So the SA path uses the estimated local gradient $\underline{\widehat{\beta}}_{-0}^T = (\widehat{\beta}_1, \ldots, \widehat{\beta}_k)$.

Unfortunately, SA search direction suffers from two well-known problems; see Myers and Montgomery (2002): (i) it is scale dependent; (ii) the step size along its path is selected intuitively. In practice, analysts try an intuitively selected value for the step size. If that value yields a lower response, then they reduce the step size. Otherwise, they take one more step. An example is the case study in Kleijnen (1993), which uses a step size that doubles the most important input.

Some disciplines interpret RSM in a completely different way: RSM becomes a one-shot approach that fits a single response surface - either a second-order polynomial or a Kriging model - to the I/O data of a random or deterministic simulation model over the whole experimental area (instead of a series of local areas). Next, that single model is used to estimate the optimal input combination. See Sacks et al. (1989); other more recent references are Jones, Schonlau, and Welch (1998), and Simpson et al. (2001).

Our research contribution is the following. We derive "adapted" steepest ascent (ASA); that is, we adjust the estimated first-order factor effects through their estimated covariance matrix. We prove that ASA is scale independent. In most of our experiments, ASA gives a better search direction than SA. We also derive and explore a possible solution for the step size.

We use underlined letters and bold letters to denote vectors and matrices, respectively. Hence, $\underline{0}$, $\underline{1}$, $\mathbf{0}$, and $\mathbf{1}$ stand for a vector of zeros, a vector of ones, a square matrix of zeros, and a square matrix of ones, respectively. For square matrices, we use subscripts to denote their dimensions; i.e., $\mathbf{I}_k$ denotes a $k \times k$ identity matrix.

The remainder of this paper is organized as follows. §2.2 summarizes those parts of linear regression analysis that we need to formulate our two techniques. §2.3 derives these techniques, and their mathematical properties and interpretation. To get further insight into these properties, §2.4 applies the step-size selection of ASA to simple numerical problems. §2.5 compares the search directions of the two techniques by means of Monte Carlo experiments. §2.6 gives conclusions and future research plans.

2.2 Linear regression basics

We define the estimated signal/noise ratio (say) $\widehat{\gamma}_j$ as

$$\widehat{\gamma}_j = \frac{\widehat{\beta}_j}{\sqrt{\widehat{var}(\widehat{\beta}_j)}} \quad j = 1, \ \ldots, \ k \tag{2.1}$$

where $\widehat{\beta}_j$ is the OLS estimator of β_j in the local first-order polynomial approximation

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j d_j + \varepsilon \tag{2.2}$$

where y is the regression predictor of the corresponding expected output, and ε is white noise; i.e., ε is normally, identically, and independently distributed with zero mean μ_ε and constant variance σ_ε^2.

The OLS estimator in (2.1) is computed as

$$\underline{\widehat{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underline{\widehat{w}} \tag{2.3}$$

with (in order of appearance)

$\underline{\widehat{\beta}}$	:	vector with the q estimated effects in the regression model ($q = 1 + k$ in (2.2))
q	:	number of regression effects including the intercept β_0
$\mathbf{X}$	:	$N \times q$ matrix of explanatory or independent regression variables including the dummy variable with constant value 1; $\mathbf{X}$ is assumed to have linearly independent columns so $\mathbf{X}$ has full column rank
$N = \sum_{i=1}^{n} m_i$	:	number of observations on real system or simulation runs
m_i	:	number of replicates at input combination or point i, with $m_i \in \mathbb{N} \wedge m_i > 0$
n	:	number of different observed or simulated combinations of the k inputs, with $n \in \mathbb{N} \wedge n \geq q$ (necessary condition for avoiding singularity in (2.3))
$\underline{\widehat{w}}$	:	vector with N outputs $\widehat{w}_{i,r}$ - real or simulated - corresponding to the N inputs.

The signal's noise (see (2.1)'s denominator) is the square root of the corresponding element on the main diagonal of

$$cov(\underline{\widehat{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma_\varepsilon^2. \tag{2.4}$$

The unknown parameter σ_ε^2 in (2.4) can be estimated through the mean squared residual (MSR) estimator

$$\widehat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{n}\sum_{r=1}^{m_i}(\widehat{w}_{i,r} - \widehat{y}_i)^2}{(N-q)} \tag{2.5}$$

where the ith component $\widehat{y}_i$ of the regression predictor $\widehat{y}$ follows from

$$\widehat{y} = \mathbf{X}\underline{\widehat{\beta}}. \tag{2.6}$$

The variance of this predictor is a function of $\underline{d}$, where $\underline{d}^T = (d_1, ..., d_k)$:

$$var(\widehat{y}|\ \underline{d}) = \left(1,\ \underline{d}^T\right) cov(\underline{\hat{\beta}}) \begin{pmatrix} 1 \\ \underline{d} \end{pmatrix}. \tag{2.7}$$

Notice that $\underline{d}$ in (2.7) may correspond with either one of the old N points in $\mathbf{X}$ - as in (2.6) - or a new point; a new point means interpolation or extrapolation.

To illustrate the implications of (2.7), suppose that the design leading to $\mathbf{X}$ is orthogonal; that is, $\mathbf{X}^T\mathbf{X} = N\mathbf{I}_{q\times q}$. Combining (2.4) and (2.7) then gives

$$var(\widehat{y}|\ \underline{d}) = \frac{\sigma_\varepsilon^2}{N}\left(1,\ \underline{d}^T\right) \begin{pmatrix} 1 \\ \underline{d} \end{pmatrix}. \tag{2.8}$$

Obviously, (2.8) shows that the regression predictor becomes less reliable (accurate), as the number of observations N decreases, or the noise σ_ε^2 increases . In Appendix 1 we derive $\underline{d}_0 = -\mathbf{C}^{-1}\underline{b}$ where $\underline{d}_0$ is the design point that minimizes the variance of the regression predictor, and where $\mathbf{C}$ and $\underline{b}$ follow from

$$cov(\underline{\hat{\beta}}) = \sigma_\varepsilon^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma_\varepsilon^2 \begin{pmatrix} a & \underline{b}^T \\ \underline{b} & \mathbf{C} \end{pmatrix}$$

where a is a scalar, $\underline{b}$ a $k \times 1$ vector, and $\mathbf{C}$ a $k \times k$ matrix. (Note that $\sigma_\varepsilon^2\mathbf{C}$ is the covariance matrix of $\underline{\widehat{\beta}}_{-0}$ where $\underline{\widehat{\beta}}_{-0}$ equals $\underline{\widehat{\beta}}$ excluding the intercept $\widehat{\beta}_0$.) If the design is orthogonal, then (2.8) is minimal at the center of the experimental area: $\underline{d}_0 = \underline{0}$ (also see the 'funnel' shape of Figure 2.1, discussed below). Hence, extrapolation should be less trusted as the extrapolated point moves farther away into regions not yet observed; this property will guide our ASA. (The term "trust region" is used in nonlinear optimization; see Conn, Gould, and Toint (2000).)

2.3 Two new search techniques

We consider a lower, one-sided $1-\alpha$ confidence interval for the predictor based on (2.2), given $\underline{d}$. This interval ranges from infinity down to

$$\widehat{y}_{\min}(\underline{d}) = \widehat{y}(\underline{d}) - t_{N-q}^{\alpha}\widehat{\sigma}(\widehat{y},\ \underline{d}) = \left(1,\ \underline{d}^T\right)\underline{\hat{\beta}} - t_{N-q}^{\alpha}\widehat{\sigma}_\varepsilon\left(\left(1,\ \underline{d}^T\right)(\mathbf{X}^T\mathbf{X})^{-1}\begin{pmatrix} 1 \\ \underline{d} \end{pmatrix}\right)^{1/2} \tag{2.9}$$

where t^{α}_{N-q} denotes the $1-\alpha$ quantile of the t distribution with $N-q$ degrees of freedom, and (2.4) through (2.7) lead to

$$\widehat{\sigma}(\widehat{y},\ \underline{d}) = (\widehat{var}[\widehat{y}(\underline{d})|\ \underline{d}])^{1/2} = \widehat{\sigma}_{\varepsilon}\left(\left(1,\ \underline{d}^{T}\right)(\mathbf{X}^{T}\mathbf{X})^{-1}\begin{pmatrix}1\\ \underline{d}\end{pmatrix}\right)^{1/2}.$$

The first term in (2.9) concerns the signal, whereas the second term concerns the noise.

When we consider a set of $\underline{d}$ values, then the set of intervals following from (2.9) has a joint (simultaneous) probability lower than $1-\alpha$. This complication is ignored in our two techniques.

- Technique 1 (ASA) finds (say) $\underline{d}^{+}$ which is the $\underline{d}$ that maximizes the minimum output predicted through (2.9). This $\underline{d}^{+}$ gives both a search direction and a step size. First we prove in Appendix 2 that the objective function in (2.9) is concave in $\underline{d}$. Next in Appendix 3 we derive the following explicit solution for the optimal input values of the next observation:

$$\underline{d}^{+} = -\mathbf{C}^{-1}\underline{b} + \lambda\mathbf{C}^{-1}\widehat{\underline{\beta}}_{-0} \tag{2.10}$$

  where $\mathbf{C}^{-1}\widehat{\underline{\beta}}_{-0}$ is the ASA direction, and λ the step size specified by

$$\lambda = \left(\frac{a - \underline{b}^{T}\mathbf{C}^{-1}\underline{b}}{(t^{\alpha}_{N-q}\widehat{\sigma}_{\varepsilon})^{2} - \widehat{\underline{\beta}}^{T}_{-0}\mathbf{C}^{-1}\widehat{\underline{\beta}}_{-0}}\right)^{1/2}. \tag{2.11}$$

We point out that when deriving this step size, we assume that the local regression model provides some guidance outside the local region currently explored. (Angün et al. (2003) uses the local regression model to make a relatively big step along the search path, and then check whether that step should be reduced.)

- Technique 2 still maximizes $\widehat{y}_{\min}(\underline{d})$, but the new point is restricted to the SA path; that is, the search direction is specified by the estimated local gradient, $\widehat{\underline{\beta}}_{-0}$. In Appendix 4, we derive the optimal step size (say) ζ^{+} along this path:

$$\zeta^{+} = \left(\frac{a - \underline{b}^{T}\mathbf{C}^{-1}\underline{b}}{\left(\frac{\widehat{\underline{\beta}}^{T}_{-0}\mathbf{C}\widehat{\underline{\beta}}_{-0}}{\widehat{\underline{\beta}}^{T}_{-0}\widehat{\underline{\beta}}_{-0}}t^{\alpha}_{N-q}\widehat{\sigma}_{\varepsilon}\right)^{2} - \widehat{\underline{\beta}}^{T}_{-0}\mathbf{C}\widehat{\underline{\beta}}_{-0}}\right)^{1/2}. \tag{2.12}$$

We derive the following mathematical properties and interpretations of these two techniques.

The first term in (2.10) means that the ASA path starts from the point with minimal predictor variance, namely $-\mathbf{C}^{-1}\underline{b}$ (also see end of §2.2). The second term means that the ASA path adjusts the classic SA direction $\underline{\widehat{\beta}}_{-0}$ (second term's last factor) through the covariance matrix of $\underline{\widehat{\beta}}_{-0}$, which is $\sigma_\varepsilon^2\mathbf{C}$ (see §2.2, last paragraph). Finally, the step size λ is quantified in (2.11).

For the orthogonal case (i.e., $(\mathbf{X}^T\mathbf{X}) = N\mathbf{I}_{q\times q}$)), it is easy to verify that $a = 1/N$, $\underline{b} = \underline{0}$, and $\mathbf{C} = \mathbf{I}/N$, so (2.10) reduces to

$$\underline{d}^+ = \frac{1}{\left(\frac{(t^\alpha_{N-q}\widehat{\sigma}_\varepsilon)^2}{N} - \underline{\widehat{\beta}}^T_{-0}\underline{\widehat{\beta}}_{-0}\right)^{1/2}}\underline{\widehat{\beta}}_{-0}. \tag{2.13}$$

This solution implies identical search directions for ASA and SA in case of orthogonality. Moreover, for the orthogonal case we prove in Appendix 4 that the two techniques coincide (both the search direction and the step size are the same), provided SA starts from the design center.

In practice, however, designs are not orthogonal. The classic textbooks on Design Of Experiments (DOE) and RSM do present many orthogonal designs (for example, 2^{k-p} designs), but these designs use standardized inputs (say) t_j; that is, inputs ranging between -1 and $+1$, with an average value of zero. In practice, we apply the following linear transformation to obtain original inputs d_j that range between L_j and H_j:

$$d_j = f_j + g_j t_j \text{ with } f_j = \frac{L_j + H_j}{2};\ g_j = \frac{L_j - H_j}{2}. \tag{2.14}$$

Consequently, the first-order polynomial regression model (2.2) implies that β_j and φ_j - the main effects of the original and standardized inputs respectively - are related as follows: $\beta_j = \varphi_j/g_j$. Hence, the steepest ascent path directions for the original and the standardized inputs differ (unless $\forall j : g_j = 1$). (The interpretation of standardization is controversial in mathematical statistics; see the many references in Kleijnen (1987, pp. 221, 345).)

We prove in Appendix 5 that ASA is scale independent. So ASA is not affected by switching from (say) inches to centimeters when measuring inputs. Driessen et al. (2001) proves that ASA is also independent of linear transformations with $f_j \neq 0$ in (2.14).

In case of large signal/noise ratios (defined in (2.1)), the denominator under the square root in (2.11) is negative. So this equation does not give a finite solution for $\underline{d}^+$; that is, (2.9) can be driven to infinity (unbounded solution). Indeed, if the noise is negligible, we have a deterministic problem, which our technique is not meant

to address (many other researchers - including Conn, Gould, and Toint (2000) - study optimization of deterministic simulation models).

In case of a small signal/noise ratio, no step is taken. Actually, we distinguish two cases: (i) the signal is small, (ii) the noise is big. These two cases are discussed next.

In case (i), the signal may be small because the first-order polynomial approximation is bad. Then we should switch to an alternative metamodel using transformations of d_j such as $\log(d_j)$ or $1/d_j$ (inexpensive alternative), a second-order polynomial, which adds d_j^2 and $d_j d_{j'}$ with $j' > j$ (expensive because many more observations are required to estimate the corresponding effects), etc.; see the RSM literature (for example, Irizarry, Wilson, and Trevino (2001)).

In case (ii), however, the first-order polynomial may fit, but the intrinsic noise may be high (also see the comment below (2.8)). To decrease this noise, we should increase the number of observations, N; see the denominator in (2.8). Hence, we should increase either n or m_i (see the definitions below (2.3)). When our technique gives a value $\underline{d}^+$ that is "close" to one of the old points, then in practice we may increase m_i. Otherwise we observe a new combination: we increase n. So our technique suggests an approach to the old problem of how to choose between either using the next observation to increase the accuracy of the current local approximation, or trusting that approximation and moving into a new area. A different approach is discussed in Kleijnen (1975, p. 360). In the literature on maximizing the output of deterministic simulation, this is called the geometry improvement problem; see Conn, Gould, and Toint (2000). More research on this problem is needed.

If we specify a different α value in t^{α}_{N-q}, then (2.11) gives a different step size (in the same direction). Obviously, t^{α}_{N-q} increases to infinity, as α decreases to zero. So, a sufficiently small α always gives a finite solution. However, if we increase α, then we make a bigger step, and we prefer to take a bigger step in order to get quicker to the top of the response surface. We feel that a reasonable maximum α value is 0.20 (so we are "80% sure"), since 0.20 corresponds to the maximum of the common values of α (i.e., 0.01, 0.05, 0.10, and 0.20) among practitioners; however, more empirical research is needed.

We assume that the noise ε has zero mean when deriving the $1-\alpha$ confidence interval in (2.9), which leads to the techniques in (2.10), (2.11), and (2.12). Actually, the locally fitted first-order polynomials may show lack of fit so the expected value of $\widehat{\sigma}^2_{\varepsilon}$ exceeds σ^2_{ε}; see the lack-of-fit tests in many RSM textbooks. Fortunately, this bias has the "right" sign; that is, this bias increases $\widehat{\sigma}_{\varepsilon}$ in (2.11) and (2.12) so that it

decreases the step size.

2.4 Numerical examples of step-size selection

To obtain a better understanding of ASA - especially its step size - we apply this technique to the following three numerical examples:

(i) single input, and orthogonal design;

(ii) two inputs, and orthogonal design;

(iii) two inputs, and one-at-a-time design.

For each example, we study several cases; that is, different signal/noise ratios. We suppose that the regression estimates happen to equal the true values: $\widehat{\underline{\beta}} = \underline{\beta}$ and $\widehat{\sigma}_\varepsilon = \sigma_\varepsilon$. Without loss of generality, we take $\beta_0 = 0$ and $\sigma_\varepsilon = 1$ (σ_ε and $\mathbf{X}$ determine the noise of $\widehat{\underline{\beta}}$; see (2.4)). We start with example (iii), because in practice designs are not orthogonal (see our discussion of (2.14)), and one-at-a-time designs are popular non-orthogonal designs (nevertheless, we do not recommend one-at-a-time designs); then we summarize results for the other two examples with orthogonal designs.

We use a one-factor-at-a-time design with $\underline{d}_1^T = (-1,\ -1)$, $\underline{d}_2^T = (-1,\ 1)$ so $n = 3$ $(= q)$. To estimate σ_ε^2 through the MSR in (2.5), we duplicate combination 1: $m_1 = 2$ so $N = 4$. We consider two extreme signal/noise cases.

- Case 1, high/low signal/noise, $\widehat{\gamma}_1 = 10$ and $\widehat{\gamma}_2 = 0.10$: The given $\mathbf{X}$ and $\widehat{\sigma}_\varepsilon$ result in $\widehat{\beta}_1 = 6.124$ and $\widehat{\beta}_2 = 0.061$. (2.11) does not give a finite step size for the traditional α values 0.20, 0.10, and 0.05.

- Case 2, low signal/noise, $\widehat{\gamma}_1 = 0.3$ and $\widehat{\gamma}_2 = 0.5$: This case implies $\widehat{\beta}_1 = 0.184$ and $\widehat{\beta}_2 = 0.306$. Then $\alpha = 0.20$ gives $(d_1^+,\ d_2^+) = (-0.404,\ -0.212)$; $\alpha = 0.05$ gives $(-0.4804,\ -0.4416)$: no move outside the local input area (the N old outputs were obtained for $-1 \leq d_j \leq 1$).

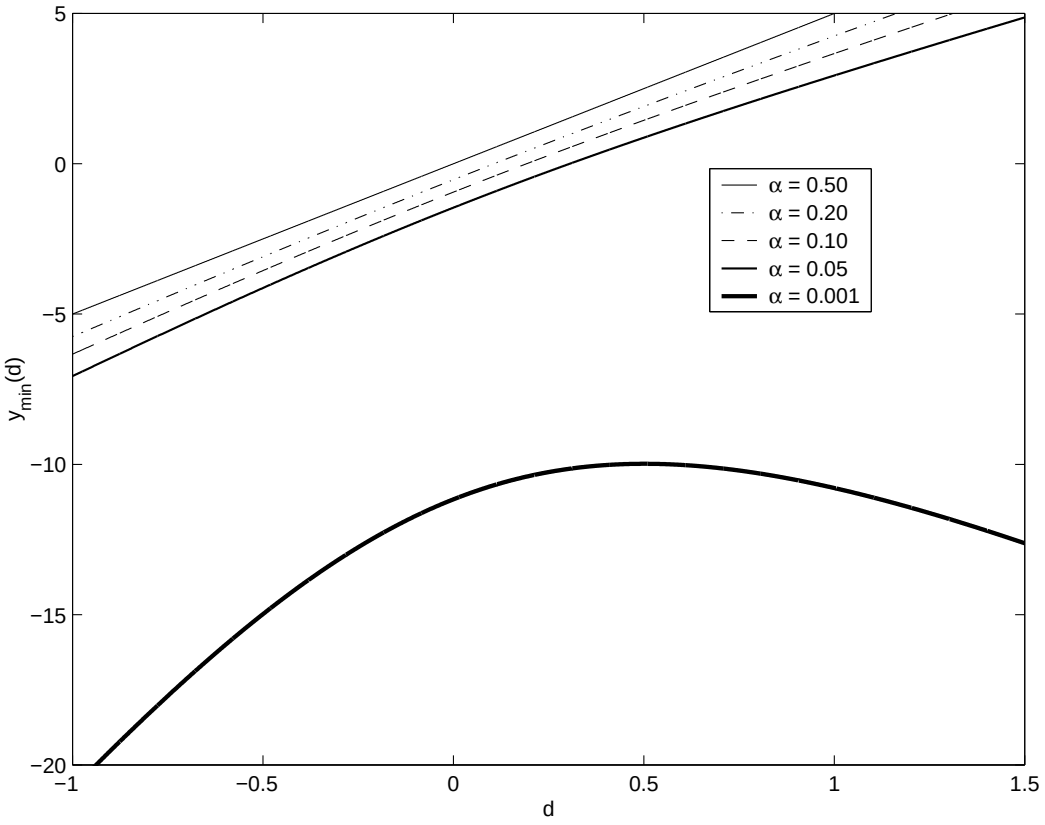
The other two examples can be summarized as follows.

- Example (i), single input and orthogonal design: Obviously, we now have $q = 2$. Suppose $\underline{d}_1^T = (1,\ -1)$ and $\underline{d}_2^T = (1,\ 1)$, so $n = 2$. Suppose $m_i = 2$, so $N = 4$. Then, (2.4) gives $\widehat{var}(\widehat{\beta}_1) = \widehat{\sigma}_\varepsilon^2/N = 1/4$. Now (2.13) reduces to

$$d_1^+ = sign(\widehat{\beta}_1)\sqrt{\frac{\widehat{\gamma}_1^2}{(t_2^\alpha)^2 - \widehat{\gamma}_1^2}},$$

which gives a finite solution if $t_2^\alpha > \widehat{\gamma}_1$.

Figure 2.1: High signal/noise case: lower $1-\alpha$ confidence intervals for the regression predictor $\widehat{y}_{\min}$ for different α



Consider a case with high signal/noise: $\widehat{\gamma}_1 = 10$; that is, $\widehat{\beta}_1 = 5$. A finite solution results only for $\alpha \leq 0.0049$; for example, $\alpha = 0.001$ gives such a solution in Figure 2.1 (where $\alpha \leq 0.50$ or $t_2^{0.50} = 0$ corresponds with $\widehat{y}_{\min}$ itself).

- Example (ii), two inputs and orthogonal design: A 2^k design gives $n = 4$; $m_i = 1$ implies $N = 4$. So (2.13) becomes

$$d_j^+ = sign(\widehat{\beta}_j)\sqrt{\frac{\widehat{\gamma}_j^2}{(t_1^\alpha)^2 - (\widehat{\gamma}_1^2 + \widehat{\gamma}_2^2)}} \text{ with } j = 1, 2.$$

We consider a case with high/low signal/noise: $\widehat{\gamma}_1 = 10$, $\widehat{\gamma}_2 = 0.10$; that is, $\widehat{\beta}_1 = 5$ and $\widehat{\beta}_2 = 0.05$. Then neither $\alpha = 0.20$ nor $\alpha = 0.10$ gives a finite solution. So we might split the inputs into two parts: a high signal/noise input 1 treated as in deterministic simulation, and a low signal/noise input 2. Further, $\alpha = 0.025$ does give a finite solution, namely $(d_1^+,\ d_2^+) = (1.2759,\ 0.0128)$; so the input with the highest signal changes almost 30%.

2.5 Comparison of the adapted steepest ascent (ASA) and steepest ascent search directions through Monte Carlo experiments

We perform some Monte Carlo experiments to compare the search directions of the two techniques, ASA and SA. The Monte Carlo method is an efficient and effective way to estimate the behavior of search techniques applied to random simulations (such as discrete-event dynamic systems, including simulated queuing and inventory systems); also see McDaniel and Ankenman (2000).

We experiment with two inputs: $k = 2$. Our Monte Carlo experiments generate output $\underline{\widehat{w}}$ (used in (2.3)) through second-order polynomials in two inputs with white noise:

$$\omega = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_{1,1} d_1^2 + \beta_{2,2} d_2^2 + \beta_{1,2} d_1 d_2 + \psi. \tag{2.15}$$

RSM fits first-order polynomials defined in (2.2) locally, and then estimates the SA. The global experimental area is the area over which the inputs of the real system can be varied, or the area over which the simulation model is assumed to be valid. We assume that this area is the unit square: $-1 \leq d_1 \leq 1$ and $-1 \leq d_2 \leq 1$. In the local area we use a specific design $\mathbf{D}$, namely a one-at-a-time design (as in §2.4). The specific local area is the upper corner of Figure 2.2, discussed below.

There are infinitely many polynomials that satisfy (2.15). We define the canonical case as $\beta_0 = \beta_1 = \beta_2 = \beta_{1,2} = 0$; $\beta_{1,1} = \beta_{2,2} = -1$; see the contour functions in Figure 2.2. In the canonical case, our local starting area is the upper-right corner (1, 1) with 0.2 input ranges: $\underline{d}_1^T = (1,\ 1)$, $\underline{d}_2^T = (1,\ 0.8)$, and $\underline{d}_3^T = (0.8,\ 1)$. Obviously, the true optimal input combination (say) $\underline{d}^*$ is $(0,\ 0)^T$.

For $\underline{d}_1^T = (1,\ 1)$ we observe the output $\underline{\widehat{w}}$ twice: $m_1 = 2$. After fitting the first-order polynomial, we start the search from $\underline{d}_0^T = (0.95,\ 0.95)$, which is the point that gives the minimal variance; see Appendix 1. In this Monte Carlo experiment we know that the truly optimal search direction is the vector (say) $\underline{e}$ that starts at $\underline{d}_0$ and ends at the true optimum $(0,\ 0)^T$; also see Figure 2.3 below. We estimate the angle (say) $\widehat{\theta}$ between this true search direction $\underline{e}$ and the estimated search direction $\underline{\widehat{v}}$, where $|.|$ and $\|.\|_2$ denote the absolute value and the two-norm, respectively:

$$\widehat{\theta} = \left| \arccos\left(\frac{\underline{e}^T \underline{\widehat{v}}}{\|\underline{e}\|_2 \, \|\underline{\widehat{v}}\|_2} \right) \right|. \tag{2.16}$$

So, the smaller the resulting $\widehat{\theta}$ is, the better the search technique performs.

Figure 2.2: Contour functions $E(\omega|\ d_1,\ d_2)$ with global and local experimental areas

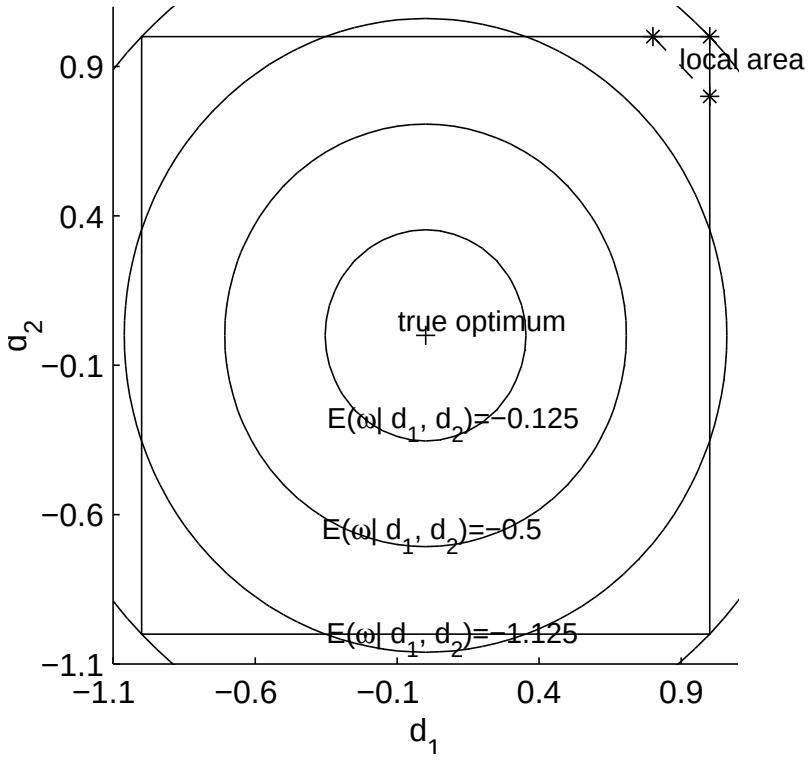
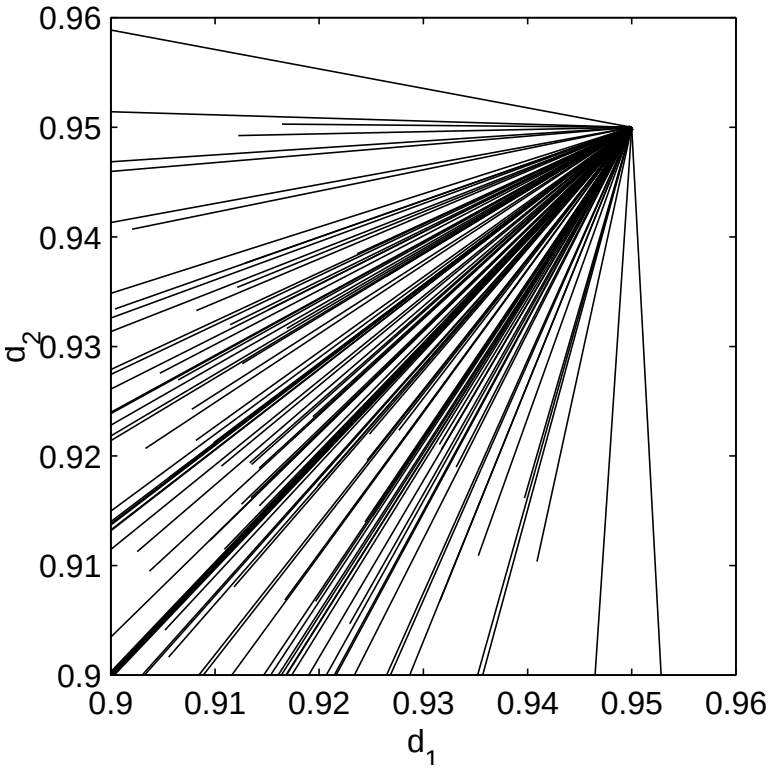


Figure 2.3: ASA's search directions when $\sigma_\psi = 0.10$



We take 100 macro-replicates. Each time we apply the two techniques to the same I/O data (d_1, d_2, $\underline{\widehat{w}}$). Then we compute the 100 search directions $\underline{\widehat{v}}$ for ASA; see Figure 2.3. We characterize $\widehat{\theta}$'s empirical distribution through its average, standard deviation, and specific quantiles. This gives Table 2.1 (left part), which demonstrates the superiority of ASA, unless we focus on the worst case for low variance or the 95% quantile for high variance.

Next we investigate the effects of σ_ψ (see ψ in (2.15)): we increase σ_ψ from 0.10 to 0.25. We use the same random numbers as we used for the smaller noise. Now the estimated search directions may be very wrong; see Figure 2.4. ASA still performs better, unless we focus on outliers: see the 95% or 100% quantiles in the right-hand part of Table 2.1.

Finally, we consider interaction between the two inputs: we take $\beta_0 = \beta_1 = \beta_2 = 0$; $\beta_{1,1} = -2$, $\beta_{2,2} = -1$, $\beta_{1,2} = 2$ in (2.15). We again find that ASA is better; see Table 2.2.

Table 2.1: Statistics for ASA and SA's estimated angle error $\widehat{\theta}$ for two noise values σ_ψ

Statistics	$\sigma_\psi = 0.10$		$\sigma_\psi = 0.25$	
	ASA	SA	ASA	SA
Average	1.83	89.97	19.83	89.84
Standard deviation	17.89	0.04	56.45	0.19
Median (50% quantile)	0.03	89.88	0.06	89.89
75% quantile	0	89.89	0.12	89.91
25% quantile	0.01	89.87	0.03	89.85
95% quantile	0.13	89.91	179.81	89.99
5% quantile	0	89.79	0.01	89.46
100% quantile	178.92	89.93	179.88	90.06
0% quantile	0	89.59	0	88.74

Figure 2.4: ASA's search directions when $\sigma_\psi = 0.25$

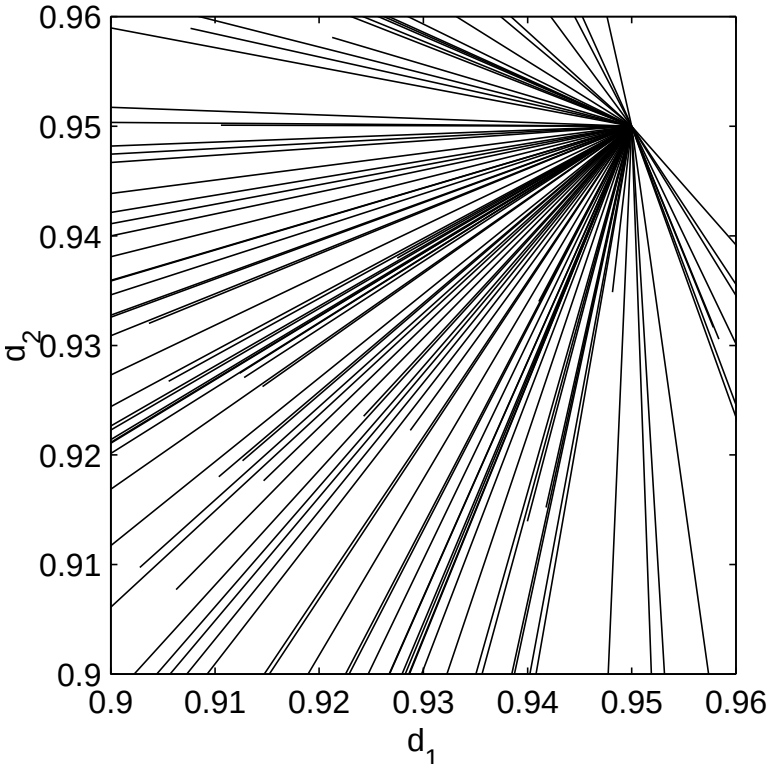


Table 2.2: Statistics in case of interaction, for ASA and SA's estimated angle error $\widehat{\theta}$, for two noise values σ_ψ

Statistics	$\sigma_\psi = 0.10$		$\sigma_\psi = 0.25$	
	ASA	SA	ASA	SA
Average	9.72	16.01	10.14	17.33
Standard deviation	3.3	6.23	7.69	12.88
Median (50% quantile)	9.68	16.02	8.99	14.94
75% quantile	12.37	21.12	16.13	27.87
25% quantile	6.99	10.76	3.21	5.84
95% quantile	15.66	27.05	24.78	41.55
5% quantile	4.99	6.8	0.61	0.81
100% quantile	17.41	30.08	32.07	50.99
0% quantile	0.85	1.46	0.04	0.25

2.6 Conclusions and future research

In this paper we addressed the problem of searching for the input combination that gives the maximum output. RSM is a classic technique for tackling this problem, but it has two well-known problems: (i) RSM uses steepest ascent (SA), which is scale dependent; (ii) RSM intuitively selects the step size on the SA path.

To address these two problems, we devised two new techniques. In technique 1 - called "adapted" SA or ASA - we select both a search direction and a step size. In technique 2, we use classic SA but we select a step size inspired by ASA.

Our main conclusion is that - because ASA is scale independent - it usually gives a better search direction than SA.

In future research, we may further investigate the step size, for which we proposed (2.11) and (2.12). Actually, Angün et al. (2003) explores this problem in the wider context of multiple responses.

2.7 Appendix 1: Derivation of the minimum variance of the regression predictor

The variance of the regression predictor $\widehat{y}$ at $(1, \underline{d}^T)$ follows from (2.4) and (2.7), where without loss of generality we take a unit variance, $\sigma_\varepsilon^2 = 1$:

$$var(\widehat{y}|\ \underline{d}) = (1, \underline{d}^T)(\mathbf{X}^T\mathbf{X})^{-1}\begin{pmatrix}1\\ \underline{d}\end{pmatrix}.$$

This can be rewritten as

$$var(\widehat{y}|\ \underline{d}) = a + 2\underline{b}^T\underline{d} + \underline{d}^T\mathbf{C}\underline{d}$$

where a, $\underline{b}$, and $\mathbf{C}$ are defined in §2.2.

Because $\mathbf{C}$ is positive definite, the necessary and sufficient condition for the point that gives minimal variance (say) $\underline{d}_0$ is

$$2\underline{b} + 2\mathbf{C}\underline{d}_0 = \underline{0}$$

which gives

$$\underline{d}_0 = -\mathbf{C}^{-1}\underline{b}.$$

If $\mathbf{X}$ is orthogonal, then $\underline{b} = \underline{0}$ so the variance is minimal at the design center: $\underline{d}_0 = \underline{0}$.

2.8 Appendix 2: Proof of the concavity of the objective function (2.9)

In (2.9), the first term $(1, \underline{d}^T)\widehat{\underline{\beta}}$ is linear and the second term has the positive factors t^α_{N-q} and $\widehat{\sigma}_\varepsilon$. Hence, it suffices to show that

$$\left((1, \underline{d}^T)\,(\mathbf{X}^T\mathbf{X})^{-1}\begin{pmatrix}1\\ \underline{d}\end{pmatrix}\right)^{1/2}$$

is convex in $\underline{d}$.

If in this expression the factor $(\mathbf{X}^T\mathbf{X})$ is not orthogonal, then we orthogonalize it through the well-known Gram-Schmidt QR method. It can be proven that convexity is preserved by linear transformations, so it suffices to show convexity for the orthogonal case.

If $(\mathbf{X}^T\mathbf{X})$ is orthogonal, then it suffices to show that

$$f(\underline{d}) = \left(1 + \underline{d}^T\underline{d}\right)^{1/2}$$

is convex in $\underline{d}$. Obviously, we have

$$\nabla f(\underline{d}) = \frac{1}{\left(1 + \underline{d}^T\underline{d}\right)^{1/2}}\underline{d}$$

and

$$\nabla^2 f(\underline{d}) = \left(1 + \underline{d}^T\underline{d}\right)^{-3/2}\left[\mathbf{I} + (\underline{d}^T\underline{d})\mathbf{I} - \underline{d}\underline{d}^T\right].$$

In this expression, $(\underline{d}^T\underline{d})\mathbf{I} - \underline{d}\underline{d}^T$ is positive semi-definite: for all (say) $\underline{u}$ we have

$$\begin{aligned}\underline{u}^T\left[(\underline{d}^T\underline{d})\mathbf{I} - \underline{d}\underline{d}^T\right]\underline{u} &= \|\underline{d}\|^2\|\underline{u}\|^2 - (\underline{d}^T\underline{u})^2 \\ &\geq \|\underline{d}\|^2\|\underline{u}\|^2 - \|\underline{d}\|^2\|\underline{u}\|^2 = \underline{0}.\end{aligned}$$

This means that $\nabla^2 f(\underline{d})$ is positive semi-definite; hence, $f(\underline{d})$ is convex. Consequently, (2.9) is an "easy" problem; that is, the local maximum is the global maximum.

2.9 Appendix 3: Maximization of the objective function (2.9)

We rewrite (2.9) as

$$\widehat{\beta}_0 + \underline{\widehat{\beta}}_{-0}^T\underline{d} - t_{N-q}^{\alpha}\widehat{\sigma}_{\varepsilon}\left(a + 2\underline{b}^T\underline{d} + \underline{d}^T\mathbf{C}\underline{d}\right)^{1/2}$$

where $\underline{\widehat{\beta}}_{-0}$, a, $\underline{b}$, and $\mathbf{C}$ are defined in §2.2. Since this function is concave (see Appendix 2), the necessary and sufficient first-order conditions for the maximizing point $\underline{d}^+$ are

$$\underline{\widehat{\beta}}_{-0} - \frac{t_{N-q}^{\alpha}\widehat{\sigma}_{\varepsilon}}{\left(a + 2\underline{b}^T\underline{d}^+ + \underline{d}^{+T}\mathbf{C}\underline{d}^+\right)^{1/2}}\left(\underline{b} + \mathbf{C}\underline{d}^+\right) = \underline{0}.$$

Substituting

$$\underline{d}^+ = -\mathbf{C}^{-1}\underline{b} + \lambda\mathbf{C}^{-1}\underline{\widehat{\beta}}_{-0}$$

in which λ is an unknown scalar, we get

$$\lambda = \left(\frac{a - \underline{b}^T\mathbf{C}^{-1}\underline{b}}{(t_{N-q}^{\alpha}\widehat{\sigma}_{\varepsilon})^2 - \underline{\widehat{\beta}}_{-0}^T\mathbf{C}^{-1}\underline{\widehat{\beta}}_{-0}}\right)^{1/2}.$$

For an orthogonal design (implying $a = 1/N$, $\underline{b} = \underline{0}$, and $\mathbf{C} = \mathbf{I}/N$), this equation simplifies to

$$\lambda = \left(\frac{1/N}{(t^{\alpha}_{N-q}\widehat{\sigma}_{\varepsilon})^2 - N\underline{\widehat{\beta}}^T_{-0}\underline{\widehat{\beta}}_{-0}} \right)^{1/2}.$$

Hence for an orthogonal design the new point is

$$\underline{d}^{+} = \frac{1}{\left(\frac{(t^{\alpha}_{N-q}\widehat{\sigma}_{\varepsilon})^2}{N} - \underline{\widehat{\beta}}^T_{-0}\underline{\widehat{\beta}}_{-0} \right)^{1/2}} \underline{\widehat{\beta}}_{-0}.$$

2.10 Appendix 4: Optimization of the step size in SA

We assume that the SA path starts from $\underline{d}_0 = -\mathbf{C}^{-1}\underline{b}$, which is the point at which the predictor variance is minimal; if $\mathbf{X}$ is orthogonal, then $\underline{b} = \underline{0}$ so $\underline{d}_0 = \underline{0}$ (see Appendix 1). In SA, we make a step of size (say) ζ in the $\underline{\widehat{\beta}}_{-0}$ direction. This means

$$\underline{d}^{+} = -\mathbf{C}^{-1}\underline{b} + \zeta\underline{\widehat{\beta}}_{-0}.$$

Substitution of $(1, \underline{d}^{+T})$ into the regression predictor (2.9) gives

$$\widehat{y}_{\min}(\underline{d}^{+}) = \widehat{\beta}_0 - \underline{b}^T\mathbf{C}^{-1}\underline{\widehat{\beta}}_{-0} + \zeta\underline{\widehat{\beta}}^T_{-0}\underline{\widehat{\beta}}_{-0} - t^{\alpha}_{N-q}\widehat{\sigma}_{\varepsilon}\left(a - \underline{b}^T\mathbf{C}^{-1}\underline{b} + \zeta^2\underline{\widehat{\beta}}^T_{-0}\mathbf{C}\underline{\widehat{\beta}}_{-0}\right)^{1/2}.$$

Since this expression is concave in ζ , it is easy to verify that ζ^{+} defined in (2.12) indeed maximizes $\widehat{y}_{\min}(\underline{d}^{+})$.

Comparison with Appendix 3 proves that in the orthogonal case the two techniques coincide - provided SA starts from the design center.

2.11 Appendix 5: Scale independence of ASA

Affine scaling implies $\mathbf{X} = \mathbf{ZA}$ with a non-singular square matrix $\mathbf{A}$ so $\mathbf{Z} = \mathbf{XA}^{-1}$. Hence (2.9) expressed in $\underline{z}$ becomes

$$\underline{z}^T\underline{\widehat{\beta}}_{\underline{z}} - t^{\alpha}_{N-q}\widehat{\sigma}_{\varepsilon}\left(\underline{z}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\underline{z}\right)^{1/2} \tag{2.17}$$

where

$$\underline{\widehat{\beta}}_{\underline{z}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\underline{\widehat{w}}$$

or

$$\begin{aligned}\widehat{\underline{\beta}}_{\underline{z}} &= \left(\mathbf{A}^{-T}\mathbf{X}^T\mathbf{X}\mathbf{A}^{-1}\right)^{-1}\mathbf{A}^{-T}\mathbf{X}^T\underline{\widehat{w}} \\ &= \mathbf{A}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{A}^T\mathbf{A}^{-T}\mathbf{X}^T\underline{\widehat{w}} \\ &= \mathbf{A}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\underline{\widehat{w}} \\ &= \mathbf{A}\underline{\widehat{\beta}}_{\underline{d}}.\end{aligned}$$

In (2.17) we write the square-root factor as

$$\begin{aligned}\left(\underline{z}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\underline{z}\right)^{1/2} &= \left[\underline{z}^T\left(\left(\mathbf{X}\mathbf{A}^{-1}\right)^T\left(\mathbf{X}\mathbf{A}^{-1}\right)\right)^{-1}\underline{z}\right]^{1/2} \\ &= \left[\underline{z}^T\left(\mathbf{A}^{-T}\left(\mathbf{X}^T\mathbf{X}\right)\mathbf{A}^{-1}\right)^{-1}\underline{z}\right]^{1/2} \\ &= \left(\underline{z}^T\mathbf{A}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{A}^T\underline{z}\right)^{1/2}.\end{aligned}$$

Hence (2.17) becomes

$$\begin{aligned}&\underline{z}^T\mathbf{A}\widehat{\underline{\beta}}_{\underline{d}} - t^{\alpha}_{N-q}\widehat{\sigma}_{\varepsilon}\left(\underline{z}^T\mathbf{A}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{A}^T\underline{z}\right)^{1/2} \\ &= (\mathbf{A}^T\underline{z})^T\widehat{\underline{\beta}}_{\underline{d}} - t^{\alpha}_{N-q}\widehat{\sigma}_{\varepsilon}\left((\mathbf{A}^T\underline{z})^T(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{A}^T\underline{z})\right)^{1/2} \\ &= \left(1,\, \underline{d}^T\right)\widehat{\underline{\beta}}_{\underline{d}} - t^{\alpha}_{N-q}\widehat{\sigma}_{\varepsilon}\left(\left(1,\, \underline{d}^T\right)(\mathbf{X}^T\mathbf{X})^{-1}\begin{pmatrix}1\\ \underline{d}\end{pmatrix}\right)^{1/2}\end{aligned}$$

which is indeed identical to the original expression in (2.9). This proves that ASA is scale independent.

Chapter 3

RSM with Stochastic Constraints for Expensive Simulation

3.1 Introduction

Optimization in simulation is attempted by many methods; see Fu (2002). In particular, when simulation is treated as a black box (that is, gradient estimation through perturbation analysis (Glasserman (1991)) or through score function (Rubinstein and Shapiro (1993)) is not applicable), metaheuristics such as simulated annealing, genetic algorithms, tabu search, or scatter search can be used for optimization. Another approach is the simultaneous perturbation stochastic approximation of Spall (2003), and Fu and Hill (1997) for black box simulation optimization. In this paper, we focus on Response Surface Methodology (RSM), which also treats simulation as a black box.

From the practitioners' point of view, RSM is broadly applicable since it can be easily integrated into both stochastic and deterministic simulations; for the optimization of stochastic systems through RSM, see Yang and Tseng (2002), and Irizarry, Wilson, and Trevino (2001), and for the optimization of a deterministic system through RSM, see Ben-Gal and Bukchin (2002).

Originally, RSM was derived for problems with a stochastic objective function and deterministic box constraints; Myers and Montgomery (2002) gave a general description for the first stage of this conventional RSM as follows. An experimental design is used to fit locally a first-order polynomial to the observed values of the stochastic objective through least squares. Then, a steepest descent search direction is estimated from the resulting model, and a number of steps are taken along the estimated steepest descent direction until no additional decrease in objective is evident.

This procedure is repeated until a first-order polynomial becomes an inadequate model, which is indicated when the gradient is not significantly different from zero.

In practice, however, optimization problems may have stochastic constraints in addition to deterministic box constraints. In RSM, there have been several approaches to solve constrained optimization problems. Khuri (1996) surveyed most of these approaches, including the desirability function (Harrington (1965), and Derringer and Suich (1980)), the generalized distance (Khuri and Conlon (1981)), and the dual response (Myers and Carter (1973), Vining and Myers (1990), Del Castillo and Montgomery (1993), and Fan and Del Castillo (1999)). Further, Wei, Olson, and White (1990) suggested another approach, namely the prediction-interval constrained goal programming, which was not mentioned in Khuri (1996). After we shall have described the methods cited in Khuri (1996) and the prediction-interval constrained goal programming in Wei, Olson, and White (1990), we will explain a major drawback of these methods.

Harrington (1965), and Derringer and Suich (1980) suggested the approach called the desirability function. They transformed each predicted response to a desirability index, and combined the transformed desirability indices into a single response function. In this way, they reduced the multiresponse optimization problem to one with a single response and no constraints. Then, Derringer and Suich (1980) optimized the overall desirability function using a pattern search method.

Later, Wei, Olson, and White (1990) compared the desirability approach with prediction-interval constrained goal programming. They minimized the sum of the weighted deviations between each response variable and its predetermined target value. They concluded that their approach gave more flexibility to the users.

Khuri and Conlon (1981) considered the generalized distance approach. They first optimized each response individually over the experimental region. Then, they defined a distance function that measures the deviations between each predicted response and its individual optimum. By minimizing the distance function, they obtained a so-called compromise optimum. Thus, the multiresponse optimization is again converted into a single response unconstrained optimization.

Myers and Carter (1973), Vining and Myers (1990), Del Castillo and Montgomery (1993), and Fan and Del Castillo (1999) applied the dual response or Taguchian approach. They locally modelled the same problem in three different ways: i) minimize the mean response subject to the variance of the response attaining a target value; ii) maximize the mean response subject to the variance of the response attaining a target value; iii) minimize the variance of the response subject to the mean response attaining

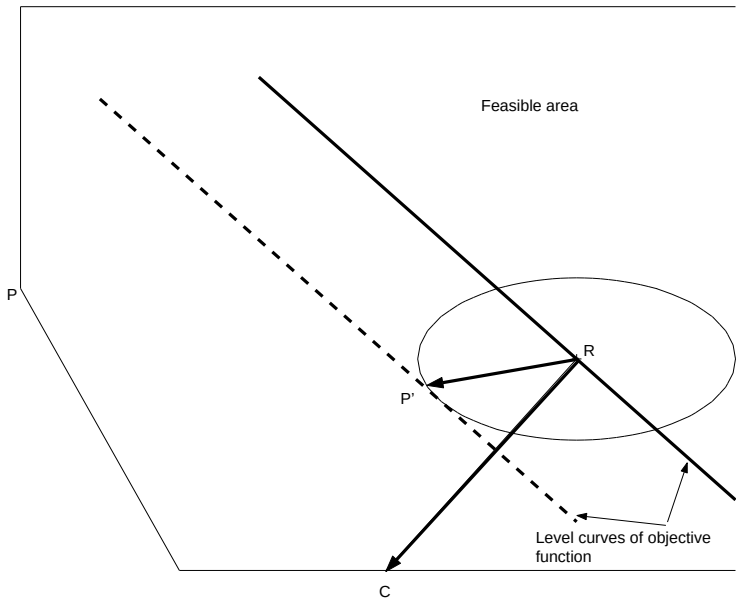
a target value. To solve dual response problems, Myers and Carter (1973), and Vining and Myers (1990) applied the Lagrangian method; Del Castillo and Montgomery (1993) used the generalized reduced gradient method.

In all these approaches, the constrained optimization problem is reformulated by combining the constraints and the original objective function into a new, single objective function by appropriate transformations. The type of transformation differs with the particular method. Next, the resulting unconstrained problem is solved using an ordinary nonlinear programming algorithm. These methods suffer from a major drawback: these transformations require arbitrary choices. To overcome this drawback, we propose an alternative approach. Rather than transforming it into an unconstrained problem, we focus on the original (possibly stochastically) constrained optimization problem.

We now introduce our approach for extending conventional RSM to handle stochastic constraints. This is achieved through the generalization of the estimated steepest descent search direction using ideas from interior point methods, more specifically the affine scaling algorithm. To obtain better insight, we illustrate the superiority of our search direction through a deterministically constrained problem, as follows. In Figure 3.1, which is inspired by Ye (1989, p. 51), suppose that the current input vector is R and the optimal point is P. Suppose further that although there are constraints, we would use conventional RSM's estimated steepest descent direction. Then, this search direction would soon hit the boundary at the input vector C. In this way, we would have slow convergence, creeping along the boundary. Instead, we propose a search direction that generates an input vector which avoids hitting the boundary. We accomplish this by introducing an ellipsoid constraint centered at the current iterate with a shape determined by the current values of the slacks. By minimizing the linearly approximated objective function over this ellipsoid, we obtain the next input vector, P'. This gives the search direction RP'. An explicit formula for the search direction obtained by this optimization is well-known in the interior point methods literature and is given in Barnes (1986), and also in Appendix 1.

The main contribution of this paper is to generalize the estimated steepest descent search direction for dealing with stochastically constrained problems. Our proposed search direction is the scaled and projected estimated steepest descent, also called the estimated affine scaling search direction. To achieve this generalization, we use standard tools from interior point methods and nonlinear programming such as scaling and projection. Scaling is a standard feature of the affine scaling algorithm. Projection of the search direction is also a standard approach when there are linear

Figure 3.1: Proposed search direction RP' versus steepest descent RC



constraints; there is a related family of algorithms called null-space methods discussed in Gill, Murray, and Wright (2000).

We prove that the proposed search direction has two important features: it is indeed a descent direction, and it is scale independent. On the other hand, it is well-known that the estimated steepest descent direction, used in conventional RSM, is scale dependent. In general, practitioners need to deal with variables with widely varying orders of magnitude. In such cases, scale independence enables them to avoid numerical complications and problems.

Furthermore, we provide a heuristic for the first stage of RSM that uses the proposed search direction iteratively. This heuristic is primarily meant for expensive simulation-based optimization problems with a tight simulation budget to reach quickly a desired neighborhood of the real optimum; that is, for studies in which each simulation run takes, say several hours or days, and the total simulation budget is limited, so that only a small number of simulation runs are available. Such problems arise in practice in diverse fields; see for example Helton et al. (1997), den Hertog and Stehouwer (2002), and Vonk Noordegraaf, Nielen, and Kleijnen (2003). In particular, the heuristic enables us to cope with optimization problems with a stochastic objective function and stochastic constraints, besides deterministic box constraints. However, the heuristic can also be easily simplified for deterministic problems that are solved iteratively.

In the heuristic, we proceed towards the solution point through the interior of the feasible region. This approach is inspired by interior point methods and has two advantages. First, interior point methods are known to perform well within the interior of the feasible region; therefore, we expect that our heuristic will also avoid creeping along the boundary. Second, in practice some simulation programs may simply crash or become invalid outside the feasible region. Such problems were reported in Bettonvil and Kleijnen (1996), Booker et al. (1999), and den Hertog and Stehouwer (2002).

We use underlined letters and bold letters to denote vectors and matrices, respectively. Hence, $\underline{0}$, $\underline{1}$, $\mathbf{0}$, and $\mathbf{1}$ stand for a vector of zeros, a vector of ones, a square matrix of zeros, and a square matrix of ones, respectively. For square matrices, we use subscripts to denote their dimensions; i.e., $\mathbf{I}_k$ denotes a $k \times k$ identity matrix.

The remainder of this paper is organized as follows. In §3.2, we formalize our problem including statistical methods. In §3.3, we give the proposed search direction and outline its properties. In §3.4, we describe our heuristic for the first stage of RSM. In §3.5, we evaluate our heuristic through a Monte Carlo example. In §3.6, we give conclusions. There are three appendices with technical details and proofs.

3.2 Problem formulation

We consider the following problem:

$$\begin{array}{ll} \text{minimize} & E_{\underline{\omega}}\left[F_0\left(\underline{d},\ \underline{\omega}\right)\right] \\ \text{subject to} & E_{\underline{\omega}}\left[F_j\left(\underline{d},\ \underline{\omega}\right)\right] \geq a_j \quad \text{for} \quad j = 1,\ ...,\ r-1 \\ & \underline{l} \leq \underline{d} \leq \underline{u} \end{array} \tag{3.1}$$

where $E_{\underline{\omega}}$ is the expectation operator with respect to the simulation's seed vector $\underline{\omega}$, $\underline{d}$ is $k \times 1$ vector of input variables, a_j is the jth component of $(r-1) \times 1$ deterministic right-hand-side vector, $\underline{l}$ is $k \times 1$ vector of lower bounds on $\underline{d}$, $\underline{u}$ is $k \times 1$ vector of upper bounds on $\underline{d}$, and the F_i $(i = 0,\ ...,\ r-1)$ are r random responses that are estimated through simulation. We assume that the F_i are continuous and continuously differentiable on the "feasible" set defined by the inequalities in (3.1). For deterministic problems where the F_i are not known explicitly, $\underline{\omega}$ and hence the expectation operator $E_{\underline{\omega}}$ will simply vanish in (3.1).

The formulation in (3.1) covers expected values, variances, coefficients of variation, and probabilities. Note that probabilities can be formulated as expected values of indicator functions. However, indicator functions are not continuous. Other relevant performance measures are quantiles; there exist more general formulations than

(3.1) that enable us to consider quantiles as responses, but for simplicity we use the formulation in (3.1).

Let n ($n \geq k+1$) be the number of distinct simulated input combinations per local area - usually specified by a resolution-3 design in the first stage of RSM. Further, let m_t be the number of replicates at the tth ($t = 1,\ ...,\ n$) input combination, and let N be the resulting total number of local runs; that is, $N = m_1 + ... + m_t + ... + m_n$. In the first stage of RSM, we locally approximate the r simulation responses F_i in (3.1) by first-order polynomial regression metamodels with additive noises:

$$\underline{G}_i\left(\underline{d},\ \underline{\omega}\right) = \mathbf{X}\underline{\beta}_i + \underline{\epsilon}_i\left(\underline{d},\ \underline{\omega}\right) \quad \text{for } i = 0,\ ...,\ r-1 \tag{3.2}$$

where $\mathbf{X}$ is $N \times (k+1)$ matrix obtained by adding a column of ones to the design matrix, say, $\mathbf{D}$ (i.e., $\mathbf{X} = (\underline{1},\ \mathbf{D})$), $\underline{\beta}_i$ denotes $(k+1) \times 1$ vector of regression coefficients for response type i, and $\underline{\epsilon}_i$ is $N \times 1$ vector of additive noise with mean vector $\underline{\mu}_{\underline{\epsilon}_i}$ and covariance matrix $\mathbf{\Sigma}_{\underline{\epsilon}_i}$ for response type i. These $\underline{\epsilon}_i$ account for both the lack of fit of the first-order polynomial approximations and the inherent randomness in stochastic simulation created through the use of the simulation's seed $\underline{\omega}$.

Classic RSM assumes a single response ($r = 1$), and $\underline{\mu}_{\underline{\epsilon}_i} = \underline{0}$ and $\mathbf{\Sigma}_{\underline{\epsilon}_i} = \sigma_{i,i}\mathbf{I}$. In the rest of this paper, we call the assumptions $\underline{\mu}_{\underline{\epsilon}_i} = \underline{0}$ and $\mathbf{\Sigma}_{\underline{\epsilon}_i} = \sigma_{i,i}\mathbf{I}$ the ordinary least squares (OLS) assumptions. When these assumptions are satisfied, the Gauss-Markov theorem guarantees the OLS estimator of $\underline{\beta}_i$ to be the best linear unbiased estimator (BLUE), where "best" means minimum variance. In our problem, however, we have r responses that are statistically dependent (correlated); moreover, the variances per responses differ in general. Hence, generalized least squares (GLS) gives the BLUE of $\underline{\beta}_i$. However, under certain conditions GLS reduces to OLS, as we explain now.

Let $\underline{\widehat{F}}_i\left(\underline{d},\ \underline{\omega}\right) = \left(\underline{\widehat{F}}_i\left(\underline{d}_1,\ \underline{\omega}_1\right)^T,\ ...,\ \underline{\widehat{F}}_i\left(\underline{d}_t,\ \underline{\omega}_t\right)^T,\ ...,\ \underline{\widehat{F}}_i\left(\underline{d}_n,\ \underline{\omega}_n\right)^T\right)^T$ be $N \times 1$ vector of estimators of $E_{\underline{\omega}}\left[F_i\left(\underline{d},\ \underline{\omega}\right)\right]$, the expected response i in (3.1), where $\underline{d}_t$ is $m_t \times 1$ vector of replicates for the tth input combination and $\underline{\omega}_t$ is $m_t \times 1$ seed vector with independent components. Moreover, the seed vectors, say $\underline{\omega}_t$ and $\underline{\omega}_{t'}$ $\left(t' = 1,\ ...,\ n,\ t' \neq t\right)$, for distinct input combinations t and t' are independent; that is, we do not use common random numbers, so all N simulation outputs are independent. Moreover, we assume that each response i satisfies the OLS assumptions within the local area where we make N simulation runs (i.e., locally). Then, Rao (1967) and Ruud (2000, p. 703) prove that if the same design is used for all r responses and if each response satisfies the OLS assumptions, then the GLS estimator reduces to the OLS estimator of the individual coefficient vector $\underline{\beta}_i$:

$$\underline{\widehat{\beta}}_i\left(\underline{d},\ \underline{\omega}\right) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\underline{\widehat{F}}_i\left(\underline{d},\ \underline{\omega}\right) \text{ for } i = 0,\ ...,\ r-1. \tag{3.3}$$

Now, let $\mathbf{\Sigma}$ be $r \times r$ covariance matrix of responses with entry $(i,\ h)$ being $\sigma_{i,h}$ $(h = 0,\ ...,\ r-1)$. Often, simulation analysts estimate $\mathbf{\Sigma}$ through replicates. However, when simulation is time-consuming, we need to have a single replicate per input combination; that is, $m_1 = ... = m_t = ... = m_n = 1$. Therefore, we prefer an alternative estimator, namely the mean squared residual (MSR). Let $\widehat{\underline{G}}_i\left(\underline{d},\ \underline{\omega}\right) = \mathbf{X}\widehat{\underline{\beta}}_i\left(\underline{d},\ \underline{\omega}\right)$ be $N \times 1$ vector of OLS estimators of $E_{\underline{\omega}}\left[\underline{G}_i\left(\underline{d},\ \underline{\omega}\right)\right]$, the expectation of the linear approximation in (3.2). Indeed, Khuri (1996, p. 385) gives the following MSR estimator of $\sigma_{i,h}$:

$$\widehat{\sigma}_{i,h}\left(\underline{d},\ \underline{\omega}\right) = \frac{\left(\widehat{\underline{F}}_i\left(\underline{d},\ \underline{\omega}\right) - \widehat{\underline{G}}_i\left(\underline{d},\ \underline{\omega}\right)\right)^T \left(\widehat{\underline{F}}_h\left(\underline{d},\ \underline{\omega}\right) - \widehat{\underline{G}}_h\left(\underline{d},\ \underline{\omega}\right)\right)}{N - (k+1)}. \tag{3.4}$$

We assume constant covariances within the local area (with its N simulation runs); i.e., $\sigma_{i,h}\left(\underline{d}\right) = \sigma_{i,h}$ locally. The MSR estimators in (3.4) are unbiased for $\sigma_{i,h}$ if the responses satisfy the OLS assumptions locally; see Theil (1971, p. 114). Furthermore, since in the rest of this paper we only need point estimates for $\widehat{\sigma}_{i,h}$, we do not need to assume multivariate normality of the responses; that is, $\widehat{\sigma}_{i,h}$ in (3.4) are not necessarily chi-squarely distributed. Note that for $i = h$, (3.4) is found in any classic RSM or regression textbook, where "classic" means no multiple responses.

After the $\widehat{\underline{\beta}}_i$ are estimated, our problem in (3.1) is locally approximated by

$$\begin{array}{ll} \text{minimize} & \underline{b}_0^T \underline{d} \\ \text{subject to} & \underline{b}_j^T \underline{d} \geq c_j \ \text{ for } \ j = 1,\ ...,\ r-1 \\ & \underline{l} \leq \underline{d} \leq \underline{u} \end{array} \tag{3.5}$$

where $\underline{b}_i = \left(b_{i,1},\ ...,\ b_{i,k}\right)^T$ denotes a realization of the OLS estimator $\widehat{\underline{\beta}}_i$ excluding the intercept $\widehat{\beta}_{i,0}$ per response i, and $c_j = a_j - b_{j,0}$. In (3.5), we leave the constant term $b_{0,0}$ out of the objective function, since we will use only the gradient of the objective in the following section.

An important characteristic of simulation is the experimenter's control over the seed $\underline{\omega}$ that drives the simulation model. The use of common or antithetic seeds as variance reduction techniques is standard in simulation; see Law and Kelton (2000). In conventional RSM context, Donohue, Houck, and Myers (1993, 1995), and Hussey, Myers, and Houck (1987) considered the use of the following three seed assignment strategies as part of their experimental design: (i) the assignment of an independent seed to each input vector; (ii) the assignment of a common seed to input vectors; (iii) Schruben and Margolin's (1978) "assignment rule" that simultaneously uses common and antithetic seeds in an orthogonally blocked experimental design. Notice that for a fixed seed $\underline{\omega}$, the unknown responses F_i in (3.1) are deterministic functions of $\underline{d}$. In

the heuristic in §3.4, we will use the common seed approach on our path towards the optimum to reduce the random variations in the estimates for the estimators of the expected responses.

3.3 The new search direction and its properties

In this section we propose an estimated search direction for problem in (3.1) and present two properties: it is a descent direction, and it is scale independent.

We rewrite (3.5) by introducing $\mathbf{B} = \left(\underline{b}_1,\ ...,\ \underline{b}_{r-1}\right)^T$, and we add slack vectors $\underline{s}$, $\underline{r}$, and $\underline{v}$:

$$\begin{array}{ll} \text{minimize} & \underline{b}_0^T \underline{d} \\ \text{subject to} & \mathbf{B}\underline{d} - \underline{s} = \underline{c}, \\ & \underline{d} + \underline{r} = \underline{u}, \\ & \underline{d} - \underline{v} = \underline{l}, \\ & \underline{s} \in R_+^{r-1},\ \underline{r},\ \underline{v} \in R_+^k. \end{array} \tag{3.6}$$

Note that our purpose is not to solve the linear programming problem in (3.6); we use the local approximation in (3.6) only to derive a search direction.

Now, we give the proposed search direction $\underline{p}$ derived in both Appendices 1 and 2, using standard tools from interior point methods and nonlinear programming:

$$\underline{p} = -\left(\mathbf{B}^T \widehat{\mathbf{S}}^{-2} \mathbf{B} + \widehat{\mathbf{R}}^{-2} + \widehat{\mathbf{V}}^{-2}\right)^{-1} \underline{b}_0 \tag{3.7}$$

where $\widehat{\mathbf{S}}$, $\widehat{\mathbf{R}}$, and $\widehat{\mathbf{V}}$ are diagonal matrices with the components of the current estimated slack vectors $\underline{\widehat{s}}, \underline{\widehat{r}}, \underline{\widehat{v}} > \underline{0}$ on the diagonals. Notice that in (3.7), the inverse of the matrix within the parentheses scales and projects the estimated steepest descent $(-\underline{b}_0)$. First, we present some intuitive ideas about the derivations of (3.7), and then we consider some special cases of (3.6), which affect (3.7).

To give some geometrical insight about the derivation of (3.7) in Appendix 1, we start with conventional RSM (with only one response), and the search direction is given by the estimated steepest descent. Finding the normalized estimated steepest descent $\underline{p}$ can then be formulated as

$$\text{minimize } \underline{b}_0^T \underline{p} \quad \text{subject to } \|\underline{p}\|_2 \leq 1 \tag{3.8}$$

where $\|.\|_2$ denotes the two-norm. The problem in (3.8) corresponds to the minimization of a linear function over a ball constraint. However, the estimated steepest descent may be very inefficient in a constrained problem, as we illustrated with Figure 3.1 in

§3.1. Hence, we replace the nonnegativity constraints on the estimated slack vectors by an ellipsoid constraint that lies in the interior of the positive orthant, and that is centered at the current estimated slacks with a shape determined by the current values of the estimated slacks; see Barnes (1986) and also Appendix 1. Then, (3.7) is obtained through the minimization of a linear function over this ellipsoid constraint.

In Appendix 2, we consider an alternative derivation of (3.7). This derivation has two major components, namely scaling and projection of the estimated steepest descent. Scaling simply means multiplying the estimated slack vectors by diagonal matrices with the current estimated slack values on the diagonals. Notice that scaling maps the current estimated slack vectors to all-one vectors $(1, ..., 1)^T$ of appropriate dimensions. This ensures that (for example) whether distance is measured in miles or in kilometers, the property of being close to a "wall" remains unchanged. Next, projection is used to maintain the feasibility of equality constraints while dealing with the estimated steepest descent.

We now discuss two special cases of (3.6).

- Case 1: If there are no simple bounds (box constraints) on the decision variables (i.e., a formulation without $\underline{d} + \underline{r} = \underline{u}$ and $\underline{d} - \underline{v} = \underline{l}$), then (3.7) reduces to

$$\underline{p} = -\left(\mathbf{B}^T \widehat{\mathbf{S}}^{-2} \mathbf{B}\right)^{-1} \underline{b}_0.$$

  This can be seen as follows. No simple bounds on $\underline{d}$ implies $-\infty < \underline{d} < +\infty$. Then, using $1/\infty = 0$ reduces $\widehat{\mathbf{R}}^{-2}$ and $\widehat{\mathbf{V}}^{-2}$ in (3.7) to zero-matrices.

- Case 2: If there are only simple bounds on the decision variables (i.e., a formulation without $\mathbf{B}\underline{d} - \underline{s} = \underline{c}$), then (3.7) simplifies to

$$\underline{p} = -\left(\widehat{\mathbf{R}}^{-2} + \widehat{\mathbf{V}}^{-2}\right)^{-1} \underline{b}_0. \tag{3.9}$$

  This is useful in general, since although it is not explicitly stated, conventional RSM assumes simple bounds on the decision variables. The scaling matrix $\left(\widehat{\mathbf{R}}^{-2} + \widehat{\mathbf{V}}^{-2}\right)^{-1}$ in (3.9) is a diagonal matrix with expressions of the estimated current slacks on the diagonal. Then, the scaling matrix in (3.9) adapts the estimated steepest descent $(-\underline{b}_0)$ with respect to the shape of the feasible region through the estimated slack values. We expect that for problems with only box constraints, (3.9) performs significantly better than the estimated steepest descent $(-\underline{b}_0)$ whenever the individual diagonal entries in $\left(\widehat{\mathbf{R}}^{-2} + \widehat{\mathbf{V}}^{-2}\right)^{-1}$ differ.

We now consider two properties of the proposed search direction. First, since $\mathbf{B}^T\widehat{\mathbf{S}}^{-2}\mathbf{B}$, $\widehat{\mathbf{R}}^{-2}$, and $\widehat{\mathbf{V}}^{-2}$ are positive definite, the estimated search direction in (3.7) is indeed a descent direction; that is,

$$-\underline{b}_0^T \left(\mathbf{B}^T\widehat{\mathbf{S}}^{-2}\mathbf{B} + \widehat{\mathbf{R}}^{-2} + \widehat{\mathbf{V}}^{-2}\right)^{-1} \underline{b}_0 < 0.$$

Second, to avoid numerical complications and problems, it is common practice to make variables have similar magnitudes. This can be done by a general linear transformation as follows:

$$\mathbf{J}\underline{z} + \underline{e} = \underline{d} \tag{3.10}$$

where $\mathbf{J}$ is a $k \times k$ non-singular matrix and $\underline{e}$ is a $k \times 1$ vector. This transformation is one-to-one, so $\underline{z} = \mathbf{J}^{-1}(\underline{d} - \underline{e})$. Such a transformation is standard in ordinary RSM literature: transform the original variables into coded ones. In Appendix 3, we prove that the estimated search direction in (3.7) is invariant under the transformation (3.10). On the other hand, the estimated steepest descent, used in conventional RSM, is scale dependent; see Myers and Montgomery (2002, pp. 218-220).

3.4 An iterative heuristic for the first stage of RSM

In this section, we describe a heuristic that uses the search direction in (3.7) iteratively. Primarily, this heuristic is intended for simulation-based optimization problems in which each simulation run is time-consuming, and the simulation budget is tight. In particular, this heuristic aims at quickly reaching a neighborhood of the minimizer of stochastically constrained problem in (3.1). Moreover, after a few simplifications, the heuristic can be easily applied to deterministic problems. We will introduce these simplifications when we detail each step of the heuristic.

Before presenting the details of the heuristic, we explain briefly the general procedure. The F_i in (3.1) are locally approximated by first-order polynomials. Then, a search direction (given in (3.7)) and a maximum step size are estimated from the resulting model. In the initialization, the heuristic fixes a set of independent seed vectors to be used when estimating function values locally. This same set is used several times when estimating function values in different local areas. The heuristic changes this fixed set of independent seed vectors only once, when there is no progress observed in a newly estimated search direction. In this way, the heuristic tries to avoid premature stopping if this stopping is caused by the inherent randomness of the problem.

At each iteration of the heuristic, there are two statistical tests - one for "feasibility" and one for "improvement" in objective - to determine whether the candidate iterate will become the next iterate. Different from the general practice, in these tests the heuristic considers relative improvements (rather than absolute improvements) in the candidate iterate with respect to the current iterate. This approach is in line with interior point methods, where at each iteration one takes a step such that the candidate iterate's slacks are equal to some percentage of the current slacks. In addition, the use of ratios instead of absolute differences avoids scale dependency.

Furthermore, the heuristic does not assume multivariate normality for the F_i. Hence, we do not use the classical statistical tests. Instead, we use Monte Carlo sampling (as described in Law and Kelton (2000, pp. 90-91)), which does not require any assumption on the joint distribution for the F_i.

We now provide a more specific and detailed description of why and how we resort to Monte Carlo sampling as part of the heuristic: suppose our current iterate is $\underline{d}_m$, and it is a "feasible" point. First, we check whether the candidate point, $\underline{d}_m + \lambda \underline{p}$, is "feasible" through a statistical test for "feasibility". If $\underline{d}_m + \lambda \underline{p}$ is "feasible", then we check whether there is significant decrease in the objective at $\underline{d}_m + \lambda \underline{p}$ compared to the objective at $\underline{d}_m$. This comparison is done through a statistical test for "improvement" in objective. In the heuristic, we consider the ratios $S_j\left(\underline{d}_m + \lambda \underline{p}\right) / S_j\left(\underline{d}_m\right)$ of the random slacks for the statistical test for "feasibility"; this is in line with the classical interior point methods. For symmetry and ease of explanation, we also look at the ratios of the objectives (that is, relative improvements in the objectives formulated as $\left(F_0\left(\underline{d}_m\right) - F_0\left(\underline{d}_m + \lambda \underline{p}\right)\right) / \mid F_0\left(\underline{d}_m\right) \mid$) instead of absolute improvements, where $|.|$ denotes the absolute value:

$$E_{\underline{\omega}}\left[F_0\left(\underline{d},\ \underline{\omega}\right) \mid \underline{d}{=}\underline{d}_m + \lambda \underline{p}\right] < E_{\underline{\omega}}\left[F_0\left(\underline{d},\ \underline{\omega}\right) \mid \underline{d}{=}\underline{d}_m\right].$$

Obviously, $E_{\underline{\omega}}\left[F_0\left(\underline{d},\ \underline{\omega}\right) \mid \underline{d}{=}\underline{d}_m + \lambda \underline{p}\right]$ and $E_{\underline{\omega}}\left[F_0\left(\underline{d},\ \underline{\omega}\right) \mid \underline{d}{=}\underline{d}_m\right]$ are unknown. Since our heuristic is concerned with cases in which simulation is very time-consuming (so we cannot make a large number of simulation runs), we consider $\widehat{f}_0\left(\underline{d}_m + \lambda \underline{p}\right)$ and $\widehat{f}_0\left(\underline{d}_m\right)$, which are the corresponding simulation runs obtained for $F_0\left(\underline{d}_m + \lambda \underline{p},\ \underline{\omega}\right)$ and $F_0\left(\underline{d}_m,\ \underline{\omega}\right)$, as the point estimates for the means of $F_0\left(\underline{d}_m + \lambda \underline{p},\ \underline{\omega}\right)$ and $F_0\left(\underline{d}_m,\ \underline{\omega}\right)$. Notice that we cannot use any of the classical statistical tests, since there is no multivariate normality assumption for the F_i. Since Monte Carlo requires an amount of computer time that is negligible in an expensive simulation study and it is easy to use and explain to the readers, we use Monte Carlo. Note that the use of Monte Carlo requires the knowledge of the distributions and the distributional parameters. We pro-

ceed as follows: we sample a large number, say $K = 1000$, of the random objectives $F_0\left(\underline{d}_m + \lambda \underline{p},\ \underline{\omega}\right)$ and $F_0\left(\underline{d}_m,\ \underline{\omega}\right)$ from their distributions, say normal (but it could be any other distribution; that is, if the practitioner has reason to believe a particular distribution may be better suited for the problem on hand, then sampling should be done from that particular distribution), with means $\widehat{f}_0\left(\underline{d}_m + \lambda \underline{p}\right)$ and $\widehat{f}_0\left(\underline{d}_m\right)$ and the variance $\widehat{\sigma}_{0,0}$, which was estimated when the search direction was last revised. A similar procedure is used for generating a Monte Carlo sample of the slacks. Further details of the statistical tests for "feasibility" and "improvement" in objective, based on these Monte Carlo samples, are explained at the end of this section.

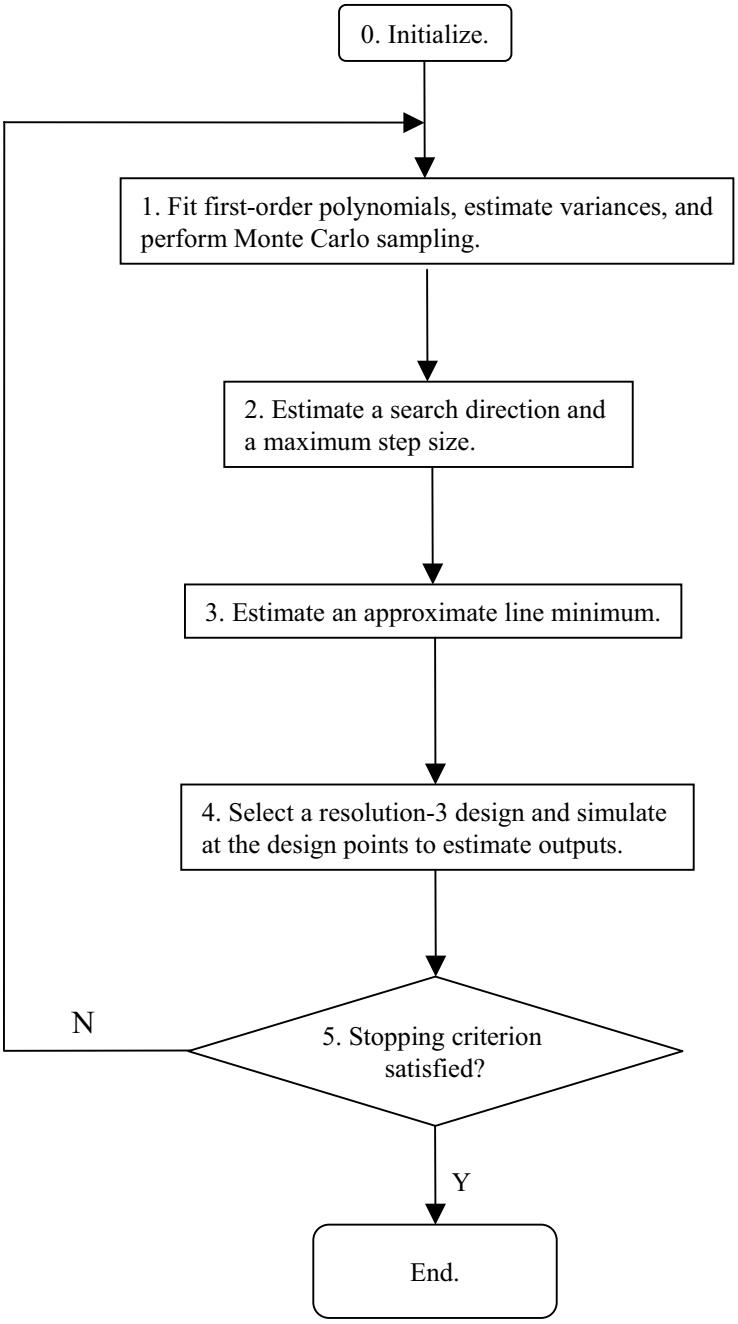
In Figure 3.2, we summarize the outer loop of the heuristic. Our heuristic neglects the statistical dependencies among the F_i. Now, we detail each step of the heuristic as follows.

- **Step 0, Initialize:** Input a user-specified fixed number of simulation runs per inner loop and a maximum total number of simulation runs for the outer loop. The maximum total number for the outer loop is determined by the time and budget constraints of an expensive simulation study. For the outer loop, set the number of simulation runs already executed to zero.

  Input also a fixed, user-specified size of the local experimental area, where the responses F_i are locally approximated by first-order polynomials. This local experimental area should lie within the global area determined by the box constraints in (3.1). The size of this local experimental area is clearly scale dependent, and there are no general guidelines to determine an appropriate size that would work in all applications; see the standard RSM textbooks by Myers and Montgomery (2002), and Khuri and Cornell (1996). Therefore, to determine an appropriate size, the users need to have insights into their applications. Notice that the arbitrary choice for the initial size of the region of interest is also a general issue in nonlinear programming: in trust-region methods, the initial size of the local experimental area, namely the trust-region, is also user-specified; see Conn, Gould, and Toint (2000).

  Furthermore, we assume that the users can determine at least one design point inside the "feasible" region. The rest of the initial design points may potentially be "infeasible", but clearly the simulation program should not crash in those points. For an existing system, a "feasible" point is provided by the current design point at which the system is operating. In general, however, there are again no guidelines that ensure that a selected design point is inside the "feasible" re-

Figure 3.2: Overview of the iterative heuristic



gion prior to simulating at that point; therefore, the users are again asked to use their insights into their specific problems. As a further remark, at least one of the "feasible" points has to be in the interior of the "feasible" region far from the boundary, since (3.7) will creep along the boundary or fail when this point is close to the boundary or on the boundary, respectively.

For the first stage of RSM, we use a resolution-3 design, since such a design type gives unbiased estimators for $\underline{\beta}_i$ in (3.2) with a small number of simulation runs, provided that first-order polynomials are adequate approximations; see Kleijnen (1998). Simulate at the design points $\underline{d}_m$ to estimate their objectives $\widehat{f}_0(\underline{d}_m)$ and their slacks $\widehat{s}_j(\underline{d}_m) = a_j - \widehat{f}_j(\underline{d}_m)$, where the a_j are the right-hand-side values in (3.1). Notice that when we use Monte Carlo sampling later, these $\widehat{f}_0(\underline{d}_m)$ and $\widehat{s}_j(\underline{d}_m)$, which are obtained through a single simulation run at $\underline{d}_m$, will be used as the point estimates for the means of the random objective $F_0(\underline{d}_m, \underline{\omega})$ and the random slacks $S_j(\underline{d}_m, \underline{\omega})$.

For later use, fix the set $\{\underline{\omega}_1, ..., \underline{\omega}_m, ..., \underline{\omega}_N\}$ of independent seed vectors for each of the N design points. The initial iterate, say $\underline{d}_m$, is the "feasible" design point (among the N design points) that has the minimum objective $\widehat{f}_0(\underline{d}_m)$ estimated through simulation; i.e., $\underline{d}_m$ is not found by minimizing the local linear model in (3.6). Fix the seed vector $\underline{\omega}_m$ corresponding to $\underline{d}_m$ as the common seed. Increase the number of already executed simulation runs for the outer loop by N, which is the number of runs used for initialization.

As we already mentioned in §3.2 when we introduced (3.1), the inherent randomness $\underline{\omega}$ vanishes for deterministic problems. Hence, the generation of independent seeds and the use of common seed are to be skipped when the heuristic is applied to such problems.

- **Step 1, Fit first-order polynomials, estimate variances, and perform Monte Carlo sampling:** Approximate (3.1) by local first-order polynomials within the local experimental area using (3.3), and obtain point estimates $\widehat{\sigma}_{i,h}$ for the variances through (3.4). As mentioned already in §3.2 below (3.4), the MSR estimators in (3.4) are unbiased for $\sigma_{i,h}$. Besides, the locally constant covariance assumption, mentioned in §3.2 below (3.4), may not hold globally. Thus, we will use only the most recent estimates $\widehat{\sigma}_{i,h}$ in the heuristic.

  After estimating the variances, we perform Monte Carlo sampling (as a reminder, the heuristic ignores the correlations among the F_i): sample a large number K of observations on $F_0(\underline{d}_m, \underline{\omega})$ from the normal distribution with mean $\widehat{f}_0(\underline{d}_m)$

and variance $\widehat{\sigma}_{0,0}$, and on $S_j\left(\underline{d}_m,\ \underline{\omega}\right)$ from the normal distributions with means $\widehat{s}_j\left(\underline{d}_m\right)$ and variances $\widehat{\sigma}_{j,j}$. Notice that our normality choice is only for clarity in describing the heuristic, and in line with conventional RSM literature. Later, these samples from $F_0\left(\underline{d}_m,\ \underline{\omega}\right)$ and $S_j\left(\underline{d}_m,\ \underline{\omega}\right)$ will be used in the statistical tests for "feasibility" and "improvement" in objective.

For deterministic problems, variance estimation and Monte Carlo sampling are not needed since the exact values for $F_0\left(\underline{d}_m + \lambda\underline{p}\right)$, $F_0\left(\underline{d}_m\right)$, $S_j\left(\underline{d}_m + \lambda\underline{p}\right)$, and $S_j\left(\underline{d}_m\right)$ are obtained through simulation.

- **Step 2, Estimate a search direction and a maximum step size:** Determine a search direction using (3.7), where the diagonal entries of $\widehat{\mathbf{S}}^{-2}$ are estimated through a single simulation run at $\underline{d}_m$; i.e., $\widehat{s}_j^{-2}\left(\underline{d}_m\right) = \left(a_j - \widehat{f}_j\left(\underline{d}_m\right)\right)^{-2}$. To determine a maximum step size into this direction, we initially assume that the approximation in (3.5) holds globally. Then we find a maximum step size $\lambda_{\max}$ as follows.

$$\begin{array}{ll} \text{maximize} & \lambda \\ \text{subject to} & \underline{b}_j^T\left[\underline{d}_m + \lambda\underline{p}\right] \geq c_j \ \text{ for } \ j = 1, ..., r-1 \\ & \underline{l} \leq \underline{d}_m + \lambda\underline{p} \leq \underline{u},\ \ \lambda \geq 0 \end{array} \tag{3.11}$$

  Notice that it is not necessary to solve (3.11) as a linear program. The maximum step size $\lambda_{\max}$ can be obtained explicitly through $\lambda_{\max} = \max\{0,\ \min\{\lambda_1,\ \lambda_2,\ \lambda_3\}\}$ where

$$\begin{aligned} \lambda_1 &= \min\left\{(c_h - \underline{b}_h^T\underline{d}_m)/\underline{b}_h^T\underline{p}\text{:}\ h \in \{1,\ ...,\ r-1\},\ \underline{b}_h^T\underline{p}<0\right\} \\ \lambda_2 &= \min\left\{(u_n - d_n)/p_n\text{:}\ n \in \{1,\ ...,\ k\},\ p_n > 0\right\} \\ \lambda_3 &= \min\left\{(l_n - d_n)/p_n\text{:}\ n \in \{1,\ ...,\ k\},\ p_n < 0\right\}. \end{aligned}$$

  To increase the probability of staying within the interior of the "feasible" region, we take only 80% of $\lambda_{\max}$ as our maximum step size λ.

- **Step 3, Estimate an approximate line minimum:** In the initialization, set the number of simulation runs already executed per inner loop to zero. Simulate at $\underline{d}_m + \lambda\underline{p}$ to estimate $\widehat{f}_0\left(\underline{d}_m + \lambda\underline{p}\right)$ and $\widehat{s}_j\left(\underline{d}_m + \lambda\underline{p}\right)$ using the common seed vector $\underline{\omega}_m$. Sample K observations on $F_0\left(\underline{d}_m + \lambda\underline{p},\ \underline{\omega}\right)$ from the normal distribution with mean $\widehat{f}_0\left(\underline{d}_m + \lambda\underline{p}\right)$ and variance $\widehat{\sigma}_{0,0}$, and on $S_j\left(\underline{d}_m + \lambda\underline{p},\ \underline{\omega}\right)$ from the normal distributions with means $\widehat{s}_j\left(\underline{d}_m + \lambda\underline{p}\right)$ and variances $\widehat{\sigma}_{j,j}$.
  At this point, the heuristic compares the current iterate $\underline{d}_m$ with the candidate

iterate $\underline{d}_m + \lambda \underline{p}$ to determine the "better" point, where "better" means "feasible" with a lower objective. This comparison is done through two statistical tests, namely one for "feasibility" and one for "improvement" in objective; these tests will be described in detail at the end of this section. As mentioned in the beginning of this section, the heuristic tests the ratios of the random slacks, which is in the spirit of interior point methods. By symmetry, the heuristic also tests the ratios of the objectives. Notice that classical statistical tests (for example, a paired-t test for the objectives) are not used, since the heuristic does not assume (multivariate) normality. Besides, normality implies Cauchy type distributions for ratios, and these distributions have no finite means. Then, the Monte Carlo samples are used for the two statistical tests that can be summarized as follows: determine the medians (not means) of the given Monte Carlo samples since ratios of two random variables may not necessarily have finite means, and test whether the lower bounds on these medians are significant (i.e., exceed prespecified values). The details are explained at the end of this section.
Determine the "better" of $\underline{d}_m + \lambda \underline{p}$ and $\underline{d}_m$. Denote the "better" by $\underline{d}_m$ and the other by $\underline{d}_m + \lambda \underline{p}$. Set the objective to $\widehat{f}_0(\underline{d}_m)$. Increase the number of already executed simulation runs for both the outer and the inner loops by one.
Now, analogous to binary search, we have the interval $\left[\underline{d}_m, \underline{d}_m + \lambda \underline{p}\right]$ to be systematically halved for a prespecified number of simulation runs in the same estimated search direction, to estimate an approximate line minimum. Repeat the procedure each time with a new interval $\left[\underline{d}_m, \underline{d}_m + \lambda \underline{p}\right]$, until the fixed number of simulation runs per inner loop is reached. Then, set the current estimated approximate line minimum's $(\underline{d}_m)$ slacks to $\widehat{s}_j(\underline{d}_m)$. At the end of this step, it is possible that the heuristic fails to find a better point than the old $\underline{d}_m$. Then, the heuristic does not leave that point.

- **Step 4, Select a resolution-3 design and simulate at the design points:** The current design point $\underline{d}_m$ and the other design points are the vertices of an N-dimensional hypercube with the side length determined by the fixed size of the local experimental area. Using the fixed set $\{\underline{\omega}_1, \ ..., \ \underline{\omega}_m, \ ..., \ \underline{\omega}_N\}$ of independent seeds (except $\underline{\omega}_m$, which was used at $\underline{d}_m$), simulate at the new design points other than $\underline{d}_m$, since we already simulated at $\underline{d}_m$. If in the line search in Step 3, the heuristic failed to find a better point than the old $\underline{d}_m$ (we already used $\underline{d}_m$ in a previous resolution-3 design), then generate a new set $\{\underline{\omega}_1, \ ..., \ \underline{\omega}_m, \ ..., \ \underline{\omega}_N\}$ of independent seeds, and select the seed $\underline{\omega}_m$ of $\underline{d}_m$ as the new common seed.

Increase the number of already executed simulation runs for the outer loop by the number of new runs.

- **Step 5, Stopping criterion satisfied:**The heuristic stops when either the number of executed simulation runs for the outer loop exceeds the maximum, or the heuristic uses the same design point more than twice in a resolution-3 design. The latter happens as follows: suppose that our current iterate is $\underline{d}_m$, and we selected a resolution-3 design for which $\underline{d}_m$ was one of the vertices. If the heuristic could not find a better point than $\underline{d}_m$ at the line search (in Step 3), it did not leave the current iterate $\underline{d}_m$. Next, the heuristic generates a new set of independent seed vectors $\{\underline{\omega}_1, ..., \underline{\omega}_N\}$ (in Step 4 or in Step 0), and use the previous resolution-3 design to obtain new local linear approximations in (3.2), and to derive a new search direction (3.7). If the heuristic again fails to find a better point than $\underline{d}_m$ at the line search, then it stops since $\underline{d}_m$ was already used twice in a resolution-3 design.
  Note that in case the current estimated iterate is far from a neighborhood of the minimizer, premature stopping may be caused by the inherent randomness of the problem. Therefore, we estimate a new search direction by generating a new set of independent seeds in Step 4. If, however, the current estimated iterate is in a neighborhood of the minimizer, the fixed size of the local experimental area becomes simply too big to estimate a good search direction. If the stopping criterion is not satisfied yet, then return to Step 1.
  For deterministic problems, the only stopping rule is the maximum number of simulation runs for the outer loop.

This finishes our discussion of Figure 3.2. Clearly, the stopping rules used in the heuristic are rather arbitrary but also simple to use. The possibility of developing a more formal stopping rule is mentioned in §3.6, but we consider it outside the scope of this paper. Besides, this heuristic is designed for a small number of simulation runs. In such a case, we cannot claim theoretical convergence to the true optimum.

Tests for "feasibility" and "improvement" in objective: The deterministic counterpart of testing for "feasibility" and "improvement" in objective will be explained in the last paragraph of this section. Therefore, the text until the last paragraph of this section should be skipped for deterministic problems.

As mentioned before, to decide on the point that has a "better" estimated objective, we use the random ratio $\left(F_0\left(\underline{d}_m\right) - F_0\left(\underline{d}_m + \lambda \underline{p}\right)\right) / \mid F_0\left(\underline{d}_m\right) \mid$, where $F_0\left(\underline{d}_m\right)$

stands for the minimum objective so far and $F_0\left(\underline{d}_m + \lambda \underline{p}\right)$ for the candidate's objective. To simplify the notation, we do not show the dependence on $\underline{\omega}$.

We use the symbol $Q_i = \left(F_{0,i}\left(\underline{d}_m\right) - F_{0,i}\left(\underline{d}_m + \lambda \underline{p}\right)\right) / \mid F_{0,i}\left(\underline{d}_m\right) \mid$, where $i \in \{1, ..., K\}$ is the ith sampled value (from Monte Carlo). Let $Q_{(1)} < ... < Q_{(K)}$ be the corresponding order statistics. Then, a point estimator for the median $Q^{0.5}$ (see Law and Kelton (2000, p. 517)) is

$$\widehat{Q}^{0.5} = \begin{cases} Q_{(0.5K)} & \text{if } 0.5K \text{ is an integer,} \\ Q_{(\lfloor 0.5K+1 \rfloor)} & \text{otherwise.} \end{cases} \tag{3.12}$$

where $\lfloor . \rfloor$ denotes the largest integer less than or equal to $0.5K + 1$.

Our null hypothesis is $H_0^{(1)} : Q^{0.5} \leq \delta$, where δ is a user-specified positive constant indicating the smallest desired improvement in the estimated objective; in the numerical examples, we have $\delta = 2.5\%$. In other words, $H_0^{(1)}$ is pessimistic: we reject $H_0^{(1)}$ only if the estimated median $\widehat{q}^{0.5}$ is significantly greater than δ. This implies that a point will be considered to be a significantly improving point only if its estimated objective is significantly smaller than the minimum objective so far. Then, the index y of the lower confidence limit $Q_{(y)}$ is given by

$$y = \lceil Kx - z^{\alpha}\sqrt{Kx\left(1-x\right)} \rceil \tag{3.13}$$

where α is the significance level, $x = 0.5$, z^{α} is the $1 - \alpha$ quantile of the standard normal distribution (see Kleijnen (1987, pp. 23-25)), and $\lceil . \rceil$ denotes the smallest integer greater than or equal to $Kx - z^{\alpha}\sqrt{Kx\left(1-x\right)}$.

We also use the random ratios $S_j\left(\underline{d}_m + \lambda \underline{p}\right) / S_j\left(\underline{d}_m\right)$ to check the "feasibility", where $S_j\left(\underline{d}_m\right)$ stands for the slack j at the current iterate $\underline{d}_m$ and $S_j\left(\underline{d}_m + \lambda \underline{p}\right)$ for the candidate's slack j. Ideally, we would take a step such that the candidate iterate's slacks would be $S_j\left(\underline{d}_m + \lambda \underline{p}\right) = \gamma S_j\left(\underline{d}_m\right)$, where γ is a user-specified constant with $\gamma < 1$. Then, the heuristic considers those points $\underline{d}_m + \lambda \underline{p}$ with the slacks $S_j\left(\underline{d}_m + \lambda \underline{p}\right) < \gamma S_j\left(\underline{d}_m\right)$ as "infeasible", although these points may indeed be "feasible" for (3.1). This way, the heuristic avoids prematurely approaching the boundary. This is important since if $\underline{d}_m$ is on the boundary or close to the boundary at an early iteration, then the search direction (3.7) will fail or creep along the boundary. However, as the heuristic moves towards the optimum, the current iterate $\underline{d}_m$ will eventually be close to the boundary, and so will be the candidate.

We follow a procedure analogous to the one described for the objective, as follows. For constraint j, we denote the random ratios by $M_{j,i} = S_{j,i}\left(\underline{d}_m + \lambda \underline{p}\right) / S_{j,i}\left(\underline{d}_m\right)$ with $i = 1, ..., K$. Now, our null hypothesis becomes $H_0^{(2)} : \min\left\{M_1^{0.5}, ..., M_{r-1}^{0.5}\right\} \leq \gamma$,

where γ is the same user-specified constant as the one in the previous paragraph, and $M_j^{0.5}$ is the median for the jth constraint. Hence, we again have a pessimistic $H_0^{(2)}$, which means that the "feasible" region considered by the heuristic is tighter than the actual "feasible" region. In our experiments, we use $\gamma = 0.2$, which is in line with our choice of 80% for determining a maximum step size (see Step 2). Let $M^{0.5} = \min\left\{M_1^{0.5},\ ...,\ M_{r-1}^{0.5}\right\}$. We then find a point estimator $\widehat{M}^{0.5}$ and an index y for the lower confidence limit through (3.12) and (3.13).

Notice that α in (3.13) for $H_0^{(2)}$ is not necessarily the same as α for $H_0^{(1)}$. Moreover, because of Bonferroni's inequality we choose the significance level α_j for each constraint j such that $\alpha_1 + ... + \alpha_{r-1} = \alpha$. Furthermore, the choice for α of $H_0^{(2)}$ is not completely arbitrary; when α is small, rejection of $H_0^{(2)}$ is more likely. In this way, we move towards the optimum through the interior of the "feasible" region, which we empirically found to improve the performance of the heuristic.

To decide on the point that has a lower objective in deterministic problems, we can check whether $\left(\widehat{f}_0\left(\underline{d}_m\right) - \widehat{f}_0\left(\underline{d}_m + \lambda\underline{p}\right)\right) / \left(\mid \widehat{f}_0\left(\underline{d}_m\right) \mid +1\right) > \delta$, where δ stands for the user-specified constant indicating the smallest desired improvement in the objective, and $\widehat{f}_0\left(\underline{d}_m\right)$ and $\widehat{f}_0\left(\underline{d}_m + \lambda\underline{p}\right)$ stand for the exact values obtained through simulation of $F_0\left(\underline{d}_m\right)$ and $F_0\left(\underline{d}_m + \lambda\underline{p}\right)$. Notice that the ratio $\left(\widehat{f}_0\left(\underline{d}_m\right) - \widehat{f}_0\left(\underline{d}_m + \lambda\underline{p}\right)\right) / \left(\mid \widehat{f}_0\left(\underline{d}_m\right) \mid +1\right)$ is defined even when $\widehat{f}_0\left(\underline{d}_m\right) = 0$. For feasibility, we can check whether

$$\left\{\min_{j=1,...,r-1} \frac{\widehat{s}_j\left(\underline{d}_m + \lambda\underline{p}\right)}{\widehat{s}_j\left(\underline{d}_m\right)}\right\} > \gamma$$

where γ is the same user-specified constant as in the stochastic case, and $\widehat{s}_j\left(\underline{d}_m\right)$ and $\widehat{s}_j\left(\underline{d}_m + \lambda\underline{p}\right)$ are the exact values obtained through simulation of $S_j\left(\underline{d}_m\right)$ and $S_j\left(\underline{d}_m + \lambda\underline{p}\right)$. Analogous to the stochastic case, the feasible point with the lower objective becomes the current iterate $\underline{d}_m$.

3.5 Numerical examples

Now, we perform Monte Carlo experiments to estimate the performance of the heuristic and the proposed search direction. Notice that Monte Carlo experiments enable us to control the intrinsic noise of the simulation and they are computationally very efficient compared to an expensive simulation study that may actually take weeks.

We consider the following example problem:

$$\begin{array}{ll} \text{minimize} & E[5(d_1-1)^2+(d_2-5)^2+4d_1d_2+\epsilon_0] \\ \text{subject to} & E[(d_1-3)^2+d_2^2+d_1d_2+\epsilon_1]\le 4 \\ & E[d_1^2+3\,(d_2+1.061)^2+\epsilon_2]\le 9 \\ & 0\le d_1\le 3,\ -2\le d_2\le 1 \end{array} \tag{3.14}$$

where ϵ_0, ϵ_1, and ϵ_2 are assumed to be multivariate normally distributed with mean vector $\underline{0}$, variances $\sigma_{0,0}=1$ ($\sigma_0=1$), $\sigma_{1,1}=0.0225$ ($\sigma_1=0.15$), $\sigma_{2,2}=0.16$ ($\sigma_2=0.4$), and correlations $\rho_{0,1}=0.6$, $\rho_{0,2}=0.3$, $\rho_{1,2}=-0.1$. It is easy to see that the unconstrained minimum occurs at $(d_1,\ d_2)=(-5,\ 15)$, whereas the constrained minimum occurs at approximately $(1.24,\ 0.52)$ with a mean objective value of 22.96 approximately.

In our Monte Carlo experiment, we arbitrarily choose the following local experimental area for a 2^2 design, expressed in the original variables: $2.4\le d_1\le 2.7$ and $-1.1\le d_2\le -0.8$. So, Δd_1 and Δd_2 are arbitrarily chosen to be 10% of the whole ranges for d_1 ($0\le d_1\le 3,\ \Delta d_1=0.3$) and d_2 ($-2\le d_2\le 1,\ \Delta d_2=0.3$). Notice that we do not shrink the experimental area while proceeding towards the optimum (although shrinking may improve the performance of the heuristic).

The user-supplied maximum number of simulation runs for the outer and the inner loops are 20 and 3, respectively. For the Monte Carlo sampling, the sample size $K=1000$ is selected. The significance level for $H_0^{(1)}$ is selected as $\alpha^{(1)}=20\%$. As mentioned in §3.4, relatively small values of $\alpha^{(2)}$ increase the probability of staying inside the "feasible" region, which enhances the performance of our heuristic; so we choose $\alpha^{(2)}=1\%$. Moreover, we choose $\delta=2.5\%$ (the smallest desired improvement in the estimated objective).

To test the performance of the heuristic, we make 100 macro-replicates since estimated medians and various quantiles are then found to stabilize. Because of the simplicity of our problem in (3.14), 100 macro-replicates take less than five minutes. Notice that at the end of each macro-replicate, we have an estimate $\left(\widehat{d}_1,\ \widehat{d}_2\right)$ for the solution of the first stage of RSM. Then we sort the macro-replicates with respect to the deviations of the corresponding objectives at $\left(\widehat{d}_1,\ \widehat{d}_2\right)$ from the true minimum 22.96. Hence, the estimated solutions $\left(\widehat{d}_1,\ \widehat{d}_2\right)$ in Figures 3.3 through 3.5 correspond to the 10th, 50th, and 90th "best" (10th, 50th, and 90th quantiles) of the estimated solutions, respectively. Furthermore, in these figures the two level curves of the objective function correspond to the true optimum value (22.96) plus three and six times the standard deviation $\sigma_0=1$, namely $E_{\underline{\omega}}\left[F_0\left(\underline{d},\ \underline{\omega}\right)\right]=25.96$ and $E_{\underline{\omega}}\left[F_0\left(\underline{d},\ \underline{\omega}\right)\right]=28.96$.

In Figures 3.3 and 3.4, the heuristic indeed reaches the desired neighborhood

Figure 3.3: The "best" (10th quantile) of 100 estimated solutions for (3.14): (7*) is estimated solution

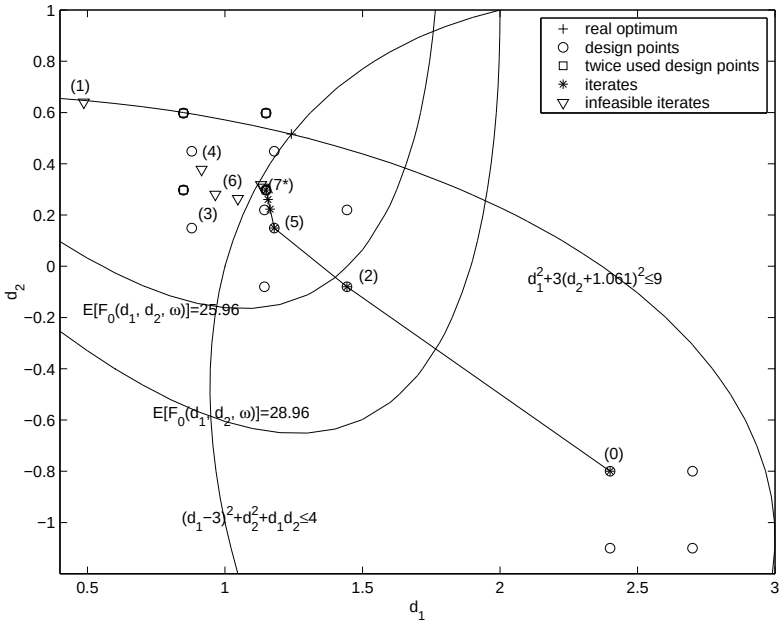


Figure 3.4: The "average" (50th quantile) of 100 estimated solutions for (3.14): (10*) is estimated solution

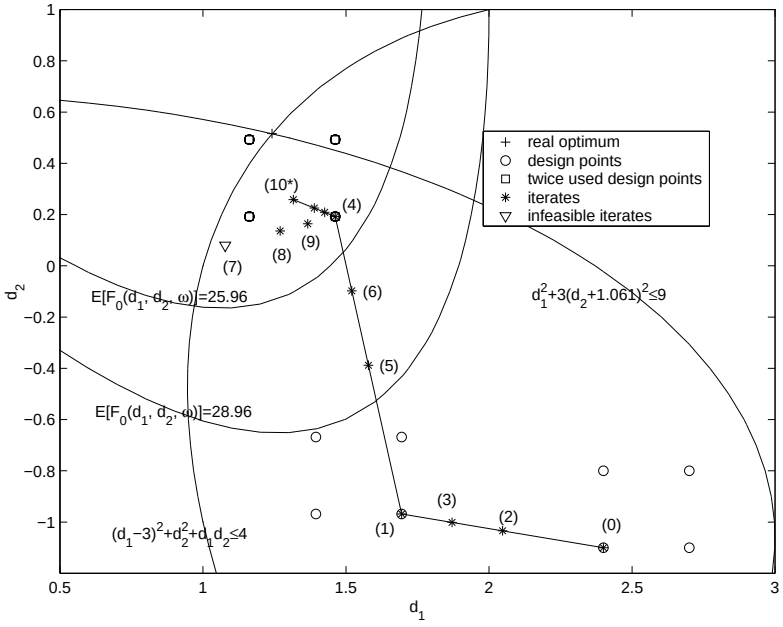
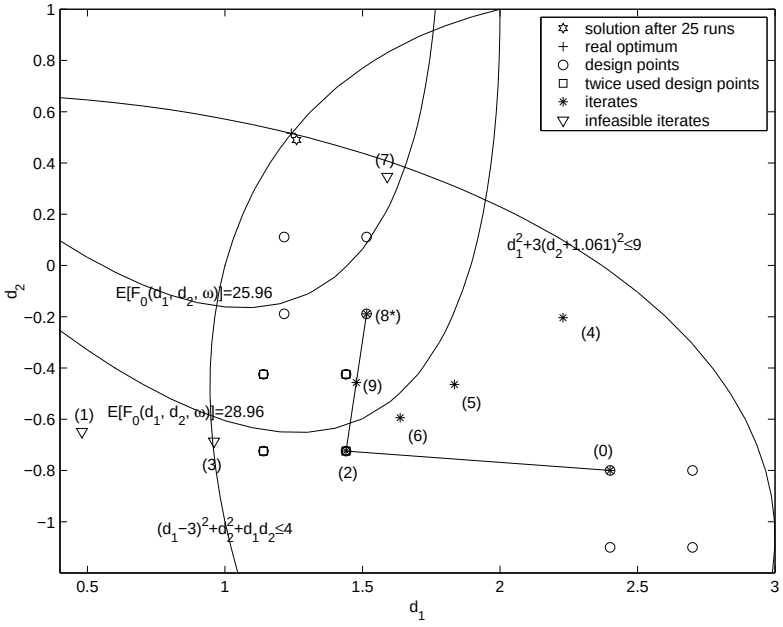


Figure 3.5: The "worst" (90th quantile) of 100 estimated solutions for (3.14): (8*) is estimated solution



of the true optimum: $\left(\widehat{d}_1, \widehat{d}_2\right) = (1.16, 0.22)$ with 23.99 and $\left(\widehat{d}_1, \widehat{d}_2\right) = (1.46, 0.19)$ with 25.30, respectively. In Figure 3.5, the initial set of independent seeds does not enable us to estimate a good search direction, so that most of the simulation runs are wasted until the 8th iterate is reached. The heuristic stops at $\left(\widehat{d}_1, \widehat{d}_2\right) = (1.52, -0.18)$ with the corresponding objective 27.09, since the number of simulations runs exceeds 20 (the maximum for the outer loop). If the budget is flexible, this problem can be overcome by simply increasing the maximum for the outer loop. (For example, in Figure 3.5, by increasing the maximum total number of simulation runs for the outer loop from 20 to 25, the estimated solution and the corresponding objective become $\left(\widehat{d}_1, \widehat{d}_2\right) = (1.26, 0.49)$ and 23.12, respectively, which is an improvement of 15% in the objective reached after 20 runs). Detailed numerical results for Figure 3.4 (the macro-replicate giving the median result) are presented in Table 3.1.

We experimented with different starting points and obtained very similar results. For example, for $1.7 \leq d_1 \leq 2.0$ and $-1.5 \leq d_2 \leq -1.2$, see Figures 3.6 through 3.8.

We summarize the results obtained from all 100 macro-replicates in Table 3.2. The heuristic tends to end at a "feasible" point since we have only positive quantiles in Table 3.2. This is clearly due to our conservative (small) significance level $\alpha^{(2)} = 1\%$

with a pessimistic null hypothesis $H_0^{(2)}$.

Our conclusion is that the heuristic reaches the desired neighborhood of the true optimum in a relatively small number of simulation runs. Once the heuristic reaches this neighborhood, it stops at a "feasible" point.

Table 3.1: Numerical results for Figure 3.4

iteration	est. iterate $\left(\widehat{d}_1,\ \widehat{d}_2\right)$	est. search direction $\underline{p}^T$	est. step size λ	$H_0^{(1)}$	$H_0^{(2)}$
0	(2.40, −1.10)	(−0.98, 0.18)	0.72		
1	(1.69, −0.97)	(−0.19, 0.98)	1.18	reject	reject
2	(2.05, −1.03)			reject	reject
3	(1.87, −1.00)			reject	reject
4	(1.46, 0.19)	(−0.91, 0.41), (−0.96, −0.28)	0.16, 0.40	reject	reject
5	(1.58, −0.39)			reject	reject
6	(1.52, −0.09)			reject	reject
7	(1.32, 0.26)			reject	reject
8	(1.39, 0.23)			reject	reject
9	(1.43, 0.21)			reject	reject
10	(1.08, 0.08)			reject	reject
11	(1.27, 0.14)			reject	reject
12	(1.37, 0.16)			reject	reject

Figure 3.6: The "best" (10th quantile) of 100 estimated solutions for (3.14): (11*) is estimated solution

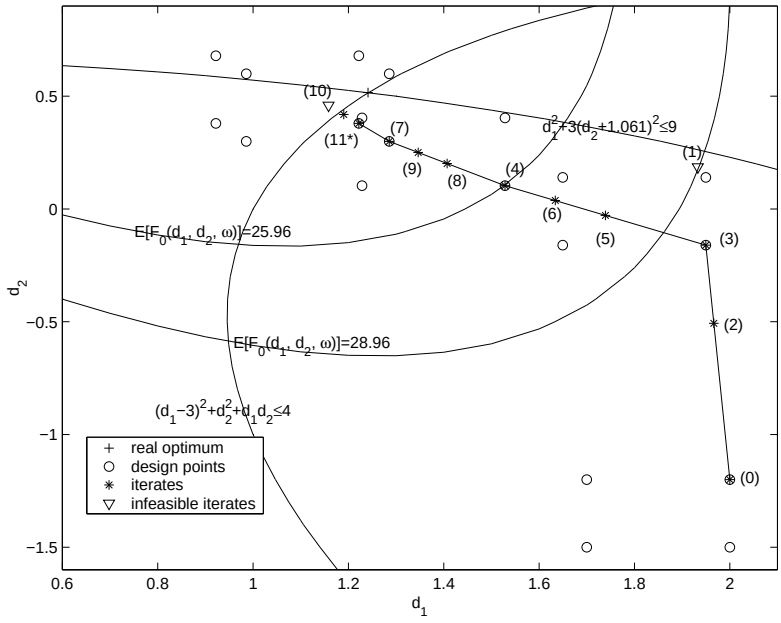


Figure 3.7: The "average" (50th quantile) of 100 estimated solutions for (3.14): (8*) is estimated solution

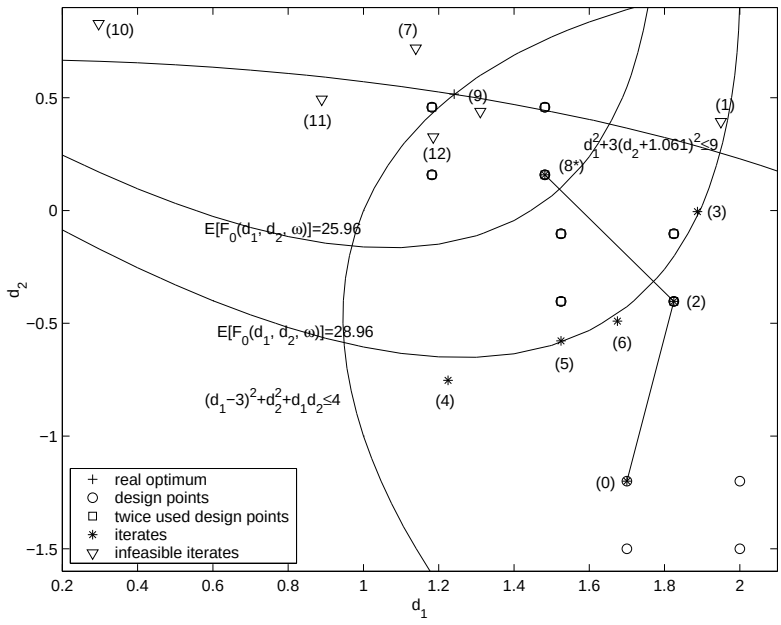


Figure 3.8: The "worst" (90th quantile) of 100 estimated solutions for (3.14): (8*) is estimated solution

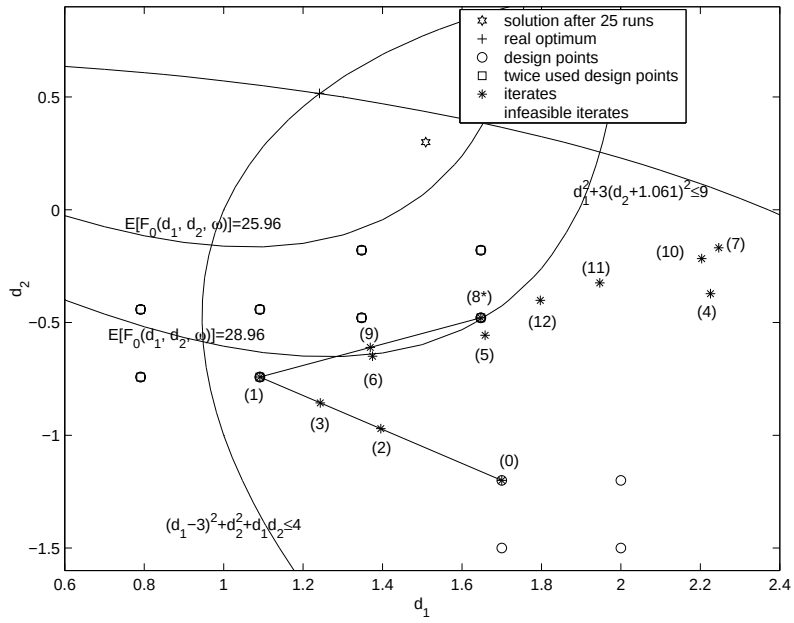


Table 3.2: Variability of the estimated objectives and slacks over 100 macro-replicates for the problem (3.14)

	10th quantile	25th quantile	50th quantile	75th quantile	90th quantile
$\frac{(E[F_0(\widehat{d}_1,\, \widehat{d}_2)]-22.96)}{22.96}$	0.0448	0.0555	0.1019	0.1858	0.1798
$\frac{(4-E[F_1(\widehat{d}_1,\, \widehat{d}_2)])}{4}$	0.0264	0.1185	0.2529	0.4348	0.6147
$\frac{(9-E[F_2(\widehat{d}_1,\, \widehat{d}_2)])}{9}$	0.0315	0.1453	0.2970	0.4921	0.5009

3.6 Conclusions and future research

In this paper we focus on RSM for problems with a stochastic objective function and stochastic constraints, as well as deterministic box constraints. At the first stage of RSM, the unknown functions are approximated locally by first-order polynomials. Then, a search direction is estimated, and along this path a number of steps are taken. The estimated search direction is a generalization of the steepest descent used in conventional RSM (with a stochastic objective function and deterministic box constraints). To achieve this generalization, we use standard ideas such as affine scaling and projection from nonlinear programming. We prove two properties of our search direction: it is indeed a descent direction, and it is invariant with respect to general linear transformations. The ordinary steepest descent, however, is scale dependent.

Next, we provide a heuristic procedure for quickly reaching the desired neighborhood of the true optimum of expensive simulation-based optimization problems. The heuristic proceeds towards this neighborhood through the interior of the feasible region. In this way, it avoids creeping along the boundary and ensures that simulation programs do not crash or become invalid. The heuristic especially aims at dealing with stochastic optimization problems with stochastic constraints. Yet, after a few simplifications, it can also be applied to deterministic problems. The empirical results of our numerical example are encouraging; that is, in general the heuristic reaches a point that is "sufficiently" close to the true optimum in a few simulation runs.

Notice that RSM treats simulation as a black box; that is, gradient estimation through simulation is not applicable. The main contribution of this paper is the generalization of the classical RSM to problems with stochastic constraints. The numerical example is only meant to illustrate the applicability of a heuristic that is based on the novel scale independent search direction. More experiments and examples may bring more insight into and better understanding of the capabilities and strengths of our approach. Furthermore, we might use certain enhancements such as better line search methods to determine the step size, or dynamically adapting the experimental area as we move towards the solution point, using ideas from trust-region methods.

In addition, it should be clear that this heuristic covers only the first stage of RSM; that is, it reaches a neighborhood of the true optimum in a few simulation runs. Afterwards the second stage, which is fine tuning and getting even closer to the true optimum with a well defined stopping rule, should be carried out. Analogous to the unconstrained deterministic optimization, where the vanishing gradient is one of the necessary conditions for optimality, in conventional RSM the first stage ends when

the Euclidean norm of the gradient of the fitted objective is not significantly larger than zero. Then, in the second stage of conventional RSM, a second-order polynomial is fitted, and canonical analysis finds an optimum, or a saddle point, or a ridge. It may be possible to develop a formal stopping rule based on ideas from constrained deterministic optimization. However, this goes beyond the aim and the scope of this paper, and is left for future research.

3.7 Appendix 1: Derivation of the search direction (3.7) by introducing the ellipsoid constraint

We consider the approximation in (3.6), and replace the nonnegativity constraints on the slack vectors (i.e., $\underline{s} \in R_+^{r-1}$, $\underline{r}$, $\underline{v} \in R_+^k$) by an ellipsoid constraint; see Barnes (1986).

$$\begin{array}{ll} \text{minimize} & \underline{b}_0^T \underline{d} \\ \text{subject to} & \mathbf{B}\underline{d} - \underline{s} = \underline{c}, \\ & \underline{d} + \underline{r} = \underline{u}, \\ & \underline{d} - \underline{v} = \underline{l}, \\ & \|\underline{q} - \widehat{\underline{q}}\|_{\widehat{\mathbf{Q}}^{-1}} \leq \rho, \end{array}$$

where $0 < \rho < 1$, $\underline{q} = \left(\underline{s}^T,\ \underline{r}^T,\ \underline{v}^T\right)^T$, $\widehat{\underline{q}} = \left(\widehat{\underline{s}}^T,\ \widehat{\underline{r}}^T,\ \widehat{\underline{v}}^T\right)^T$, $\widehat{\mathbf{Q}} = diag\left(\widehat{\underline{s}},\ \widehat{\underline{r}},\ \widehat{\underline{v}}\right)$ with $\underline{q} \in R^{r-1+2k}$, $\widehat{\underline{q}} \in R^{r-1+2k}$, $\underline{d} \in R^k$, and $\widehat{\mathbf{Q}} \in R^{(r-1+2k)\times(r-1+2k)}$. Note that the entries in $\widehat{\mathbf{Q}}$ are the components of the current slack vectors $\widehat{\underline{s}}$, $\widehat{\underline{r}}$, $\widehat{\underline{v}} > \underline{0}$. The ellipsoid constraint in the above formulation can be rewritten as

$$\left(\underline{s} - \widehat{\underline{s}}\right)^T \widehat{\mathbf{S}}^{-2} \left(\underline{s} - \widehat{\underline{s}}\right) + \left(\underline{r} - \widehat{\underline{r}}\right)^T \widehat{\mathbf{R}}^{-2} \left(\underline{r} - \widehat{\underline{r}}\right) + \left(\underline{v} - \widehat{\underline{v}}\right)^T \widehat{\mathbf{V}}^{-2} \left(\underline{v} - \widehat{\underline{v}}\right) \leq \rho^2.$$

By substituting the slack vectors in terms of $\underline{d}$ (i.e., $\underline{s} = \mathbf{B}\underline{d} - \underline{c}$, $\underline{r} = \underline{u} - \underline{d}$, $\underline{v} = \underline{d} - \underline{l}$) into the ellipsoid constraint, we obtain

$$\left(\underline{d} - \widehat{\underline{d}}\right)^T \mathbf{B}^T \widehat{\mathbf{S}}^{-2} \mathbf{B} \left(\underline{d} - \widehat{\underline{d}}\right) + \left(\underline{d} - \widehat{\underline{d}}\right)^T \widehat{\mathbf{R}}^{-2} \left(\underline{d} - \widehat{\underline{d}}\right) + \left(\underline{d} - \widehat{\underline{d}}\right)^T \widehat{\mathbf{V}}^{-2} \left(\underline{d} - \widehat{\underline{d}}\right) \leq \rho^2.$$

Now the approximation can be rewritten as

$$\begin{array}{ll} \text{minimize} & \underline{b}_0^T \underline{d} \\ \text{subject to} & \left(\underline{d} - \widehat{\underline{d}}\right)^T \mathbf{B}^T \widehat{\mathbf{S}}^{-2} \mathbf{B} \left(\underline{d} - \widehat{\underline{d}}\right) + \left(\underline{d} - \widehat{\underline{d}}\right)^T \widehat{\mathbf{R}}^{-2} \left(\underline{d} - \widehat{\underline{d}}\right) \\ & \qquad + \left(\underline{d} - \widehat{\underline{d}}\right)^T \widehat{\mathbf{V}}^{-2} \left(\underline{d} - \widehat{\underline{d}}\right) \leq \rho^2. \end{array}$$

Since the problem is to minimize a linear function over an ellipsoid, the optimum will occur at the boundary. So we can replace the $\leq$-constraint by a $=$-constraint and form the Lagrange function $L\left(\underline{d},\ \mu\right)$, where μ stands for the Lagrange multiplier. Hence, from the first-order necessary conditions (that is, $\partial L/\partial d_i = 0$ and $\partial L/\partial \mu = 0$), we derive

$$\underline{b}_0 - 2\mu[\mathbf{B}^T \widehat{\mathbf{S}}^{-2} \mathbf{B} \left(\underline{d} - \widehat{\underline{d}}\right) + \widehat{\mathbf{R}}^{-2} \left(\underline{d} - \widehat{\underline{d}}\right) + \widehat{\mathbf{V}}^{-2} \left(\underline{d} - \widehat{\underline{d}}\right)] = \underline{0}$$

$$\left(\underline{d} - \widehat{\underline{d}}\right)^T \mathbf{B}^T \widehat{\mathbf{S}}^{-2} \mathbf{B} \left(\underline{d} - \widehat{\underline{d}}\right) + \left(\underline{d} - \widehat{\underline{d}}\right)^T \widehat{\mathbf{R}}^{-2} \left(\underline{d} - \widehat{\underline{d}}\right) + \left(\underline{d} - \widehat{\underline{d}}\right)^T \widehat{\mathbf{V}}^{-2} \left(\underline{d} - \widehat{\underline{d}}\right) = \rho^2.$$

From the first equality above, we derive

$$\underline{d} = \underline{\widehat{d}} - \frac{1}{2\mu}\left[-\left(\mathbf{B}^T\widehat{\mathbf{S}}^{-2}\mathbf{B} + \widehat{\mathbf{R}}^{-2} + \widehat{\mathbf{V}}^{-2}\right)^{-1}\underline{b}_0\right].$$

In this expression, the term within the brackets gives the proposed search direction, denoted as $\underline{p}$ in (3.7) in §3.3.

3.8 Appendix 2: An alternative derivation of the search direction (3.7)

We include an alternative derivation of the proposed search direction, which shows the generalization of the steepest descent used in conventional RSM to our proposed search direction, namely scaled and projected steepest descent. We follow the same steps as in Monma and Morton (1987) to derive the search direction. Consider the approximation in (3.6):

$$\begin{array}{ll} \text{minimize} & \underline{b}_0^T\underline{d} \\ \text{subject to} & \mathbf{B}\underline{d} - \underline{s} = \underline{c}, \\ & \underline{d} + \underline{r} = \underline{u}, \\ & \underline{d} - \underline{v} = \underline{l}, \\ & \underline{s} \in R_+^{r-1},\ \underline{r},\ \underline{v} \in R_+^k. \end{array}$$

where $\mathbf{B} \in R^{(r-1)\times k}$, $\underline{c} \in R^{r-1}$, and $\underline{d}$, $\underline{b}_0$, $\underline{l}$, and $\underline{u} \in R^k$. Given a feasible solution $\left(\underline{\widehat{d}},\ \underline{\widehat{s}},\ \underline{\widehat{r}},\ \underline{\widehat{v}}\right)$ with $\underline{\widehat{s}}, \underline{\widehat{r}}, \underline{\widehat{v}} > \underline{0}$, we define $\widehat{\mathbf{S}} = diag\left(\underline{\widehat{s}}\right)$, $\widehat{\mathbf{R}} = diag\left(\underline{\widehat{r}}\right)$, and $\widehat{\mathbf{V}} = diag\left(\underline{\widehat{v}}\right)$.

Next we perform an affine transformation on the slack vectors: $\underline{t} = \widehat{\mathbf{S}}^{-1}\underline{s}$, $\underline{z} = \widehat{\mathbf{R}}^{-1}\underline{r}$, and $\underline{y} = \widehat{\mathbf{V}}^{-1}\underline{v}$. Then, we scale the constraints:

$$\begin{array}{ll} \text{minimize} & \underline{b}_0^T\underline{d} \\ \text{subject to} & \widehat{\mathbf{S}}^{-1}\mathbf{B}\underline{d} - \underline{t} = \widehat{\mathbf{S}}^{-1}\underline{c}, \\ & \widehat{\mathbf{R}}^{-1}\underline{d} + \underline{z} = \widehat{\mathbf{R}}^{-1}\underline{u}, \\ & \widehat{\mathbf{V}}^{-1}\underline{d} - \underline{y} = \widehat{\mathbf{V}}^{-1}\underline{l}, \\ & \underline{t} \in R_+^{r-1},\ \underline{z},\ \underline{y} \in R_+^k. \end{array}$$

By simplifying the formulation above, we obtain

$$\text{minimize } \underline{b}_0^T\underline{d} \quad \text{subject to } \mathbf{H}\underline{d} + \underline{h} = \underline{g}$$

where $\mathbf{H} = \left(\mathbf{B}^T\widehat{\mathbf{S}}^{-1},\ \widehat{\mathbf{R}}^{-1},\ \widehat{\mathbf{V}}^{-1}\right)^T$, $\underline{h} = \left(-\underline{t}^T,\ \underline{z}^T,\ -\underline{y}^T\right)^T$ and $\underline{g} = \left(\underline{c}^T\widehat{\mathbf{S}}^{-1},\ \underline{u}^T\widehat{\mathbf{R}}^{-1},\ \underline{l}^T\widehat{\mathbf{V}}^{-1}\right)^T$. Let $\mathbf{I}$ and $\mathbf{0}$ be the identity and zero-matrix, respectively. Assuming that $\mathbf{H}$ has full

rank, we obtain a QR factorization given by the orthonormal matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ with $\mathbf{Q}_1^T\mathbf{Q}_1 = \mathbf{I}$, $\mathbf{Q}_2^T\mathbf{Q}_2 = \mathbf{I}$, and $\mathbf{Q}_2^T\mathbf{Q}_1 = \mathbf{0}$, and a nonsingular upper triangular matrix $\mathbf{R}$ so that $\mathbf{H} = (\mathbf{Q}_1, \mathbf{Q}_2)\left(\mathbf{R}^T, \mathbf{0}\right)^T = \mathbf{Q}_1\mathbf{R}$. After this QR-factorization for $\mathbf{H}$, the formulation can be rewritten as

$$\text{minimize } \underline{b}_0^T\underline{d} \quad \text{subject to } \mathbf{Q}_1\mathbf{R}\underline{d} + \underline{h} = \underline{g}.$$

We multiply the equality constraint first by $\mathbf{Q}_1^T$ and then by $\mathbf{Q}_2^T$, to obtain

$$\mathbf{Q}_1^T\mathbf{Q}_1\mathbf{R}\underline{d} + \mathbf{Q}_1^T\underline{h} = \mathbf{Q}_1^T\underline{g} \text{ and } \mathbf{Q}_2^T\mathbf{Q}_1\mathbf{R}\underline{d} + \mathbf{Q}_2^T\underline{h} = \mathbf{Q}_2^T\underline{g}.$$

Finally we solve for $\underline{d}$: $\underline{d} = \mathbf{R}^{-1}\mathbf{Q}_1^T\left(\underline{g} - \underline{h}\right)$ to obtain

$$\text{minimize } -\underline{b}_0^T\mathbf{R}^{-1}\mathbf{Q}_1^T\underline{h} + \underline{b}_0^T\mathbf{R}^{-1}\mathbf{Q}_1^T\underline{g} \quad \text{subject to } \mathbf{Q}_2^T\underline{h} = \mathbf{Q}_2^T\underline{g}.$$

Now, we obtain the search direction $\underline{p^h}$ in the slack space. The steepest descent direction is given by $\underline{b}_0^T\mathbf{R}^{-1}\mathbf{Q}_1^T$. To maintain the feasibility of the equality constraint, we project the steepest descent direction onto the null-space of $\mathbf{Q}_2^T$. Note that $\left(\mathbf{I} - \mathbf{Q}_2\left(\mathbf{Q}_2^T\mathbf{Q}_2\right)^{-1}\mathbf{Q}_2^T\right)$ is the matrix that projects any vector onto the null-space of $\mathbf{Q}_2^T$. Hence the search direction $\underline{p^h}$ in the slack space is given by:

$$\begin{aligned}\underline{p^h} &= \left(\mathbf{I} - \mathbf{Q}_2\left(\mathbf{Q}_2^T\mathbf{Q}_2\right)^{-1}\mathbf{Q}_2^T\right)\mathbf{Q}_1\mathbf{R}^{-T}\underline{b}_0 \\ &= \mathbf{Q}_1\mathbf{R}^{-T}\underline{b}_0.\end{aligned}$$

Without loss of generality, we assume unit step length at both directions $\underline{p^h}$ and $\underline{p^d}$. Suppose that $\underline{d}$ and $\underline{h}$ are feasible; that is, $\mathbf{Q}_1\mathbf{R}\underline{d} + \underline{h} = \underline{g}$. To maintain feasibility, the search direction $\underline{p^d}$ in the $\underline{d}$ space has to satisfy the following equalities:

$$\begin{aligned}\mathbf{Q}_1\mathbf{R}\left(\underline{d} + \underline{p^d}\right) + \left(\underline{h} + \underline{p^h}\right) &= \underline{g}, \\ \mathbf{Q}_1\mathbf{R}\underline{p^d} + \underline{p^h} &= \underline{0}, \\ \mathbf{R}\underline{p^d} + \mathbf{Q}_1^T\underline{p^h} &= \underline{0}.\end{aligned}$$

Therefore $\underline{p^d} = -\mathbf{R}^{-1}\mathbf{Q}_1^T\underline{p^h} = -\mathbf{R}^{-1}\mathbf{Q}_1^T\mathbf{Q}_1\mathbf{R}^{-T}\underline{b}_0$. From $\mathbf{Q}_1^T\mathbf{Q}_1 = \mathbf{I}$, we have $\mathbf{Q}_1^T = \mathbf{Q}_1^{-1}$ and $\mathbf{Q}_1 = \mathbf{Q}_1^{-T}$. Replacing $\mathbf{Q}_1^T$ by $\mathbf{Q}_1^{-1}$, $\mathbf{Q}_1$ by $\mathbf{Q}_1^{-T}$ and $\mathbf{Q}_1\mathbf{R}$ by $\mathbf{H}$, we obtain

$$\begin{aligned}\underline{p^d} &= -\mathbf{R}^{-1}\mathbf{Q}_1^{-1}\mathbf{Q}_1^{-T}\mathbf{R}^{-T}\underline{b}_0 \\ &= -\left(\mathbf{Q}_1\mathbf{R}\right)^{-1}\left(\mathbf{R}^T\mathbf{Q}_1^T\right)^{-1}\underline{b}_0 \\ &= -\mathbf{H}^{-1}\mathbf{H}^{-T}\underline{b}_0 \\ &= -\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\underline{b}_0.\end{aligned}$$

Setting back $\mathbf{H} = \left(\mathbf{B}^T\widehat{\mathbf{S}}^{-1},\ \widehat{\mathbf{R}}^{-1},\ \widehat{\mathbf{V}}^{-1}\right)^T$ leads to the proposed search direction, denoted as $\underline{p}$ in (3.7) in §3.3.

3.9 Appendix 3: Scale independence of the search direction (3.7)

We consider a general linear transformation of the variables: $\mathbf{J}\underline{z}+\underline{e}=\underline{d}$ where $\mathbf{J}\in R^{k\times k}$ is non-singular, and $\underline{e}$ R^k. This transformation is one-to-one, and $\underline{z}=\mathbf{J}^{-1}(\underline{d}-\underline{e})$. The final result proven in this appendix will be a simple relation between the search directions $\underline{p^d}$ and $\underline{p^z}$ in the $\underline{d}$ and $\underline{z}$ spaces, which implies scale independence: $\underline{p^z}=\mathbf{J}^{-1}\underline{p^d}$.

Defining $\widetilde{\underline{z}}=\left(1,\ \underline{z}^T\right)^T$ and $\widetilde{\underline{d}}=\left(1,\ \underline{d}^T\right)^T$, $\mathbf{J}\underline{z}+\underline{e}=\underline{d}$ can be written as: $\mathbf{G}\widetilde{\underline{z}}=\widetilde{\underline{d}}$ where

$$\mathbf{G}=\begin{pmatrix}1 & \mathbf{0}\\ \underline{e} & \mathbf{J}\end{pmatrix}$$

and $\mathbf{G}$ is non-singular. Given that $\mathbf{Z}=\mathbf{X}\mathbf{G}^{-T}$, the least squares estimator $\widehat{\beta}_i^{\widetilde{z}}$ in the $\underline{z}$ space is given by:

$$\begin{aligned}\widehat{\beta}_i^{\widetilde{z}}&=\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\widehat{\underline{F}}_i\\&=\left(\mathbf{G}^{-1}\mathbf{X}^T\mathbf{X}\mathbf{G}^{-T}\right)^{-1}\mathbf{G}^{-1}\mathbf{X}^T\widehat{\underline{F}}_i\\&=\mathbf{G}_i^T\widehat{\beta}_i^{\widetilde{d}}\end{aligned}$$

where $\widehat{\beta}_i^{\widetilde{d}}$ is given in (3.3). Now, we can write (3.5) in terms of $\underline{z}$, as follows:

$$\begin{array}{ll}\text{minimize} & \underline{b}_0^T\mathbf{J}\underline{z}\\ \text{subject to} & \underline{b}_j^T\mathbf{J}\underline{z}\geq c_j-\underline{b}_j^T\underline{e}\ \text{ for }\ j=1,\ ...,\ r-1,\\ & \mathbf{J}^{-1}(\underline{l}-\underline{e})\leq\underline{z}\leq\mathbf{J}^{-1}(\underline{u}-\underline{e}).\end{array}$$

Hence, from the above formulation we have: $\mathbf{B}_{\underline{z}}=\mathbf{B}_{\underline{d}}\mathbf{J}$. By further simplifying the formulation and adding the slacks $\underline{s_z}$, $\underline{r_z}$, and $\underline{v_z}$ we obtain

$$\begin{array}{ll}\text{minimize} & \underline{b}_0^T\mathbf{J}\underline{z}\\ \text{subject to} & \mathbf{B}_{\underline{d}}\mathbf{J}\underline{z}-\underline{s_z}=\underline{c}-\mathbf{B}_{\underline{d}}\underline{e},\\ & \underline{z}+\underline{r_z}=\mathbf{J}^{-1}(\underline{u}-\underline{e}),\\ & \underline{z}-\underline{v_z}=\mathbf{J}^{-1}(\underline{l}-\underline{e}),\\ & \underline{s_z},\ \underline{r_z},\ \underline{v_z}\geq 0.\end{array}$$

Hence

$$\begin{aligned}
\underline{s}_{\underline{z}} &= \mathbf{B}_{\underline{d}}\mathbf{J}\underline{z} - \underline{c} + \mathbf{B}_{\underline{d}}\underline{e} \\
&= \mathbf{B}_{\underline{d}}\mathbf{J}\left[\mathbf{J}^{-1}\left(\underline{d} - \underline{e}\right)\right] - \underline{c} - \mathbf{B}_{\underline{d}}\underline{e} \\
&= \mathbf{B}_{\underline{d}}\underline{d} - \underline{c} = \underline{s}_{\underline{d}}, \\
\underline{r}_{\underline{z}} &= \mathbf{J}^{-1}\left(\underline{u} - \underline{e}\right) - \underline{z} \\
&= \mathbf{J}^{-1}\left(\underline{u} - \underline{e}\right) - \mathbf{J}^{-1}\left(\underline{d} - \underline{e}\right) \\
&= \mathbf{J}^{-1}\left(\underline{u} - \underline{d}\right) = \mathbf{J}^{-1}\underline{r}_{\underline{d}}, \\
\underline{v}_{\underline{z}} &= \underline{z} - \mathbf{J}^{-1}\left(\underline{l} - \underline{e}\right) \\
&= \mathbf{J}^{-1}\left(\underline{d} - \underline{e}\right) - \mathbf{J}^{-1}\left(\underline{l} - \underline{e}\right) \\
&= \mathbf{J}^{-1}\left(\underline{d} - \underline{l}\right) = \mathbf{J}^{-1}\underline{v}_{\underline{d}}.
\end{aligned}$$

Thus, $\mathbf{B}_{\underline{z}} = \mathbf{B}_{\underline{d}}\mathbf{J}$, $\widehat{\mathbf{S}}_{\underline{z}} = \widehat{\mathbf{S}}_{\underline{d}}$, $\widehat{\mathbf{R}}_{\underline{z}} = \widehat{\mathbf{R}}_{\underline{d}}\mathbf{J}^{-T}$, and $\widehat{\mathbf{V}}_{\underline{z}} = \widehat{\mathbf{V}}_{\underline{d}}\mathbf{J}^{-T}$. Remember that the search direction in $\underline{d}$ is given by $\underline{p}^{\underline{d}} = -\left(\mathbf{B}_{\underline{d}}^{T}\widehat{\mathbf{S}}_{\underline{d}}^{-2}\mathbf{B}_{\underline{d}} + \widehat{\mathbf{R}}_{\underline{d}}^{-2} + \widehat{\mathbf{V}}_{\underline{d}}^{-2}\right)^{-1}\underline{b}_0$. Therefore the search direction in $\underline{z}$ space is given by

$$\begin{aligned}
\underline{p}^{\underline{z}} &= -\left(\mathbf{J}^{T}\mathbf{B}_{\underline{d}}^{T}\widehat{\mathbf{S}}_{\underline{d}}^{-2}\mathbf{B}_{\underline{d}}\mathbf{J} + \mathbf{J}^{T}\widehat{\mathbf{R}}_{\underline{d}}^{-2}\mathbf{J} + \mathbf{J}^{T}\widehat{\mathbf{V}}_{\underline{d}}^{-2}\mathbf{J}\right)^{-1}\mathbf{J}^{T}\underline{b}_0 \\
&= -\mathbf{J}^{-1}\left(\mathbf{B}_{\underline{d}}^{T}\widehat{\mathbf{S}}_{\underline{d}}^{-2}\mathbf{B}_{\underline{d}} + \widehat{\mathbf{R}}_{\underline{d}}^{-2} + \widehat{\mathbf{V}}_{\underline{d}}^{-2}\right)^{-1}\mathbf{J}^{-T}\mathbf{J}^{T}\underline{b}_0 \\
&= \mathbf{J}^{-1}\underline{p}^{\underline{d}}.
\end{aligned}$$

Chapter 4

An Asymptotic Stopping Rule for Simulation Optimization

4.1 Introduction

Optimization in simulation is attempted by many methods; for a recent survey, see Fu (2002). In particular, when simulation is treated as a black box (that is, gradient estimation through methods such as perturbation analysis (Glasserman (1991)) or likelihood ratio score function (Rubinstein and Shapiro (1993)) is not applicable), metaheuristics such as simulated annealing, genetic algorithms, tabu search, or scatter search (see, for example, Fouskakis and Draper (2002)), the simultaneous perturbation stochastic approximation of Spall (2003), and Kleinman, Spall, and Naiman (1999), or Response Surface Methodology (RSM) of Myers and Montgomery (2002) can be used for optimization.

We deal with problems that have a stochastic objective function and stochastic constraints. More specifically, we derive a statistical stopping rule that assumes large samples. Under a constraint qualification (regularity condition), this rule tests for the well-established first-order necessary optimality conditions of deterministic optimization (see, for example, Gill, Murray, and Wright (2000, p. 81)) at a feasible point. To derive this rule, we use the Delta method, which shows that under certain conditions, nonlinear statistics are approximately multivariate normally distributed. Also, we apply the procedure in Kodde and Palm (1986) that uses Wald's statistic, which enables us to test composite hypotheses (that is, hypotheses with vectorial equalities and inequalities). Notice that in a very recent paper, Bettonvil and Kleijnen (2004) derives a novel procedure to test the first-order necessary optimality conditions for com-

putationally expensive, black box simulation optimization problems, using bootstrap procedure.

We also present two alternative lack-of-fit tests, namely Roy's largest root test and the classic F test combined with Bonferroni's inequality, which are only used in the RSM literature. Notice that the former test was already introduced to RSM with multiple responses by Khuri (1996). Additionally, we show in Appendix 1 that Roy's test is a generalization of the classic F test to multiple responses.

We use underlined letters and bold letters to denote vectors and matrices, respectively. In addition, $\underline{0}$, $\underline{1}$, $\mathbf{0}$, and $\mathbf{1}$ stand for a vector of zeros, a vector of ones, a square matrix of zeros, and a square matrix of ones, respectively. For square matrices, we use subscripts to denote their dimensions; for example, $\mathbf{I}_k$ denotes a $k \times k$ identity matrix.

The remainder of this paper is organized as follows. In §4.2, we formalize our problem including some statistical issues. In §4.3, we present two alternative lack-of-fit tests. In §4.4, we present the proposed stopping rule. In §4.5, we evaluate our stopping rule through a Monte Carlo example. In §4.6, we give conclusions. In two appendices, we give technical details and proofs.

4.2 Problem formulation

Our problem is as follows:

$$\begin{array}{ll} \text{minimize} & E_{\underline{\omega}}\left[F_0\left(\underline{d},\ \underline{\omega}\right)\right] \\ \text{subject to} & E_{\underline{\omega}}\left[F_j\left(\underline{d},\ \underline{\omega}\right)\right] \leq a_j \quad \text{for} \quad j = 1,\ ...,\ r-1 \end{array} \tag{4.1}$$

where $E_{\underline{\omega}}$ is the expectation operator with respect to the simulation's seed vector $\underline{\omega}$, the F_i $(i = 0,\ ...,\ r-1)$ are r random responses that are estimated through simulation, $\underline{d}$ is $k \times 1$ vector of input variables, and a_j is the jth component of the $(r-1) \times 1$ deterministic right-hand-side vector. An example of (4.1) is the minimization of the expected inventory-carrying and ordering costs (excluding stock-out costs) such that a prespecified customer service level is satisfied; obviously, in this example $r = 2$.

In the following, we explain how to approximate the F_i in (4.1) locally by first-order polynomials. For clarity in the description, we use the RSM's approach to estimate the coefficients of these polynomials; that is, we use linear regression. In general, one can use different methods - other than linear regression - to obtain these coefficients, provided that certain conditions, which will be given in Appendix 2, are satisfied. The important issue is to approximate the F_i in (4.1) locally. Let n

$(n \geq k+1)$ be the number of distinct simulated input combinations per local area - usually specified by a resolution-3 design in the first stage of RSM. Further, let m_l be the number of replicates at the lth $(l = 1,\ ...,\ n)$ input combination; let N be the resulting total number of local runs; that is, $N = m_1 + ... + m_l + ... + m_n$. By definition, the first stage of RSM means that we locally approximate the r simulation responses F_i in (4.1) by r first-order polynomial regression metamodels with additive noises:

$$\underline{G}_i\left(\underline{d},\ \underline{\omega}\right) = \mathbf{X}\underline{\beta}_i + \underline{\epsilon}_i\left(\underline{d},\ \underline{\omega}\right) \quad \text{for } i = 0,\ ...,\ r-1 \tag{4.2}$$

where $\mathbf{X}$ is the $N \times (k+1)$ matrix obtained by adding a column of ones to the design matrix, say, $\mathbf{D}$ (i.e., $\mathbf{X} = (\underline{1},\ \mathbf{D})$), $\underline{\beta}_i$ denotes the $(k+1) \times 1$ vector of regression coefficients for response type i, and $\underline{\epsilon}_i$ is the $N \times 1$ vector of additive noise with mean vector $\underline{\mu}_{\underline{\epsilon}_i}$ and covariance matrix $\mathbf{\Sigma}_{\underline{\epsilon}_i}$ for response type i. This $\underline{\epsilon}_i$ accounts for both the lack of fit of the first-order polynomial approximations and the inherent randomness in stochastic simulation created through the use of the simulation's seed $\underline{\omega}$. The metamodel i is called "adequate" (no lack-of-fit) if $\underline{\mu}_{\underline{\epsilon}_i} = E\left(\underline{\epsilon}_i\right) = \underline{0}$. Furthermore, we assume that $\underline{\epsilon} = (\epsilon_0,\ ...,\ \epsilon_i,\ ...,\ \epsilon_{r-1})^T$ is multivariate normally distributed.

Classic RSM assumes a single response $(r = 1)$ and $\mathbf{\Sigma}_{\underline{\epsilon}_i} = \sigma_{i,i}\mathbf{I}$. In the rest of this paper, we call the assumptions $\underline{\mu}_{\underline{\epsilon}_i} = \underline{0}$ and $\mathbf{\Sigma}_{\underline{\epsilon}_i} = \sigma_{i,i}\mathbf{I}$ the ordinary least squares (OLS) assumptions. When these assumptions are satisfied, OLS gives the best linear unbiased estimator (BLUE) of $\underline{\beta}_i$, where "best" means minimum variance. In our problem, however, we have r responses that are statistically dependent (correlated); moreover, the variances per responses differ in general. Hence, generalized least squares (GLS) gives the BLUE of $\underline{\beta}_i$. However, under certain conditions GLS reduces to OLS, as we explain now.

Let $\underline{\widehat{F}}_i\left(\underline{d},\ \underline{\omega}\right) = \left(\underline{\widehat{F}}_i\left(\underline{d}_1,\ \underline{\omega}_1\right)^T,\ ...,\ \underline{\widehat{F}}_i\left(\underline{d}_l,\ \underline{\omega}_l\right)^T,\ ...,\ \underline{\widehat{F}}_i\left(\underline{d}_n,\ \underline{\omega}_n\right)^T\right)^T$ be the $N \times 1$ vector of estimators of $E\left[F_i\left(\underline{d},\ \underline{\omega}\right)\right]$, the expected response i in (4.1), where $\underline{d}_l$ is the $m_l \times 1$ vector of replicates for the lth input combination and $\underline{\omega}_l$ is the $m_l \times 1$ seed vector with independent components. Moreover, the seed vectors, say $\underline{\omega}_l$ and $\underline{\omega}_{l'}$ $\left(l' = 1,\ ...,\ n,\ l' \neq l\right)$, for distinct input combinations l and l' are independent. In other words, we assume that the m_l seeds give non-overlapping pseudo-random number (PRN) streams, so the m_l replicates are independent. Moreover, we do not use common random numbers, so all N simulation outputs are independent.

Rao (1967) proves that if the same design is used for all r responses and if each response satisfies the OLS assumptions, then the GLS estimator of the r coefficient vectors $\left(\underline{\beta}_0^T,\ ...,\ \underline{\beta}_{r-1}^T\right)^T$ reduces to the OLS estimator of the individual coefficient

vector $\underline{\beta}_i$:

$$\widehat{\underline{\beta}}_i\left(\underline{d},\ \underline{\omega}\right) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\widehat{\underline{F}}_i\left(\underline{d},\ \underline{\omega}\right) \text{ for } i = 0,\ ...,\ r-1. \tag{4.3}$$

The r simulation responses of a specific replicate are correlated, since they map the same PRN using different transformation functions. For example, the inventory simulation example records the realized ordering costs and the service percentage per simulation run or replicate. The r simulation responses have different variances. For example, ordering costs and service percentage even have different dimensions.

The OLS assumptions imply $\mathbf{\Sigma}_{\underline{\epsilon}_i} = \sigma_{i,i}\mathbf{I}$. Now, let $\mathbf{\Sigma}$ be the $r \times r$ covariance matrix with entry $(i,\ h)$ being $\sigma_{i,h}$ $(h = 0,\ ...,\ r-1)$. Often, simulation analysts estimate $\mathbf{\Sigma}$ through the m_l replicates, which results in $\sum_{l=1}^{n}(m_l - 1) = N - n$ degrees of freedom. However, we prefer an alternative estimator, namely the mean squared residual (MSR), since this alternative has more degrees of freedom, namely $N - (k+1)$, provided $n > k+1$; $n > k+1$ implies a non-saturated design (some resolution-3 designs are saturated). In the following, $\widehat{\underline{G}}_i\left(\underline{d},\ \underline{\omega}\right) = \mathbf{X}\widehat{\underline{\beta}}_i\left(\underline{d},\ \underline{\omega}\right)$ denotes the $N \times 1$ vector of OLS estimators of $E\left[\underline{G}_i\left(\underline{d},\ \underline{\omega}\right)\right]$, the expectation of the linear approximation in (4.2). Indeed, Khuri (1996, p. 385) gives the following MSR estimator of $\sigma_{i,h}$:

$$\widehat{\sigma}_{i,h}\left(\underline{d},\ \underline{\omega}\right) = \frac{\left(\widehat{\underline{F}}_i\left(\underline{d},\ \underline{\omega}\right) - \widehat{\underline{G}}_i\left(\underline{d},\ \underline{\omega}\right)\right)^T\left(\widehat{\underline{F}}_h\left(\underline{d},\ \underline{\omega}\right) - \widehat{\underline{G}}_h\left(\underline{d},\ \underline{\omega}\right)\right)}{N - (k+1)}. \tag{4.4}$$

We assume constant covariances within the local area (with its N simulation runs); i.e., $\sigma_{i,h}\left(\underline{d}\right) = \sigma_{i,h}$ locally. The MSR estimators in (4.4) are unbiased if the responses satisfy the OLS assumptions locally; see Theil (1971, p. 114). Note that for $i = h$, (4.4) is found in any classic RSM or regression textbook, where "classic" means "no multiple responses". Notice also that (4.4) is specific to the RSM literature. In general, one can use a pooled sample variance estimate, instead of (4.4).

4.3 Testing the lack of fit

Since RSM uses locally fitted first-order polynomials, it is prudent to test the local first-order approximations in (4.2) for lack of fit; see Myers and Montgomery (2002, p. 51). If there is no lack of fit, then we use the OLS estimators in (4.3) to test the optimality conditions in §4.4. However, if there is lack of fit, then classic RSM switches to alternative regression metamodels, using transformations of $\underline{d}$ such as $\log d_{k'}$ or $1/d_{k'}$, where $d_{k'}$ is the k'th component of $\underline{d}$, or to second-order polynomials. In this paper, we do not consider switching to different models; instead, we refer to, for example, Irizarry et al. (2003).

We discuss two alternative lack-of-fit tests for multivariate responses, namely Roy's largest root test (§4.3.1) and the classic F test and Bonferroni's inequality (§4.3.2). We remind that the results of this section belong only to the RSM literature.

4.3.1 Roy's largest root test

Based on classic RSM, we extend the lack-of-fit test for multiple responses, as follows. We define a joint, "optimistic" null hypothesis:

$$H_0 : \forall i : E\left[F_i\left(\underline{d},\ \underline{\omega}\right)\right] = E\left[G_i\left(\underline{d},\ \underline{\omega}\right)\right]. \tag{4.5}$$

where F_i and G_i were defined in (4.1) and (4.2). H_0 is "optimistic" since it is rejected only if there is significant lack of fit for any of the r responses. Note that failing to reject the lack-of-fit test H_0 in (4.5) does not guarantee that this H_0 holds (type I error). We test (4.5) by Roy's largest root test, explained in Roy, Gnanadesikan, and Srivastava (1971). This test is a multivariate version of the ordinary lack-of-fit F test; i.e., it reduces to the ordinary test when there is a single response - as we show in Appendix 1.

As in the classic lack-of-fit F test, Roy's test requires replicates at some design points - to estimate the pure error. (As a reminder, n $(n > k+1)$ is the number of distinct simulated input combinations, m_l is the number of replicates at the lth input combination, and N is the total number of local runs.) Let q $(1 \leq q \leq n)$ be the number of distinct input combinations with at least two replicates. The total number of replicates at these q input combinations must satisfy the following condition: $(m_1 - 1) + ... + (m_q - 1) \geq r$, where r is the number of response types; see Roy, Gnanadesikan, and Srivastava (1971, p. 35). Below we summarize the version in Roy, Gnanadesikan, and Srivastava (1971) with $q = n$ and $m_1 = ... = m_n = m$; we explain the general setting in Roy, Gnanadesikan, and Srivastava (1971) and the relation of this test with the classic F test (novel results) in Appendix 1.

Roy's test considers the weighted sums of all r responses $b_0\underline{G}_0\left(\underline{d},\ \underline{\omega}\right) + ... + b_{r-1}\underline{G}_{r-1}\left(\underline{d},\ \underline{\omega}\right)$ for all $r \times 1$ nonzero weight vectors $\underline{b}$. Roy, Gnanadesikan, and Srivastava (1971) defines $r \times r$ matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ as follows:

$$\mathbf{Q}_1 = \widehat{\mathbf{F}}^T\left[\mathbf{I}_N - \mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-}\mathbf{Z}^T - \mathbf{K}\right]\widehat{\mathbf{F}} \text{ and } \mathbf{Q}_2 = \widehat{\mathbf{F}}^T\mathbf{K}\widehat{\mathbf{F}}$$

where $\widehat{\mathbf{F}} = \left[\underline{\widehat{F}}_0,\ ...,\ \underline{\widehat{F}}_{r-1}\right]$ is the $N \times r$ matrix of simulation estimators for the expected responses in (4.1), $\mathbf{Z}$ is the $N \times (r\,(k+1))$ matrix consisting of r identical $\mathbf{X}$ matrices,

$\left(\mathbf{Z}^T\mathbf{Z}\right)^-$ is a generalized inverse of $\mathbf{Z}^T\mathbf{Z}$, and $\mathbf{K}$ is the $N \times N$ block-diagonal matrix with $\mathbf{K}_l = \mathbf{I}_m - (1/m)\,\mathbf{1}_m$ on the diagonal, since there are m replicates at each input combination l; i.e., $\mathbf{K} = diag\,(\mathbf{K}_1,\ ...,\ \mathbf{K}_n)$. Notice that the condition on the total number of replicates (namely, $(m_1 - 1) + ... + (m_q - 1) \geq r$) ensures that $\mathbf{Q}_2$ is positive definite; see Roy, Gnanadesikan, and Srivastava (1971, p. 35).

For readers familiar with the classic lack-of-fit F test, it is clear that $\mathbf{Q}_1$ is the difference between the sum of squared residuals SS_R and the sum of squared pure errors SS_{PE} for r responses; i.e., $\mathbf{Q}_1$ represents the sum of squared lack of fit (say) SS_{LOF} (see Appendix 1). Appendix 1 further shows that $\mathbf{Q}_2$ represents the sum of squared pure errors SS_{PE} for r responses. $\mathbf{Q}_2$ accounts for both the correlations between the responses and their different variances. Since Roy, Gnanadesikan, and Srivastava (1971) considers the weighted sums of all responses, it also considers the weighted sums of SS_{LOF} and SS_{PE} corresponding to all r responses; that is, $SS_{LOF}\,(\underline{b})$ and $SS_{PE}\,(\underline{b})$.

If H_0 in (4.5) and the multivariate normality assumption hold, then the statistic in (4.6) has the F distribution with ν_1 and ν_2 degrees of freedom, where ν_1 and ν_2 correspond to the degrees of freedom for $SS_{LOF}\,(\underline{b})$ and $SS_{PE}\,(\underline{b})$; that is, $\nu_1 = n - (k+1)$ and $\nu_2 = N - n$. Hence

$$F_{\nu_1,\nu_2}\,(\underline{b}) = \frac{SS_{LOF}\,(\underline{b})\,/\nu_1}{SS_{PE}\,(\underline{b})\,/\nu_2} = \frac{\underline{b}^T\mathbf{Q}_1\underline{b}/\nu_1}{\underline{b}^T\mathbf{Q}_2\underline{b}/\nu_2}. \tag{4.6}$$

Note that to avoid division by $\nu_2 = 0$ in (4.6), n has to satisfy the following condition: $n > (k+1)$. H_0 in (4.5) is rejected if $\underset{\underline{b}\neq\underline{0}}{max} F_{\nu_1,\nu_2}\,(\underline{b})$ exceeds a prespecified critical value, F_{α,ν_1,ν_2}, where α is the significance level (type I error rate). In Appendix 1, we explain how to transform the maximization problem $\underset{\underline{b}\neq\underline{0}}{max} F_{\nu_1,\nu_2}\,(\underline{b})$ into the standard problem in linear algebra of finding the largest eigenvalue, say $\kappa_{\max}\left(\mathbf{Q}_2^{-1}\mathbf{Q}_1\right)$, of the matrix $\mathbf{Q}_2^{-1}\mathbf{Q}_1$. So, the test reduces to checking whether $\kappa_{\max}\left(\mathbf{Q}_2^{-1}\mathbf{Q}_1\right)$ exceeds F_{α,ν_1,ν_2}; that is, H_0 in (4.5) is rejected if

$$\kappa_{\max}\left(\mathbf{Q}_2^{-1}\mathbf{Q}_1\right) \geq F_{\alpha,\nu_1,\nu_2}. \tag{4.7}$$

4.3.2 Classic F test and Bonferroni's inequality

An alternative to Roy's test formulated in (4.7) is the ordinary lack-of-fit F test per response - combined with Bonferroni's inequality; that is, select α_i for each response i such that $\alpha_0 + ... + \alpha_{r-1} = \alpha$. When testing the responses individually, we have r null hypotheses $H_{0,i} : E\,[F_i\,(\underline{d},\ \underline{\omega})] = E\,[G_i\,(\underline{d},\ \underline{\omega})]$ instead of the single, joint H_0 in (4.5). Bonferroni's inequality implies that α constitutes an upper bound on the actual

significance level of the alternative test; i.e., the null hypothesis tends to be rejected less frequently by the alternative test. Whenever the condition $(m_1 - 1) + ... + (m_q - 1) \geq r$ is not satisfied, $\mathbf{Q}_2$ in (4.7) becomes singular, and the alternative F test becomes a good alternative. In our asymptotic cases ($m = 100$ or 250), however, this condition always holds; so Roy's test can be applied.

4.4 The statistical stopping rule

The application of the rule requires the knowledge of the index set $A(\underline{d})$ of the constraints that are binding (active) at $\underline{d}$; that is, slacks s_j are zero. Hence, per constraint j we test the corresponding null hypothesis:

$$H_{0,j}^{(1)} : E\left[s_j\left(\underline{d},\ \underline{\omega}\right)\right] = 0 \tag{4.8}$$

Notice that s_j in (4.8) satisfies $s_j\left(\underline{d},\ \underline{\omega}\right) = a_j - F_j\left(\underline{d},\ \underline{\omega}\right)$, where a_j and F_j were defined in (4.1). Because we assume multivariate normality of $(F_1,\ ...,\ F_{r-1})^T$, the slack vector $(s_1,\ ...,\ s_{r-1})^T$ also has a multivariate normal distribution. Under $H_{0,j}^{(1)}$ in (4.8), the following statistic has a t distribution with $\nu = N - (k+1)$ degrees of freedom:

$$t_{j,\nu} = \frac{\widehat{s}_j\left(\underline{d},\ \underline{\omega}\right)}{\sqrt{\widehat{\sigma}_{j,j}\left(\underline{d},\ \underline{\omega}\right)}} \tag{4.9}$$

where $\widehat{\sigma}_{j,j}$ follows from (4.4) with $i = h = j$. Let $t_{\alpha/2,\nu}$ be the critical value for the two-sided test of (4.8); i.e., $H_{0,j}^{(1)}$ is rejected if the absolute value of $t_{j,\nu}$ exceeds $t_{\alpha/2,\nu}$. These $r-1$ tests give an index set $A(\underline{d})$ that consists of the indices of the binding constraints at $\underline{d}$. We proclaim a constraint to be binding whenever its observed slack value is not significantly positive or negative. If the slack is very positive, then the point $\underline{d}$ is inside the feasible area - away from the boundary; so there is no need to test for optimality (we assume that the optimum occurs on the boundary), and we continue along the search path. If the slack is very negative, then $\underline{d}$ is infeasible; so we do not test for optimality, but back-up along the search path.

Let $\underline{d}^0$ denote a local minimizer of (4.1). Then, - given a constraint qualification - the first-order necessary optimality conditions for $\underline{d}^0$ are:

$$\begin{aligned} &\nabla E\left[F_0\left(\underline{d}^0,\ \underline{\omega}\right)\right] + \sum_{j' \in A\left(\underline{d}^0\right)} \lambda_{j'}^0 \nabla E\left[F_{j'}\left(\underline{d}^0,\ \underline{\omega}\right)\right] = \underline{0} \\ &\lambda_{j'}^0 \geq 0,\ E\left[s_{j'}\left(\underline{d}^0,\ \underline{\omega}\right)\right]\lambda_{j'}^0 = 0 \quad j' \in A\left(\underline{d}^0\right) \end{aligned} \tag{4.10}$$

where $\lambda_{j'}^0$ is the Lagrange multiplier for the j'th binding constraint at $\underline{d}^0$. The conditions in (4.10) imply that at $\underline{d}^0$, the gradient of the objective can be expressed as a

nonnegative linear combination of the gradients of the binding constraints at $\underline{d}^0$. The constraint qualification is relevant when there are nonlinear constraints in (4.1); see Gill, Murray, and Wright (2000, p. 81). There are several forms of constraint qualification, many of which are of theoretical interest. A practical constraint qualification for nonlinear constraints is the condition that the constraint gradients at $\underline{d}^0$ are linearly independent.

Since we do not know the functional forms of $E\left[F_i\left(\underline{d},\ \underline{\omega}\right)\right]$ in (4.1) explicitly, we use the local approximations in (4.2), so $\nabla E\left[F_i\left(\underline{d},\ \underline{\omega}\right)\right] = \underline{\beta}_{i,-0}$, where $\underline{\beta}_{i,-0}$ denotes $\underline{\beta}_i$ excluding the intercept $\beta_{i,0}$. Based on (4.10), we test for the existence of Lagrange multipliers $\lambda_{j'}$, using the following null hypothesis $H_0^{(2)}$ and the alternative hypothesis $H_1^{(2)}$:

$$\begin{aligned} H_0^{(2)} &: \underline{\beta}_{0,-0} + \sum_{j' \in A(\underline{d})} \lambda_{j'} \underline{\beta}_{j',-0} = \underline{0} \text{ and } \lambda_{j'} \geq 0 \text{ for } j' \in A\left(\underline{d}\right) \\ H_1^{(2)} &: \underline{\beta}_{0,-0} + \sum_{j' \in A(\underline{d})} \lambda_{j'} \underline{\beta}_{j',-0} \neq \underline{0} \text{ or } \lambda_{j'} \ngeq 0 \text{ for } j' \in A\left(\underline{d}\right). \end{aligned} \tag{4.11}$$

We consider the point $\underline{d}$ that satisfies (4.11) as an approximation to $\underline{d}^0$. In (4.11), we omit the complementarity conditions $E\left[s_{j'}\left(\underline{d},\ \underline{\omega}\right)\right]\lambda_{j'} = 0$: from (4.8), the realizations of the slacks for the active constraints are not significantly different from zero.

Moreover, to determine the active set $A\left(\underline{d}\right)$, we use the $H_{0,j}^{(1)}$ in (4.8), which tends to consider the constraints to be "active" at $\underline{d}$. Considering "inactive" constraints as "active" may cause fewer problems than considering "active" constraints as "inactive" - as we explain now. Suppose that there are $\left|A\left(\underline{d}^0\right)\right|$ active constraints at $\underline{d}^0$, where $|.|$ denotes the cardinality of the finite set $A\left(\underline{d}^0\right)$. However, (4.9) gives us these $\left|A\left(\underline{d}^0\right)\right|$ constraints plus - falsely - one extra constraint (say) the $\left(\left|A\left(\underline{d}^0\right)\right|+1\right)$th constraint. Then, (4.11) may have an infinite number of solutions, where the solution with the $\left(\left|A\left(\underline{d}^0\right)\right|+1\right)$th Lagrange multiplier equal to zero corresponds to the true solution. On the other hand, if we found $\left|A\left(\underline{d}^0\right)\right|-1$ "active" constraints at $\underline{d}^0$, then (4.11) would have no solution. Obviously, (4.9) gives only a tentative solution to determine $A\left(\underline{d}\right)$, and it is very important to determine the true $A\left(\underline{d}\right)$ for the rest of this rule.

We now introduce the stopping rule that holds asymptotically. This rule uses the Delta method to show that under certain conditions, nonlinear statistics are approximately multivariate normally distributed. Furthermore, this rule uses the procedure in Kodde and Palm (1986) to test composite hypotheses such as the one in (4.11). Kodde and Palm (1986) uses Wald's statistic, which is shown to have a so-called chi-

bar-squared distribution. A detailed explanation of the test is given in Appendix 2. In the rest of this section, we summarize this asymptotic test.

To apply Kodde and Palm (1986), we obtain a slightly different form of (4.11), as follows. Suppose that the gradients $\underline{\beta}_{j',-0}$ for active constraints at $\underline{d}$ are linearly independent (i.e., the regularity condition holds). Let $\mathbf{\Gamma}$ be a $k \times |A(\underline{d})|$ matrix with the $\underline{\beta}_{j',-0}$ for $j' \in A(\underline{d})$ as columns. Then, $\underline{\varepsilon}$ is defined as

$$\underline{\varepsilon}\left(\underline{\beta}_{0,-0},\ \mathbf{\Gamma}\right) = \underline{\beta}_{0,-0} - \mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^T\underline{\beta}_{0,-0}. \tag{4.12}$$

Under $H_0^{(2)}$ in (4.11), the residual component, say $\underline{\varepsilon}$, of $\underline{\beta}_{0,-0}$ that lies in the null space of the linear space generated by those $\underline{\beta}_{j',-0}$ corresponding to the active constraints has to be a zero-vector. Notice that in (4.12), $\mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^T$ is the orthogonal projection matrix that projects $\underline{\beta}_{0,-0}$ onto the linear space generated by the gradients $\underline{\beta}_{j',-0}$ of the active constraints. Besides, $\underline{\varepsilon}$ can be rewritten as $\underline{\varepsilon} = \underline{\beta}_{0,-0} + \mathbf{\Gamma}\underline{\lambda}$, where $\underline{\lambda}$ is the vector of Lagrange multipliers $\lambda_{j'}$. Hence, from (4.12) we have

$$\underline{\lambda}\left(\underline{\beta}_{0,-0},\ \mathbf{\Gamma}\right) = -\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^T\underline{\beta}_{0,-0}. \tag{4.13}$$

Now, we replace (4.11) by

$$\begin{aligned} H_0^{(2)} &: \underline{\varepsilon}\left(\underline{\beta}_{0,-0},\ \mathbf{\Gamma}\right) = \underline{0} \text{ and } \underline{\lambda}\left(\underline{\beta}_{0,-0},\ \mathbf{\Gamma}\right) \geq \underline{0} \\ H_1^{(2)} &: \underline{\varepsilon}\left(\underline{\beta}_{0,-0},\ \mathbf{\Gamma}\right) \neq \underline{0} \text{ or } \underline{\lambda}\left(\underline{\beta}_{0,-0},\ \mathbf{\Gamma}\right) \ngeq \underline{0}. \end{aligned} \tag{4.14}$$

In practice, the unknown deterministic values such as $\underline{\varepsilon}$, $\underline{\lambda}$, $\mathbf{\Gamma}$, $\underline{\beta}_{0,-0}$, and $\mathbf{\Psi}$ are replaced by their consistent estimators $\widehat{\underline{\varepsilon}}_N$, $\widehat{\underline{\lambda}}_N$, $\widehat{\mathbf{\Gamma}}_N$, $\widehat{\underline{\beta}}^N_{0,-0}$, and $\widehat{\mathbf{\Psi}}_N$. Using the Delta method, we show in Appendix 2 that $\sqrt{N}\left(\widehat{\underline{\varepsilon}}_N^T - \underline{\varepsilon}^T,\ \widehat{\underline{\lambda}}_N^T - \underline{\lambda}^T\right)^T$ is approximately multivariate normally distributed with zero mean vector and variance-covariance matrix $\widehat{\mathbf{\Sigma}}$ that can be partitioned as

$$\widehat{\mathbf{\Sigma}} = \begin{bmatrix} \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\varepsilon}}_N} & \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\varepsilon}}_N,\widehat{\underline{\lambda}}_N} \\ \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\lambda}}_N,\widehat{\underline{\varepsilon}}_N} & \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\lambda}}_N} \end{bmatrix}. \tag{4.15}$$

In (4.15), these four components are defined as

$$\widehat{\mathbf{\Sigma}}_{\widehat{\underline{\varepsilon}}_N} = \left(\nabla\widehat{\underline{\varepsilon}}_N\right)^T\widehat{\mathbf{\Psi}}_N\left(\nabla\widehat{\underline{\varepsilon}}_N\right),\ \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\lambda}}_N} = \left(\nabla\widehat{\underline{\lambda}}_N\right)^T\widehat{\mathbf{\Psi}}_N\left(\nabla\widehat{\underline{\lambda}}_N\right)$$
$$\widehat{\mathbf{\Sigma}}_{\widehat{\underline{\varepsilon}}_N,\widehat{\underline{\lambda}}_N} = \left(\nabla\widehat{\underline{\varepsilon}}_N\right)^T\widehat{\mathbf{\Psi}}_N\left(\nabla\widehat{\underline{\lambda}}_N\right),\ \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\lambda}}_N,\widehat{\underline{\varepsilon}}_N} = \left(\nabla\widehat{\underline{\lambda}}_N\right)^T\widehat{\mathbf{\Psi}}_N\left(\nabla\widehat{\underline{\varepsilon}}_N\right)$$

where $\nabla\widehat{\underline{\varepsilon}}_N$ is a consistent estimator of the $(k \times |A'(\underline{d})|) \times k$ Jacobian matrix of $\underline{\varepsilon}$, $\nabla\widehat{\underline{\lambda}}_N$ is a consistent estimator of the $(k \times |A'(\underline{d})|) \times |A(\underline{d})|$ Jacobian matrix of $\underline{\lambda}$, and $\widehat{\mathbf{\Psi}}_N$

is the $(k \times |A'(\underline{d})|) \times (k \times |A'(\underline{d})|)$ matrix with $\widehat{\sigma}_{i',i'}\mathbf{C}\left(i' \in A'(\underline{d})\right)$ on the diagonal and $\widehat{\sigma}_{i',h'}\mathbf{C}\left(i' \neq h', h' \in A'(\underline{d})\right)$ on the off-diagonal, with $\mathbf{C}$ being the matrix $\left(1/N\left(\mathbf{X}^T\mathbf{X}\right)\right)^{-1}$ after deleting its first row and column ($A'(\underline{d})$ is the index set of all binding constraints at $\underline{d}$ plus the index of the objective function (i.e., $A'(\underline{d}) = A(\underline{d}) \cup \{0\}$)); Appendix 2 gives formulas for $\nabla\widehat{\underline{\varepsilon}}_N$ and $\nabla\widehat{\underline{\lambda}}_N$.

For simplicity of notation, we omit the dependencies on $\underline{\beta}_{0,-0}$, $\mathbf{\Gamma}$, $\widehat{\underline{\beta}}^N_{0,-0}$, and $\widehat{\mathbf{\Gamma}}_N$ in the rest of this section. Define $\underline{\gamma}_1 = \sqrt{N}\underline{\varepsilon}$, $\underline{\gamma}_2 = \sqrt{N}\underline{\lambda}$, $\widehat{\underline{\gamma}}^N_1 = \sqrt{N}\widehat{\underline{\varepsilon}}_N$, and $\widehat{\underline{\gamma}}^N_2 = \sqrt{N}\widehat{\underline{\lambda}}_N$. Rewriting (4.14), we obtain the null and alternative hypotheses

$$H_0^{(2)} : \underline{\gamma}_1 = \underline{0} \text{ and } \underline{\gamma}_2 \geq \underline{0},\ H_1^{(2)} : \underline{\gamma}_1 \neq \underline{0} \text{ or } \underline{\gamma}_2 \ngeq \underline{0}. \tag{4.16}$$

Let $\|.\|$ denote the distance function in the metric $\widetilde{\mathbf{\Sigma}}$, which is a realized value of $\widehat{\mathbf{\Sigma}}$ defined in (4.15); that is, $\|\underline{x}\| = \underline{x}^T\widetilde{\mathbf{\Sigma}}^{-1}\underline{x}$ denotes the distance from the origin of an arbitrary vector $\underline{x}$. Furthermore, let S_0 and S_1 denote the feasible spaces for $\underline{\gamma} = \left(\underline{\gamma}_1^T, \underline{\gamma}_2^T\right)^T$ under $H_0^{(2)}$ and $H_1^{(2)}$ in (4.16), respectively. Kodde and Palm (1986) uses Wald's test statistic, (say) $\widehat{W} = \widehat{W}_0 - \widehat{W}_1$, where under $H_0^{(2)}$ in (4.16)

$$\widetilde{W}_0 = \underset{\underline{\gamma}^N \in S_0}{\text{minimize}} \left\|\widetilde{\underline{\gamma}}^N - \underline{\gamma}^N\right\|, \tag{4.17}$$

and under $H_1^{(2)}$ in (4.16)

$$\widetilde{W}_1 = \underset{\underline{\gamma}^N \in S_1}{\text{minimize}} \left\|\widetilde{\underline{\gamma}}^N - \underline{\gamma}^N\right\| \tag{4.18}$$

with $\widetilde{\underline{\gamma}}^N = \left[\left(\widetilde{\underline{\gamma}}^N_1\right)^T, \left(\widetilde{\underline{\gamma}}^N_2\right)^T\right]^T$, $\widetilde{W}_0$, and $\widetilde{W}_1$ being the realized values of $\widehat{\underline{\gamma}}^N$, $\widehat{W}_0$, and $\widehat{W}_1$. $\widetilde{W}_0$ and $\widetilde{W}_1$ in (4.17) and (4.18) are the distances of $\widetilde{\underline{\gamma}}^N$ to its orthogonal projections onto S_0 and S_1, respectively. Then, Kodde and Palm (1986) gives the statistic $\widehat{W}$ as

$$\begin{aligned}\widehat{W} = &\left(\widehat{\underline{\gamma}}^N_1\right)^T \widehat{\mathbf{\Sigma}}^{-1}_{\widehat{\underline{\varepsilon}}_N}\left(\widehat{\underline{\gamma}}^N_1\right) + \left(\widehat{\underline{\gamma}}^N_2 - \overline{\underline{\gamma}}^N_2 - \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\lambda}}_N,\widehat{\underline{\varepsilon}}_N}\widehat{\mathbf{\Sigma}}^{-1}_{\widehat{\underline{\varepsilon}}_N}\widehat{\underline{\gamma}}^N_1\right)^T \\ &\left(\widehat{\mathbf{\Sigma}}_{\widehat{\underline{\lambda}}_N} - \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\lambda}}_N,\widehat{\underline{\varepsilon}}_N}\widehat{\mathbf{\Sigma}}^{-1}_{\widehat{\underline{\varepsilon}}_N}\widehat{\mathbf{\Sigma}}_{\widehat{\underline{\varepsilon}}_N,\widehat{\underline{\lambda}}_N}\right)^{-1}\left(\widehat{\underline{\gamma}}^N_2 - \overline{\underline{\gamma}}^N_2 - \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\lambda}}_N,\widehat{\underline{\varepsilon}}_N}\widehat{\mathbf{\Sigma}}^{-1}_{\widehat{\underline{\varepsilon}}_N}\widehat{\underline{\gamma}}^N_1\right)\end{aligned} \tag{4.19}$$

where $\overline{\underline{\gamma}}^N_2$ denotes the orthogonal projection of $\widetilde{\underline{\gamma}}^N_2$ onto S_0.

Kodde and Palm (1986) proves that the statistic $\widehat{W}$ in (4.19) has the following so-called chi-bar-squared distribution under $H_0^{(2)}$ in (4.16) (remember, k and $|A(\underline{d})|$ are the dimensions of $\underline{\varepsilon}$ in (4.12) and $\underline{\lambda}$ in (4.13), respectively):

$$\Pr\left\{\widehat{W} \geq u \middle| \widehat{\mathbf{\Sigma}}\right\} = \sum_{t=0}^{|A(\underline{d})|} w_t \Pr\left\{\chi^2_{k+|A(\underline{d})|-t} \geq u\right\} \tag{4.20}$$

where u is the critical value, $\chi^2_{k+|A(\underline{d})|-t}$ is the central chi-squared random variable with $k+|A(\underline{d})|-t$ degrees of freedom, and w_t is the weight denoting the probability that t of $|A(\underline{d})|$ components of $\underline{\widetilde{\gamma}}_2^N$ are strictly positive. In (4.20), the weights w_t are such that $w_t \geq 0$ for each t, and $w_0 + ... + w_{|A(\underline{d})|} = 1$; see Kodde and Palm (1986). We now comment on these weights.

In general, the computation of the weights w_t can be very complicated; see Shapiro (1988). Kodde and Palm (1986) provides lower and upper bound critical values, say u_1 and u_2 respectively, for which the computation of w_t is not necessary; we give the formulas for u_1 and u_2 in Appendix 2. The problem with Kodde and Palm (1986)'s approach is that the test is inconclusive whenever $u_1 < \widetilde{W} < u_2$, where $\widetilde{W}$ is a realization of $\widehat{W}$. In such a case, the weights w_t in (4.20) must be computed.

Notice that the weights w_t in (4.20) are functions of the following arguments: (i) $|A(\underline{d})|$, (ii) the conditional covariance matrix of $\underline{\widehat{\gamma}}_2^N$ given $\underline{\widehat{\gamma}}_1^N$, and (iii) the nonnegative orthant $R_+^{|A(\underline{d})|} = \left\{\underline{\gamma}_2^N : \underline{\gamma}_2^N \geq \underline{0}\right\}$; i.e., $w_t = w_t\left(|A(\underline{d})|,\ \widehat{\mathbf{\Sigma}}_{\underline{\widehat{\lambda}}_N} - \widehat{\mathbf{\Sigma}}_{\underline{\widehat{\lambda}}_N,\underline{\widehat{\varepsilon}}_N}\widehat{\mathbf{\Sigma}}^{-1}_{\underline{\widehat{\varepsilon}}_N}\widehat{\mathbf{\Sigma}}_{\underline{\widehat{\varepsilon}}_N,\underline{\widehat{\lambda}}_N},\ R_+^{|A(\underline{d})|}\right)$. When the covariance matrix is equal to $\mathbf{I}$, Gouriéroux, Holly, and Monfort (1982) provides the following easy formula:

$$w_t\left(|A(\underline{d})|,\ \mathbf{I},\ R_+^{|A(\underline{d})|}\right) = \binom{|A(\underline{d})|}{t} 2^{-|A(\underline{d})|} \quad (t = 0,\ ...,\ |A(\underline{d})|)\,.$$

However, since the conditional covariance matrix may not be $\mathbf{I}$ for our case, we will compute the weights through a Monte Carlo sampling. Once the weights are computed, the critical value u can be obtained through (4.20), and $H_0^{(2)}$ is rejected if $\widetilde{W} > u$.

Now, we explain how we can compute the weights w_t in (4.20), as follows. Let $\underline{\widetilde{\lambda}}_N$ be a vector of $|A(\underline{d})|$ estimates of $\underline{\widehat{\lambda}}_N$ obtained through (4.13); let $\widetilde{\mathbf{\Sigma}}_{\underline{\widehat{\lambda}}_N}$, $\widetilde{\mathbf{\Sigma}}_{\underline{\widehat{\lambda}}_N,\underline{\widehat{\varepsilon}}_N}$, $\widetilde{\mathbf{\Sigma}}_{\underline{\widehat{\varepsilon}}_N}$, and $\widetilde{\mathbf{\Sigma}}_{\underline{\widehat{\varepsilon}}_N,\underline{\widehat{\lambda}}_N}$ be estimates of $\widehat{\mathbf{\Sigma}}_{\underline{\widehat{\lambda}}_N}$, $\widehat{\mathbf{\Sigma}}_{\underline{\widehat{\lambda}}_N,\underline{\widehat{\varepsilon}}_N}$, $\widehat{\mathbf{\Sigma}}_{\underline{\widehat{\varepsilon}}_N}$, and $\widehat{\mathbf{\Sigma}}_{\underline{\widehat{\varepsilon}}_N,\underline{\widehat{\lambda}}_N}$ obtained through the formulas that were defined below (4.15). We sample a large number, say $K = 1000$, of $\sqrt{N}\underline{\widehat{\lambda}}_N = \underline{\widehat{\gamma}}_2^N$ from the multivariate normal distribution with mean vector $\sqrt{N}\underline{\widetilde{\lambda}}_N = \underline{\widetilde{\gamma}}_2^N$ and covariance matrix $\widetilde{\mathbf{\Sigma}}_{\underline{\widehat{\lambda}}_N} - \widetilde{\mathbf{\Sigma}}_{\underline{\widehat{\lambda}}_N,\underline{\widehat{\varepsilon}}_N}\widetilde{\mathbf{\Sigma}}^{-1}_{\underline{\widehat{\varepsilon}}_N}\widetilde{\mathbf{\Sigma}}_{\underline{\widehat{\varepsilon}}_N,\underline{\widehat{\lambda}}_N}$. Next, the number of times (out of 1000) at which the sampled $\underline{\widehat{\gamma}}_2^N$ have no positive components (say, k_0 times), a single positive component (say, k_1 times), ..., and $|A(\underline{d})|$ positive components (say, $k_{|A(\underline{d})|}$ times) are determined. Now, the fractions $k_0/1000$, $k_1/1000$, ..., $k_{|A(\underline{d})|}/1000$ give the weights w_0, w_1, ..., $w_{|A(\underline{d})|}$, respectively.

A special case of (4.16) occurs if the number of binding constraints at $\underline{d}_0$ is equal to the number of input variables (i.e., $|A(\underline{d}_0)| = k$), since then $\underline{\beta}_{0,-0}$ lies in the linear space generated by the $\underline{\beta}_{j,-0}$ (i.e., the columns of $\mathbf{\Gamma}$); that is, the residual component $\underline{\varepsilon}$ of $\underline{\beta}_{0,-0}$, defined in (4.12), becomes a zero-vector. This is obvious from (4.12) by

considering $\mathbf{\Gamma}$ as a square (i.e., $k \times k$), nonsingular (i.e., the regularity condition) matrix. Hence, we may replace (4.16) by

$$H_0^{(2)} : \underline{\gamma}_2 \geq \underline{0},\ H_1^{(2)} : \underline{\gamma}_2 \ngeq \underline{0}.$$

To test this simplified $H_0^{(2)}$, we can still use Kodde and Palm (1986)'s procedure. The resulting Wald's test, first defined in (4.20), reduces to

$$\Pr\left\{\widehat{W} \geq u |\ \widehat{\mathbf{\Sigma}}\right\} = \sum_{t=0}^{|A(\underline{d})|} w_t \Pr\left\{\chi^2_{|A(\underline{d})|-t} \geq u\right\}$$

where χ^2_0 corresponds to the origin $\underline{0}$.

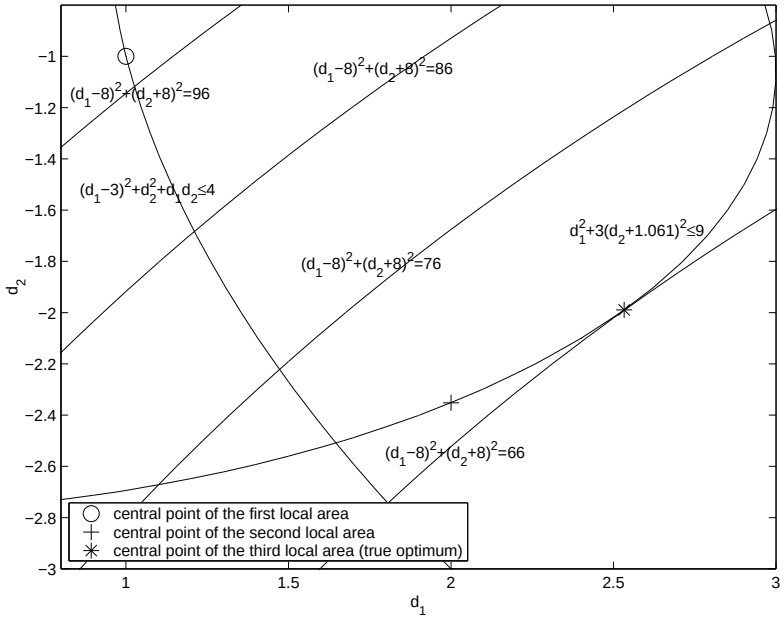
Now, we explain two simplifications for the test procedure that will result in a chi-square test statistic. The first simplification is to have only equality constraints in (4.1) so that there are no sign restrictions for $\lambda_{j'}$; see, for example, Gill, Murray, and Wright (2000, p. 81). The second simplification is to assume strict complementarity for $\lambda_{j'}$ (i.e., $E\left[s_{j'}(\underline{d},\ \underline{\omega})\right]\lambda_{j'} = 0$ implies $\lambda_{j'} > 0$ for each binding constraint j'); see, for example, Shapiro and Homem-de-Mello (1998). Obviously, for any one of these simplifications, we may replace the composite hypothesis in (4.11) by

$$\begin{aligned} H_0 &: \underline{\varepsilon}\left(\underline{\beta}_{0,-0},\ \mathbf{\Gamma}\right) = \underline{0} \\ H_1 &: \underline{\varepsilon}\left(\underline{\beta}_{0,-0},\ \mathbf{\Gamma}\right) \neq \underline{0}. \end{aligned} \tag{4.21}$$

Furthermore, the Delta method shows that $\sqrt{N}\left(\underline{\widehat{\varepsilon}}_N^T - \underline{\varepsilon}^T\right)^T$ is approximately multivariate normally distributed with zero mean vector and covariance matrix $\widehat{\mathbf{\Sigma}}_{\underline{\widehat{\varepsilon}}_N}$. Then, under the null hypothesis in (4.21), the statistic $N\underline{\widehat{\varepsilon}}_N^T \widehat{\mathbf{\Sigma}}_{\underline{\widehat{\varepsilon}}_N}^{-1} \underline{\widehat{\varepsilon}}_N$ has approximately a central chi-squared distribution with k degrees of freedom; see, for example, Muirhead (1982, p. 26). An alternative test statistic is $\underline{\widehat{\varepsilon}}_N^T \underline{\widehat{\varepsilon}}_N$; numerically, $\underline{\widehat{\varepsilon}}_N^T \underline{\widehat{\varepsilon}}_N$ is more convenient since it does not involve the inversion of the matrix $\widehat{\mathbf{\Sigma}}_{\underline{\widehat{\varepsilon}}_N}$, which may be ill-conditioned. To test the statistic $\underline{\widehat{\varepsilon}}_N^T \underline{\widehat{\varepsilon}}_N$, we refer to Shapiro and Homem-de-Mello (1998), and Mathai and Provost (1992, p. 164).

As we have already mentioned in the discussion of (4.10), the point $\underline{d}$ at which $H_0^{(2)}$ in (4.16) or (4.21) is not rejected will be considered as an approximation to the local minimizer $\underline{d}^0$. On the other hand, if $H_0^{(2)}$ is rejected, the realizations of the OLS estimators in (4.3) can be used to obtain an appropriate search direction and a step size; see, for example, Angün et al. (2003).

Figure 4.1: Location of the three central points and level curves of expected objective



4.5 Numerical examples

Now, we perform Monte Carlo experiments to estimate the type I and type II errors of the test in (4.20). Notice that Monte Carlo experiments enable us to control the intrinsic noise of the simulation. More importantly, since we have an asymptotic case, they are computationally very efficient compared to a time-consuming simulation study, which may actually take weeks.

We consider the following simple example problem:

$$\begin{array}{ll} \text{minimize} & E[(d_1-8)^2+(d_2+8)^2+\epsilon_0] \\ \text{subject to} & E[(d_1-3)^2+d_2^2+d_1d_2+\epsilon_1] \le 4 \\ & E[d_1^2+3\,(d_2+1.061)^2+\epsilon_2] \le 9 \end{array} \tag{4.22}$$

where ϵ_0, ϵ_1, and ϵ_2 are assumed to be multivariate normally distributed with mean vector $\underline{0}$. For $(\epsilon_0,\ \epsilon_1,\ \epsilon_2)^T$, we use different covariance matrices that will be explained later in this section. Obviously, the unconstrained minimum occurs at $(d_1,\ d_2)^T = (8,\ -8)^T$. Figure 4.1 shows that the constrained minimum occurs at approximately $(2.5328,\ -1.9892)^T$; substitution into (4.22) gives a mean objective value of 66.02 approximately.

In our Monte Carlo experiments, we use a 2^2 full-factorial design augmented with a central point. Although 2^2 is not saturated (see the discussion of (4.4)), we add

the central point since we test for lack of fit at that point. We choose three different local areas with central points expressed in the original variables as $(2.5328, \ -1.9892)^T$, $(1, \ -1)^T$, and $(2, \ -2.352)^T$; see points $\circ$, $+$, and $*$ in Figure 4.1. We find the other four design points by (arbitrarily) changing the coordinates of the central points by 0.5% (i.e., a total change of 1% in the local area size) and 0.25% (i.e., a total change of 0.5% in the local area size); in Figure 4.1, 2^2 points are situated very close to the central points. Changing by only 0.5% gives a very small signal; however, the noise is also very small in asymptotic situations.

We use two different covariance matrices of $(\epsilon_0, \ \epsilon_1, \ \epsilon_2)^T$ for each local area. For the first local area with central point $(2.5328, \ -1.9892)^T$, we select $\sigma_{0,0} = 3$, $\sigma_{1,1} = 1$, and $\sigma_{2,2} = 1.5$ (low noise), and $\sigma_{0,0} = 6$, $\sigma_{1,1} = 2$, and $\sigma_{2,2} = 3$ (high noise). For the second local area around $(1, \ -1)^T$, we select $\sigma_{0,0} = 0.25$, $\sigma_{1,1} = 0.1$, and $\sigma_{2,2} = 0.05$ (low noise), and $\sigma_{0,0} = 0.5$, $\sigma_{1,1} = 0.2$, and $\sigma_{2,2} = 0.1$ (high noise). For the third local area around $(2, \ -2.352)^T$, we select $\sigma_{0,0} = 0.6$, $\sigma_{1,1} = 0.35$, and $\sigma_{2,2} = 0.45$ (low noise), and $\sigma_{0,0} = 1.2$, $\sigma_{1,1} = 0.7$, and $\sigma_{2,2} = 0.9$ (high noise). For all three local areas, we let ρ denote the correlation between the responses, and select $\rho_{0,1} = -0.3$, $\rho_{0,2} = 0.4$, and $\rho_{1,2} = 0.6$ (for low noise case), and $\rho_{0,1} = -0.2$, $\rho_{0,2} = 0.7$, and $\rho_{1,2} = -0.8$ (for high noise case). We select these specific values for variances since they gave acceptable results for signal/noise ratios.

Depending on the type of noise (low or high), we select different number of replicates, namely 100 for low noise and 250 for high noise, for all three local areas. For problems where the null hypothesis is expressed as in (4.21) and Wald's statistic is used, Shieh (2003) provides a formula for the sample size as a function of the user-defined significance level and power.

Now, we summarize our results for the local area with the central point corresponding to the true optimum $(2.5328, \ -1.9892)^T$, where we change the local area size by 0.5% and make 100 replicates per input combination. Changing the coordinates of $(2.5328, \ -1.9892)^T$ by 0.25%, we obtain the other four design points, namely $(2.5391, \ -1.9842)^T$, $(2.5391, \ -1.9942)^T$, $(2.5265, \ -1.9842)^T$, $(2.5265, \ -1.9942)^T$. Given this input/output data $\left(\underline{d}, \ \widetilde{\underline{F}}_i\right)$, where $\underline{\widetilde{F}}_i$ is a vector of estimates of $\underline{\widehat{F}}_i$ in (4.3), we fit linear regression metamodels; see (4.2) and (4.3). Next, we test these metamodels for lack-of-fit using Roy's test in (4.7) (given 100 replicates, the classic F test with Bonferroni's inequality is not needed). If there is no lack-of-fit, then we test for binding constraints at the central point $(2.5328, \ -1.9892)^T$ using the t statistic in (4.9). If there is at least one binding constraint, then we test for the first-order necessary optimality conditions at the central point using Wald's test in (4.20). Note that if

the true optimum occurs inside the feasible area, then (4.20) does not apply; in such a case, the classic t test can be used to test whether the objective gradient is zero. Therefore, we assume that the true optimum occurs on the boundary; see also Figure 4.1. We repeat the procedure described above for 100 macro-replicates. Notice that if there is lack-of-fit or there is no binding constraint, then we proceed with the next macro-replicate.

For simplicity, we use the same significance level $\alpha = 10\%$ for all three statistical tests, namely for lack-of-fit, binding constraints, and first-order necessary optimality conditions. In Table 4.1, we present the results of the 100 macro-replicates at the local area around the true optimal point, $(2.5328, \ -1.9892)^T$. We examine different local area sizes (0.5% and 1%), noise types (low and high), and number of replicates (100 and 250). In Table 4.1, the number 93 in the cell corresponding with the second row and column represents the number of macro-replicates (out of 100 macro-replicates) at which we have no lack of fit for the first-order regression metamodels through Roy's test in (4.7). 93/100 is close to the expected value, $1-\alpha = 0.9$. The second constraint in (4.22) is active; see Figure 4.1. Hence, the number 84 in the second row and third column represents the number of macro-replicates (out of 93 macro-replicates) at which we detect $d_1^2 + 3\,(d_2 + 1.061)^2 \leq 9$ as the only active constraint at $(2.5328, \ -1.9892)^T$ through the t test in (4.9). 84/93 is again close to $1-\alpha = 0.9$. Furthermore, the number 13 in the second row and fourth column represents the number of macro-replicates (out of 84 macro-replicates) at which the null hypothesis $H_0^{(2)}$ in (4.16) is rejected through Wald's test in (4.20). Notice that when testing for the first-order necessary optimality conditions, we have better results for smaller local areas (3 versus 13 out of 84); the expected quadratic functions in (4.22) can be better approximated by linear functions in smaller local areas.

We also obtain numerical results at the other two local areas. At $(1, \ -1)^T$, the binding constraint is $(d_1-3)^2+d_2^2+d_1d_2 \leq 4$, whereas at $(2, \ -2.352)^T$, the binding constraint is $d_1^2 + 3\,(d_2 + 1.061)^2 \leq 9$. Note that $(1, \ -1)^T$ is further away from the true optimum $(2.5328, \ -1.9892)^T$ than $(2, \ -2.352)^T$ is; see Figure 4.1. Hence, we expect higher rejection rates for $H_0^{(2)}$ in (4.16) at $(1, \ -1)^T$; that is, the power of Wald's test in (4.20) should increase as we move away from the true optimum. The numerical values in Tables 4.2 and 4.3 illustrate this increase in power: $69/82 \simeq 0.8415$ versus $56/84 \simeq 0.6667$, $60/85 \simeq 0.7059$ versus $47/89 \simeq 0.5281$ etc., where $\simeq$ means approximately equal.

Table 4.1: Results for 100 macro-replicates at the local area where the true optimal point $(2.5328,\ -1.9892)^T$ is the central point

(local area size, noise, replicates)	no lack of fit	$d_1^2 + 3\left(d_2 + 1.061\right)^2 \leq 9$ active	H_0^2 rejected
(0.5%, low noise, 100)	93	84	3
(1%, low noise, 100)	93	84	13
(0.5%, high noise, 250)	92	89	10
(1%, high noise, 250)	92	89	15

Table 4.2: Results for 100 macro-replicates at the worst local area around $(1,\ -1)^T$

(local area size, noise, replicates)	no lack of fit	$(d_1 - 3)^2 + d_2^2 + d_1 d_2 \leq 4$ active	H_0^2 rejected
(0.5%, low noise, 100)	91	82	69
(1%, low noise, 100)	91	82	82
(0.5%, high noise, 250)	92	85	60
(1%, high noise, 250)	92	85	79

Table 4.3: Results for 100 macro-replicates at the better local area around $(2, -2.352)^T$

(local area size, noise, replicates)	no lack of fit	$d_1^2 + 3\left(d_2 + 1.061\right)^2 \leq 9$ active	H_0^2 rejected
(0.5%, low noise, 100)	93	84	56
(1%, low noise, 100)	93	84	78
(0.5%, high noise, 250)	92	89	47
(1%, high noise, 250)	92	89	79

4.6 Conclusions

In this paper, we derive an asymptotic, statistical stopping rule that tests the first-order necessary optimality conditions at a feasible point for black box simulation optimization. The derivation of this stopping rule uses the Delta method and Wald's statistic. We estimate the performance of this rule through a simple Monte Carlo example. The numerical results of this example are encouraging; that is, the null hypothesis (i.e., the first-order necessary optimality conditions hold) is "accepted" at the true optimum very often. Furthermore, the estimated power increases as the points move away from the true optimum.

Moreover, we introduce two alternative lack-of-fit tests, namely Roy's largest root test and the classic F test combined with Bonferroni's inequality, which are only relevant to the RSM literature. Further, we show that Roy's test is a direct generalization of the classic F test to multiple responses. Note that Roy's test should be preferred if the condition on the number of replicates, which was mentioned in this paper, is satisfied. If this condition is not satisfied, then the classic F test with Bonferroni's inequality is a better alternative. In an asymptotic case, Roy's test is preferred.

4.7 Appendix 1: Roy's largest root test

In this appendix, we present the details of Roy's largest root test, which was summarized in §4.3.1; see Roy, Gnanadesikan, and Srivastava (1971), and Khuri (1985). We further show that Roy's test is a generalization of the ordinary lack-of-fit F test in classic RSM to multiple responses.

Let n with $n \geq k+1$ be the number of distinct simulated input combinations per local area. Further, let m_l be the number of replicates at the lth ($l = 1,\ ...,\ n$) input combination. Suppose that we obtain at least two replicates at q ($1 \leq q \leq n$) distinct input combinations. Without loss of generality, suppose that these q input combinations are the first q design points: i.e., m_1 replicates for the first input combination, ..., m_q replicates for the qth input combination, $m_{q+1} = ... = m_n = 1$. Roy, Gnanadesikan, and Srivastava (1971, p. 35) shows that the total number of replicates at these q input combinations has to satisfy the following condition: $(m_1 - 1) + ... + (m_q - 1) \geq r$, where r is the number of response types. The total number of local runs is $N = m_1 + ... + m_n$.

Let $\underline{b} = (b_0,\ ...,\ b_{r-1})^T$ be an arbitrary nonzero weight vector. Then, to test the joint hypotheses H_0 in (4.5), we consider the weighted sums of all responses

$$b_0 \underline{G}_0(\underline{d},\ \underline{\omega}) + ... + b_{r-1}\underline{G}_{r-1}(\underline{d},\ \underline{\omega})$$

for all nonzero $\underline{b}$. In the following, we describe the components of the lack-of-fit test - namely, the sum of squared residuals (denoted by SS_R), the sum of squared pure errors (denoted by SS_{PE}), and the sum of squared lack of fit (denoted by SS_{LOF}) - for a single response. Further, we generalize these components to multiple responses.

Notice that $\widehat{\underline{F}}_i$, $\widehat{\underline{G}}_i$, $\widehat{\underline{\beta}}_i$, and $\mathbf{X}$ were already defined in §4.2. The classic F test for a single response i uses

$$\begin{aligned} SS_R &= \left(\widehat{\underline{F}}_i - \widehat{\underline{G}}_i\right)^T \left(\widehat{\underline{F}}_i - \widehat{\underline{G}}_i\right) \\ &= \left(\widehat{\underline{F}}_i - \mathbf{X}\widehat{\underline{\beta}}_i\right)^T \left(\widehat{\underline{F}}_i - \mathbf{X}\widehat{\underline{\beta}}_i\right) \\ &= \left(\widehat{\underline{F}}_i - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\widehat{\underline{F}}_i\right)^T \left(\widehat{\underline{F}}_i - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\widehat{\underline{F}}_i\right) \\ &= \widehat{\underline{F}}_i^T \left(\mathbf{I}_N - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\widehat{\underline{F}}_i. \end{aligned} \tag{4.23}$$

The last line in (4.23) is obtained using the fact that $\left(\mathbf{I}_N - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)$ is a projection matrix, say, $\mathbf{P}$ with the following properties: $\mathbf{PP} = \mathbf{P}$ and $\mathbf{P} = \mathbf{P}^T$.

We now explain the SS_R for multiple responses. Let $\mathbf{Z}$ be an $N \times (r(k+1))$ matrix consisting of r identical $\mathbf{X}$ matrices; further, $\widehat{\mathbf{F}} = \left[\widehat{\underline{F}}_0,\ ...,\ \widehat{\underline{F}}_{r-1}\right]$ and $\widehat{\mathbf{G}} = \left[\widehat{\underline{G}}_0,\ ...,\ \widehat{\underline{G}}_{r-1}\right]$

are $N \times r$ matrices, and $\mathbf{B}$ is an $(r\,(k+1) \times r)$ block-diagonal matrix with $\underline{\widehat{\beta}}_i$ on the diagonal; i.e., $\mathbf{B} = diag\left(\underline{\widehat{\beta}}_0, ..., \underline{\widehat{\beta}}_{r-1}\right)$. Analogous to (4.23), SS_R with the weight vector $\underline{b}$ can be written as

$$\begin{aligned} SS_R\left(\underline{b}\right) &= \underline{b}^T \left(\widehat{\mathbf{F}} - \widehat{\mathbf{G}}\right)^T \left(\widehat{\mathbf{F}} - \widehat{\mathbf{G}}\right) \underline{b} \\ &= \underline{b}^T \left(\widehat{\mathbf{F}} - \mathbf{ZB}\right)^T \left(\widehat{\mathbf{F}} - \mathbf{ZB}\right) \underline{b} \\ &= \underline{b}^T \left(\widehat{\mathbf{F}} - \mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)^- \mathbf{Z}^T\widehat{\mathbf{F}}\right)^T \left(\widehat{\mathbf{F}} - \mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)^- \mathbf{Z}^T\widehat{\mathbf{F}}\right) \underline{b} \\ &= \underline{b}^T\widehat{\mathbf{F}}^T \left[\mathbf{I}_N - \mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)^- \mathbf{Z}^T\right] \widehat{\mathbf{F}}\underline{b} \end{aligned} \tag{4.24}$$

where $\left(\mathbf{Z}^T\mathbf{Z}\right)^-$ denotes a generalized inverse of $\mathbf{Z}^T\mathbf{Z}$.

Now, we derive the pure error SS_{PE}, in case of a single response i. Let $\mathbf{K}$ be the $N \times N$ block-diagonal matrix with $\mathbf{K}_p = \mathbf{I}_{m_p} - (1/m_p)\,\mathbf{1}_{m_p}$ $(p = 1, ..., q)$ on the diagonal; i.e., $\mathbf{K} = diag\left(\mathbf{K}_1,\ \mathbf{K}_2,\ ...,\ \mathbf{K}_q,\ \mathbf{0}_{N-m_1-...-m_q}\right)$. Notice that $\mathbf{K}$ is a projection matrix. Let $\overline{\widehat{F}}_i$ be the $N \times 1$ vector of response i averaged over the m_l replicates; i.e., its first m_1 rows consist of $\overline{\widehat{F}}_{i,m_1}$, where $\overline{\widehat{F}}_{i,m_1}$ corresponds to the average of the first m_1 components of $\underline{\widehat{F}}_i$, its next m_2 rows consist of $\overline{\widehat{F}}_{i,m_2}$, where $\overline{\widehat{F}}_{i,m_2}$ corresponds to the average of the next m_2 components of $\underline{\widehat{F}}_i$, etc. This gives

$$\begin{aligned} SS_{PE} &= \left(\underline{\widehat{F}}_i - \overline{\underline{\widehat{F}}}_i\right)^T \left(\underline{\widehat{F}}_i - \overline{\underline{\widehat{F}}}_i\right) \\ &= \left(\mathbf{K}\underline{\widehat{F}}_i\right)^T \left(\mathbf{K}\underline{\widehat{F}}_i\right) \\ &= \underline{\widehat{F}}_i^T \mathbf{K}\underline{\widehat{F}}_i. \end{aligned} \tag{4.25}$$

Let $\overline{\widehat{\mathbf{F}}}$ be the $N \times r$ matrix with $\overline{\underline{\widehat{F}}}_i$ as its columns. Similar to (4.25), SS_{PE} for multiple responses is

$$SS_{PE}\left(\underline{b}\right) = \underline{b}^T\widehat{\mathbf{F}}^T\mathbf{K}\widehat{\mathbf{F}}\underline{b} = \underline{b}^T \left(\mathbf{K}\widehat{\mathbf{F}}\right)^T \left(\mathbf{K}\widehat{\mathbf{F}}\right) \underline{b} = \underline{b}^T \left(\widehat{\mathbf{F}} - \overline{\widehat{\mathbf{F}}}\right)^T \left(\widehat{\mathbf{F}} - \overline{\widehat{\mathbf{F}}}\right) \underline{b}. \tag{4.26}$$

The diagonal of the $r \times r$ matrix $\left(\widehat{\mathbf{F}} - \overline{\widehat{\mathbf{F}}}\right)^T \left(\widehat{\mathbf{F}} - \overline{\widehat{\mathbf{F}}}\right)$ consists of the r sum of squared pure errors corresponding to the r responses; the off-diagonal has estimators for the covariances between the responses of the same replicate.

Now (4.24) and (4.26) give the lack of fit (LOF) error:

$$\begin{aligned} SS_{LOF}\left(\underline{b}\right) &= SS_R\left(\underline{b}\right) - SS_{PE}\left(\underline{b}\right) \\ &= \underline{b}^T\widehat{\mathbf{F}}^T \left[\mathbf{I}_N - \mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)^- \mathbf{Z}^T - \mathbf{K}\right] \widehat{\mathbf{F}}\underline{b}. \end{aligned}$$

If H_0 in (4.5), the independence of the columns of $\widehat{\mathbf{F}}$ (see the discussion on page 3 about the non-overlapping PRN streams and common random numbers), and the multivariate normality assumption for the simulation responses $\widehat{\mathbf{F}}$ hold, then the statistic in (4.27) has the F distribution with ν_1 and ν_2 degrees of freedom:

$$F_{\nu_1,\nu_2}\left(\underline{b}\right) = \frac{SS_{LOF}\left(\underline{b}\right)/\nu_1}{SS_{PE}\left(\underline{b}\right)/\nu_2} \tag{4.27}$$

where $\nu_1 = n - (k+1)$ and $\nu_2 = N - n$ are the degrees of freedom for $SS_{LOF}\left(\underline{b}\right)$ and $SS_{PE}\left(\underline{b}\right)$, respectively. Further, Roy, Gnanadesikan, and Srivastava (1971) defines $r \times r$ matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ as follows:

$$\mathbf{Q}_1 = \widehat{\mathbf{F}}^T\left[\mathbf{I}_N - \mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-}\mathbf{Z}^T - \mathbf{K}\right]\widehat{\mathbf{F}} \text{ and } \mathbf{Q}_2 = \widehat{\mathbf{F}}^T\mathbf{K}\widehat{\mathbf{F}}$$

so that $F_{\nu_1,\nu_2}\left(\underline{b}\right)$ in (4.27) can be rewritten as

$$F_{\nu_1,\nu_2}\left(\underline{b}\right) = \frac{\underline{b}^T\mathbf{Q}_1\underline{b}/\nu_1}{\underline{b}^T\mathbf{Q}_2\underline{b}/\nu_2}.$$

H_0 in (4.5) is rejected if $\underset{\underline{b}\neq\underline{0}}{max}\left(\underline{b}^T\mathbf{Q}_1\underline{b}/\underline{b}^T\mathbf{Q}_2\underline{b}\right)$ exceeds a prespecified critical value, F_{α,ν_1,ν_2}, where α is the significance level.

Roy, Gnanadesikan, and Srivastava (1971, p. 35) shows that when the condition on the number of replicates is satisfied (that is, $(m_1 - 1) + ... + (m_q - 1) \geq r$), then $\mathbf{Q}_2$ is positive definite. So, $\underline{b}$ can be defined as $\underline{b} = \mathbf{Q}_2^{-1/2}\underline{z}$. Then,

$$\underset{\underline{b}\neq\underline{0}}{max}\left(\frac{\underline{b}^T\mathbf{Q}_1\underline{b}}{\underline{b}^T\mathbf{Q}_2\underline{b}}\right) = \underset{\underline{z}\neq\underline{0}}{max}\left(\frac{\underline{z}^T\mathbf{Q}_2^{-1/2}\mathbf{Q}_1\mathbf{Q}_2^{-1/2}\underline{z}}{\underline{z}^T\underline{z}}\right). \tag{4.28}$$

Note that in (4.28), $\left(\underline{z}^T\mathbf{Q}_2^{-1/2}\mathbf{Q}_1\mathbf{Q}_2^{-1/2}\underline{z}/\underline{z}^T\underline{z}\right)$ is known as Rayleigh quotient in linear algebra. Furthermore, the maximization problem on the right-hand-side of the equality in (4.28) is a standard problem in linear algebra for finding the largest eigenvalue (say) $\kappa_{\max}$ of the $r \times r$ matrix $\mathbf{Q}_2^{-1/2}\mathbf{Q}_1\mathbf{Q}_2^{-1/2}$. However, the matrices $\mathbf{Q}_2^{-1/2}\mathbf{Q}_1\mathbf{Q}_2^{-1/2}$ and $\mathbf{Q}_2^{-1}\mathbf{Q}_1$ have the same eigenvalues; see Magnus and Neudecker (1988, p. 13). Thus, H_0 in (4.5) is rejected if $\kappa_{\max}\left(\mathbf{Q}_2^{-1}\mathbf{Q}_1\right) \geq F_{\alpha,\nu_1,\nu_2}$.

4.8 Appendix 2: Derivation of the statistical stopping rule

In this appendix, we detail the procedure described in Kodde and Palm (1986) for testing composite hypotheses such as the one in (4.16). For completeness in the description

of its procedure, before introducing this procedure, we present two well-known properties of OLS estimators, which we need for the Delta method. Next, we introduce the (finite-dimensional) Delta method in §4.8.1 and Kodde and Palm (1986)'s procedure in §4.8.2.

Now, we state two well-known properties of OLS estimators that are required for the application of the Delta method. Suppose that the OLS assumptions (i.e., $\underline{\mu}_{\underline{\epsilon}_i} = \underline{0}$ and $\mathbf{\Sigma}_{\underline{\epsilon}_i} = \sigma_{i,i}\mathbf{I}$) hold per response type i. If $\left(1/N\left(\mathbf{X}^T\mathbf{X}\right)\right)$ converges to a positive definite matrix as $N \longrightarrow \infty$, then $\widehat{\underline{\beta}}_i^N$ converges in probability to the (unobservable) true value $\underline{\beta}_i$, denoted as $\widehat{\underline{\beta}}_i^N \xrightarrow{P} \underline{\beta}_i$; see Theil (1971, p. 362). In other words, $\widehat{\underline{\beta}}_i^N$ is a consistent estimator of $\underline{\beta}_i$. Furthermore, $\sqrt{N}\left(\widehat{\underline{\beta}}_i^N - \underline{\beta}_i\right)$ has a multivariate normal limiting distribution with zero mean vector and covariance matrix given by $\sigma_{i,i}\left(1/N\left(\mathbf{X}^T\mathbf{X}\right)\right)^{-1}$; see Theil (1971, p. 378).

Let $A'\left(\underline{d}\right)$ denote the index set of all binding constraints at $\underline{d}$ plus the index of the objective function; i.e., $A'\left(\underline{d}\right) = A\left(\underline{d}\right) \cup \{0\}$. Let $\underline{z}$ denote a $\left(k \times \left|A'\left(\underline{d}\right)\right|\right) \times 1$ vector with $\widehat{\underline{\beta}}_{i',-0}^N - \underline{\beta}_{i',-0}$ as rows for $i' \in A'\left(\underline{d}\right)$, where $\widehat{\underline{\beta}}_{i',-0}^N$ denotes $\widehat{\underline{\beta}}_{i'}^N$ excluding the intercept $\widehat{\beta}_{i',0}^N$. It is known from the previous paragraph that $\sqrt{N}\underline{z}$ has a multivariate normal limiting distribution with zero mean vector and covariance matrix, say, $\mathbf{\Psi}$. Notice that $\mathbf{\Psi}$ is a $\left(k \times \left|A'\left(\underline{d}\right)\right|\right) \times \left(k \times \left|A'\left(\underline{d}\right)\right|\right)$ matrix with $\sigma_{i',i'}\mathbf{C}$ on the diagonal and $\sigma_{i',h'}\mathbf{C}$ $\left(i' \neq h',\ h' \in A'\left(\underline{d}\right)\right)$ on the off-diagonal, where $\mathbf{C}$ is the matrix $\left(1/N\left(\mathbf{X}^T\mathbf{X}\right)\right)^{-1}$ after deleting its first row and column.

4.8.1 Delta method

Now, we introduce the (finite-dimensional) Delta method; see, for example, Rubinstein and Shapiro (1993, p. 259). Later in this appendix, the Delta method is used to show that under certain conditions, nonlinear test statistics are approximately multivariate normally distributed.

Let $\left\{\underline{\gamma}_N\right\}$ be a sequence of $m \times 1$ random vectors converging in probability to a vector $\underline{\mu}$. Let $\{\tau_N\}$ be a sequence of positive numbers tending to infinity such that $\tau_N\left(\underline{\gamma}_N - \underline{\mu}\right)$ converges in distribution to a random vector $\underline{\delta}$. Let $g : R^m \longrightarrow R^n$ be a mapping that is differentiable at $\underline{\mu}$. Then, $\tau_N\left(g\left(\underline{\gamma}_N\right) - g\left(\underline{\mu}\right)\right)$ has the same limiting distribution as $\left(\nabla g\left(\underline{\mu}\right)\right)^T \underline{\delta}$ where $\nabla g\left(\underline{\mu}\right)$ denotes the Jacobian matrix of g at $\underline{\mu}$.

Before applying the Delta method, it is necessary to show that $\underline{\varepsilon}$ and $\underline{\lambda}$ are differentiable with respect to $\underline{\beta}_{i',-0}$ for $i' \in A'\left(\underline{d}\right)$. Apparently, $\underline{\varepsilon}$ is everywhere partially

differentiable with respect to $\underline{\beta}_{0,-0}$:

$$\partial\underline{\varepsilon} = \left\{\mathbf{I}_k - \mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^T\right\}\partial\underline{\beta}_{0,-0} \tag{4.29}$$

where the $k \times k$ matrix within the curly brackets in (4.29) denotes the first partial derivative of $\underline{\varepsilon}$ with respect to $\underline{\beta}_{0,-0}$. To show that $\underline{\varepsilon}$ and $\underline{\lambda}$ are also partially differentiable with respect to $\underline{\beta}_{j',-0}$, we use the following theorem in Magnus and Neudecker (1988, p. 151).

Theorem: Let T be the set of non-singular real $m \times m$ matrices. Let S be an open subset of $R^{n\times q}$. If the matrix function $F : S \to T$ is k times (continuously) differentiable on S, then so is the matrix function $F^{-1} : T \to S$ defined by $F^{-1}(\mathbf{X}) = (F(\mathbf{X}))^{-1}$, and the differential is $dF^{-1} = -F^{-1}(dF)F^{-1}$.

Since $\underline{\beta}_{j',-0}$ for $j' \in A(\underline{d})$ are columns of the matrix $\mathbf{\Gamma}$, $\mathbf{\Gamma}$ is everywhere partially differentiable with respect to $\underline{\beta}_{j',-0}$. Furthermore, because of our assumption of the linearly independent binding constraints at $\underline{d}$, $\mathbf{\Gamma}^T\mathbf{\Gamma}$ is non-singular. Hence, using the above theorem, the partial differential of $\underline{\varepsilon}$ with respect to $\underline{\beta}_{j',-0}$ is given by

$$\begin{aligned}
\partial\underline{\varepsilon} &= -\partial\mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^T\underline{\beta}_{0,-0} + \mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}\partial\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^T\underline{\beta}_{0,-0} \\
&\quad - \mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}(\partial\mathbf{\Gamma})^T\underline{\beta}_{0,-0} \\
&= -\partial\mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^T\underline{\beta}_{0,-0} + \mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}\left[(\partial\mathbf{\Gamma})^T\mathbf{\Gamma} + \mathbf{\Gamma}^T\partial\mathbf{\Gamma}\right]\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^T\underline{\beta}_{0,-0} \\
&\quad - \mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}(\partial\mathbf{\Gamma})^T\underline{\beta}_{0,-0} \\
&= -\partial\mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^T\underline{\beta}_{0,-0} + \mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}(\partial\mathbf{\Gamma})^T\mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^T\underline{\beta}_{0,-0} \\
&\quad + \mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^T\partial\mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^T\underline{\beta}_{0,-0} - \mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{\Gamma}\right)^{-1}(\partial\mathbf{\Gamma})^T\underline{\beta}_{0,-0}.
\end{aligned} \tag{4.30}$$

Note that $\partial\mathbf{\Gamma} = \partial\underline{\beta}_{j',-0}\underline{e}_{j'}^T$, where $\underline{e}_{j'}$ is $|A(\underline{d})| \times 1$ vector of zeros, except a one in the j'th position. Replacing $\partial\mathbf{\Gamma}$ by $\partial\underline{\beta}_{j',-0}\underline{e}_{j'}^T$, we can rewrite (4.30) as follows:

$$
\begin{aligned}
\partial\underline{\varepsilon} &= -\left(\partial\underline{\beta}_{\acute{\jmath},-0}\underline{e}_{\acute{\jmath}}^{T}\right)\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0} \\
&+ \mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\left(\partial\underline{\beta}_{\acute{\jmath},-0}\underline{e}_{\acute{\jmath}}^{T}\right)^{T}\mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0} \\
&+ \mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\left(\partial\underline{\beta}_{\acute{\jmath},-0}\underline{e}_{\acute{\jmath}}^{T}\right)\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0} - \mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\left(\partial\underline{\beta}_{\acute{\jmath},-0}\underline{e}_{\acute{\jmath}}^{T}\right)^{T}\underline{\beta}_{0,-0} \\
&= -\partial\underline{\beta}_{\acute{\jmath},-0}\left[\underline{e}_{\acute{\jmath}}^{T}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0}\right] + \mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\underline{e}_{\acute{\jmath}}\left[\left(\partial\underline{\beta}_{\acute{\jmath},-0}\right)^{T}\mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0}\right] \\
&+ \mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\partial\underline{\beta}_{\acute{\jmath},-0}\left[\underline{e}_{\acute{\jmath}}^{T}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0}\right] - \mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\underline{e}_{\acute{\jmath}}\left[\left(\partial\underline{\beta}_{\acute{\jmath},-0}\right)^{T}\underline{\beta}_{0,-0}\right] \\
&= -\left[\underline{e}_{\acute{\jmath}}^{T}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0}\right]\partial\underline{\beta}_{\acute{\jmath},-0} + \mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\underline{e}_{\acute{\jmath}}^{T}\underline{\beta}_{0,-0}^{T}\mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\partial\underline{\beta}_{\acute{\jmath},-0} \\
&+ \left[\underline{e}_{\acute{\jmath}}^{T}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0}\right]\mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\partial\underline{\beta}_{\acute{\jmath},-0} - \mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\underline{e}_{\acute{\jmath}}\underline{\beta}_{0,-0}^{T}\partial\underline{\beta}_{\acute{\jmath},-0} \\
&= \left\{-\left[\underline{e}_{\acute{\jmath}}^{T}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0}\right]\left(\mathbf{I}_{k}-\mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\right)\right. \\
&\left.-\mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\underline{e}_{\acute{\jmath}}\underline{\beta}_{0,-0}^{T}\left(\mathbf{I}_{k}-\mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\right)\right\}\partial\underline{\beta}_{\acute{\jmath},-0}
\end{aligned}
\tag{4.31}
$$

where the expressions within the square brackets are scalars, and the $k \times k$ matrix within the curly brackets on the last two lines in (4.31) is the first partial derivative of $\underline{\varepsilon}$ with respect to $\underline{\beta}_{\acute{\jmath},-0}$.

Analogously, the partial differential of $\underline{\lambda}$ with respect to $\underline{\beta}_{0,-0}$ is given by

$$
\partial\underline{\lambda} = \left\{-\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\right\}\partial\underline{\beta}_{0,-0} \tag{4.32}
$$

where the $|A(\underline{d})| \times k$ matrix within the curly brackets in (4.32) denotes the first partial derivative of $\underline{\lambda}$ with respect to $\underline{\beta}_{0,-0}$. Furthermore, the partial differential of $\underline{\lambda}$ with respect to $\underline{\beta}_{\acute{\jmath},-0}$ is given by

$$
\begin{aligned}
\partial\underline{\lambda} &= \left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\partial\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0} - \left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\left(\partial\mathbf{\Gamma}\right)^{T}\underline{\beta}_{0,-0} \\
&= \left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\left[\left(\partial\mathbf{\Gamma}\right)^{T}\mathbf{\Gamma} + \mathbf{\Gamma}^{T}\partial\mathbf{\Gamma}\right]\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0} - \left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\left(\partial\mathbf{\Gamma}\right)^{T}\underline{\beta}_{0,-0} \\
&= \left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\left(\partial\mathbf{\Gamma}\right)^{T}\mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0} + \left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\partial\mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0} \\
&- \left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\left(\partial\mathbf{\Gamma}\right)^{T}\underline{\beta}_{0,-0}.
\end{aligned}
\tag{4.33}
$$

Replacing again $\partial\mathbf{\Gamma}$ by $\partial\underline{\beta}_{\acute{j},-0}\underline{e}_{\acute{j}}^{T}$, we rewrite (4.33) as follows:

$$\begin{aligned}
\partial\underline{\lambda} &= \left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\left(\partial\underline{\beta}_{\acute{j},-0}\underline{e}_{\acute{j}}^{T}\right)^{T}\mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0} \\
&+ \left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\left(\partial\underline{\beta}_{\acute{j},-0}\underline{e}_{\acute{j}}^{T}\right)\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0} - \left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\left(\partial\underline{\beta}_{\acute{j},-0}\underline{e}_{\acute{j}}^{T}\right)^{T}\underline{\beta}_{0,-0} \\
&= \left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\underline{e}_{\acute{j}}\left[\left(\partial\underline{\beta}_{\acute{j},-0}\right)^{T}\mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0}\right] \\
&+ \left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\partial\underline{\beta}_{\acute{j},-0}\left[\underline{e}_{\acute{j}}^{T}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0}\right] - \left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\underline{e}_{\acute{j}}\left[\left(\partial\underline{\beta}_{\acute{j},-0}\right)^{T}\underline{\beta}_{0,-0}\right] \\
&= \left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\underline{e}_{\acute{j}}^{T}\underline{\beta}_{0,-0}^{T}\mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\partial\underline{\beta}_{\acute{j},-0} \\
&+ \left[\underline{e}_{\acute{j}}^{T}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0}\right]\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\partial\underline{\beta}_{\acute{j},-0} - \left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\underline{e}_{\acute{j}}\underline{\beta}_{0,-0}^{T}\partial\underline{\beta}_{\acute{j},-0} \\
&= \left\{\left[\underline{e}_{\acute{j}}^{T}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\underline{\beta}_{0,-0}\right]\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T}\right. \\
&\left.+ \left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\underline{e}_{\acute{j}}\underline{\beta}_{0,-0}^{T}\left(\mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^{T} - \mathbf{I}_{k}\right)\right\}\partial\underline{\beta}_{\acute{j},-0}
\end{aligned} \tag{4.34}$$

where the expressions within the square brackets are scalars, and the $|A(\underline{d})| \times k$ matrix within the curly brackets on the last two lines in (4.34) is the first partial derivative of $\underline{\lambda}$ with respect to $\underline{\beta}_{\acute{j},-0}$.

In practice, the unknown deterministic values such as $\underline{\varepsilon}$, $\underline{\lambda}$, $\mathbf{\Gamma}$, $\underline{\beta}_{0,-0}$, and $\mathbf{\Psi}$ are replaced by their consistent estimators $\widehat{\underline{\varepsilon}}_N$, $\widehat{\underline{\lambda}}_N$, $\widehat{\mathbf{\Gamma}}_N$, $\widehat{\underline{\beta}}_{0,-0}^{N}$, and $\widehat{\mathbf{\Psi}}_N$. Now, the Delta method shows that $\sqrt{N}\left(\widehat{\underline{\varepsilon}}_N^{T} - \underline{\varepsilon}^{T}, \widehat{\underline{\lambda}}_N^{T} - \underline{\lambda}^{T}\right)^{T}$ is approximately multivariate normally distributed with zero mean vector and variance-covariance matrix $\widehat{\mathbf{\Sigma}}$ that can be partitioned as

$$\widehat{\mathbf{\Sigma}} = \begin{bmatrix} \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\varepsilon}}_N} & \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\varepsilon}}_N,\widehat{\underline{\lambda}}_N} \\ \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\lambda}}_N,\widehat{\underline{\varepsilon}}_N} & \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\lambda}}_N} \end{bmatrix}. \tag{4.35}$$

In (4.35), these four components are defined as

$$\widehat{\mathbf{\Sigma}}_{\widehat{\underline{\varepsilon}}_N} = \left(\nabla\widehat{\underline{\varepsilon}}_N\right)^{T}\widehat{\mathbf{\Psi}}_N\left(\nabla\widehat{\underline{\varepsilon}}_N\right), \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\lambda}}_N} = \left(\nabla\widehat{\underline{\lambda}}_N\right)^{T}\widehat{\mathbf{\Psi}}_N\left(\nabla\widehat{\underline{\lambda}}_N\right)$$
$$\widehat{\mathbf{\Sigma}}_{\widehat{\underline{\varepsilon}}_N,\widehat{\underline{\lambda}}_N} = \left(\nabla\widehat{\underline{\varepsilon}}_N\right)^{T}\widehat{\mathbf{\Psi}}_N\left(\nabla\widehat{\underline{\lambda}}_N\right), \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\lambda}}_N,\widehat{\underline{\varepsilon}}_N} = \left(\nabla\widehat{\underline{\lambda}}_N\right)^{T}\widehat{\mathbf{\Psi}}_N\left(\nabla\widehat{\underline{\varepsilon}}_N\right)$$

where $\nabla\widehat{\underline{\varepsilon}}_N$ is a consistent estimator of the $(k \times |A'(\underline{d})|) \times k$ Jacobian matrix of $\underline{\varepsilon}$, which is formed by the transposes of the matrices within the curly brackets in (4.29) and in (4.31) for each $\acute{j} \in A(\underline{d})$, $\nabla\widehat{\underline{\lambda}}_N$ is a consistent estimator of the $(k \times |A'(\underline{d})|) \times |A(\underline{d})|$ Jacobian matrix of $\underline{\lambda}$, which is formed by the transposes of the matrices within the curly brackets in (4.32) and in (4.34) for each $\acute{j} \in A(\underline{d})$, and $\widehat{\mathbf{\Psi}}_N$ is the $(k \times |A'(\underline{d})|) \times (k \times |A'(\underline{d})|)$ matrix with $\widehat{\sigma}_{\acute{i},\acute{i}}\mathbf{C}\left(\acute{i} \in A'(\underline{d})\right)$ on the diagonal and $\widehat{\sigma}_{\acute{i},\acute{h}}\mathbf{C}\left(\acute{i} \neq \acute{h}, \acute{h} \in A'(\underline{d})\right)$

on the off-diagonal, with $\mathbf{C}$ being the matrix $\left(1/N\left(\mathbf{X}^T\mathbf{X}\right)\right)^{-1}$ after deleting its first row and column (remember, $A'\left(\underline{d}\right) = A\left(\underline{d}\right) \cup \{0\}$). Notice that for the application of the Delta method, $\tau_N = \sqrt{N}$ and $\sqrt{N} \longrightarrow \infty$ as $N \longrightarrow \infty$.

4.8.2 Kodde and Palm (1986)'s procedure

We rewrite the null and alternative hypotheses in (4.16) in §4.4:

$$H_0^{(2)} : \underline{\gamma}_1 = \underline{0} \text{ and } \underline{\gamma}_2 \geq \underline{0},\ H_1^{(2)} : \underline{\gamma}_1 \neq \underline{0} \text{ or } \underline{\gamma}_2 \ngeq \underline{0}. \tag{4.36}$$

Remember that $\underline{\gamma}_1 = \sqrt{N}\underline{\varepsilon}$, $\underline{\gamma}_2 = \sqrt{N}\underline{\lambda}$, $\widehat{\underline{\gamma}}_1^N = \sqrt{N}\widehat{\underline{\varepsilon}}_N$, and $\widehat{\underline{\gamma}}_2^N = \sqrt{N}\widehat{\underline{\lambda}}_N$, where $\underline{\varepsilon}$ and $\underline{\lambda}$ were already defined in (4.12) and (4.13). Remember also from §4.4 that S_0 and S_1 denote the feasible spaces for $\underline{\gamma}=\left(\underline{\gamma}_1^T, \underline{\gamma}_2^T\right)^T$ under $H_0^{(2)}$ and $H_1^{(2)}$ in (4.36), respectively. Notice that when $H_0^{(2)}$ holds, S_0 is convex in $\underline{\gamma}$, and so is S_1 under $H_1^{(2)}$. Furthermore, define the distance function in the metric $\mathbf{V}$ of an arbitrary vector $\underline{x}$ from the origin by $\|\underline{x}\| = \underline{x}^T\mathbf{V}^{-1}\underline{x}$, where $\mathbf{V}$ is a given positive definite symmetric matrix. In the rest of this appendix, we consider $\widetilde{\mathbf{\Sigma}}$, which is a realized value of $\widehat{\mathbf{\Sigma}}$ defined in (4.35), for the distance function.

To test (4.36), Kodde and Palm (1986) suggests Wald's test statistic, (say) $\widehat{W} = \widehat{W}_0 - \widehat{W}_1$, where under $H_0^{(2)}$ in (4.36)

$$\widetilde{W}_0 = \underset{\underline{\gamma}^N \in S_0}{\text{minimize}} \left\|\widetilde{\underline{\gamma}}^N - \underline{\gamma}^N\right\|, \tag{4.37}$$

and under $H_1^{(2)}$ in (4.36)

$$\widetilde{W}_1 = \underset{\underline{\gamma}^N \in S_1}{\text{minimize}} \left\|\widetilde{\underline{\gamma}}^N - \underline{\gamma}^N\right\| \tag{4.38}$$

with $\widetilde{\underline{\gamma}}^N = \left[\left(\widetilde{\underline{\gamma}}_1^N\right)^T,\ \left(\widetilde{\underline{\gamma}}_2^N\right)^T\right]^T$, $\widetilde{W}_0$, and $\widetilde{W}_1$ being the realized values of $\widehat{\underline{\gamma}}^N$, $\widehat{W}_0$, and $\widehat{W}_1$. $\widetilde{W}_0$ and $\widetilde{W}_1$ in (4.37) and (4.38) are the distances of $\widetilde{\underline{\gamma}}^N$ to its orthogonal projections onto S_0 and S_1, respectively. Since S_0 and S_1 are convex under the relevant hypotheses, these orthogonal projections are uniquely determined.

If $H_1^{(2)}$ in (4.36) holds, $\widetilde{W}_1 = 0$ since S_1 is unrestricted. Therefore, $\widehat{W} = \widehat{W}_0$. If $H_0^{(2)}$ in (4.36) holds, $\widetilde{W}_0$ is obtained through solving the following quadratic optimization problem:

$$\begin{aligned}\underset{\underline{\gamma}_2^N \geq \underline{0}}{\text{minimize}} \quad & \left(\widetilde{\underline{\gamma}}_2^N - \underline{\gamma}_2^N - \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\lambda}}_N, \widehat{\underline{\varepsilon}}_N}\widehat{\mathbf{\Sigma}}_{\widehat{\underline{\varepsilon}}_N}^{-1}\widetilde{\underline{\gamma}}_1^N\right)^T \left(\widehat{\mathbf{\Sigma}}_{\widehat{\underline{\lambda}}_N} - \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\lambda}}_N, \widehat{\underline{\varepsilon}}_N}\widehat{\mathbf{\Sigma}}_{\widehat{\underline{\varepsilon}}_N}^{-1}\widehat{\mathbf{\Sigma}}_{\widehat{\underline{\varepsilon}}_N, \widehat{\underline{\lambda}}_N}\right)^{-1} \\ & \left(\widetilde{\underline{\gamma}}_2^N - \underline{\gamma}_2^N - \widehat{\mathbf{\Sigma}}_{\widehat{\underline{\lambda}}_N, \widehat{\underline{\varepsilon}}_N}\widehat{\mathbf{\Sigma}}_{\widehat{\underline{\varepsilon}}_N}^{-1}\widetilde{\underline{\gamma}}_1^N\right).\end{aligned}$$

Then, Kodde and Palm (1986) gives the statistic $\widehat{W} = \widehat{W}_0$ as

$$\widehat{W} = \left(\underline{\widehat{\gamma}}_1^N\right)^T \widehat{\mathbf{\Sigma}}_{\underline{\widehat{\varepsilon}}_N}^{-1} \left(\underline{\widehat{\gamma}}_1^N\right) + \left(\underline{\widehat{\gamma}}_2^N - \underline{\overline{\gamma}}_2^N - \widehat{\mathbf{\Sigma}}_{\underline{\widehat{\lambda}}_N, \underline{\widehat{\varepsilon}}_N} \widehat{\mathbf{\Sigma}}_{\underline{\widehat{\varepsilon}}_N}^{-1} \underline{\widehat{\gamma}}_1^N\right)^T \left(\widehat{\mathbf{\Sigma}}_{\underline{\widehat{\lambda}}_N} - \widehat{\mathbf{\Sigma}}_{\underline{\widehat{\lambda}}_N, \underline{\widehat{\varepsilon}}_N} \widehat{\mathbf{\Sigma}}_{\underline{\widehat{\varepsilon}}_N}^{-1} \widehat{\mathbf{\Sigma}}_{\underline{\widehat{\varepsilon}}_N, \underline{\widehat{\lambda}}_N}\right)^{-1} \left(\underline{\widehat{\gamma}}_2^N - \underline{\overline{\gamma}}_2^N - \widehat{\mathbf{\Sigma}}_{\underline{\widehat{\lambda}}_N, \underline{\widehat{\varepsilon}}_N} \widehat{\mathbf{\Sigma}}_{\underline{\widehat{\varepsilon}}_N}^{-1} \underline{\widehat{\gamma}}_1^N\right) \tag{4.39}$$

where $\underline{\overline{\gamma}}_2^N$ denotes the orthogonal projection of $\underline{\widetilde{\gamma}}_2^N$ onto S_0.

Kodde and Palm (1986) proves that the statistic $\widehat{W}$ in (4.39) has the following so-called chi-bar-squared distribution under $H_0^{(2)}$ in (4.36) (remember, k and $|A(\underline{d})|$ are the dimensions of $\underline{\varepsilon}$ in (4.12) and $\underline{\lambda}$ in (4.13), respectively):

$$\Pr\left\{\widehat{W} \geq u | \ \widehat{\mathbf{\Sigma}}\right\} = \sum_{t=0}^{|A(\underline{d})|} w_t \Pr\left\{\chi^2_{k+|A(\underline{d})|-t} \geq u\right\} \tag{4.40}$$

where u is the critical value, $\chi^2_{k+|A(\underline{d})|-t}$ is the central chi-squared random variable with $k + |A(\underline{d})| - t$ degrees of freedom, and w_t is the weight denoting the probability that t of $|A(\underline{d})|$ components of $\underline{\overline{\gamma}}_2^N$ are strictly positive. In (4.40), the weights w_t are such that $w_t \geq 0$ for each t, and $w_0 + ... + w_{|A(\underline{d})|} = 1$; see Kodde and Palm (1986).

In general, the computation of the weights w_t can be very complicated; see Shapiro (1988). For a given significance level α, Kodde and Palm (1986) provides the following formulas to obtain lower and upper bound critical values, say u_1 and u_2 respectively, for which the computation of w_t is not necessary:

$$\begin{aligned}
\alpha &= \inf_{\widetilde{\mathbf{\Sigma}} > \mathbf{0}} \Pr\left\{\widehat{W} \geq u_1 | \ \widehat{\mathbf{\Sigma}}\right\} \\
&= 0.5 \Pr\left\{\chi^2_k \geq u_1\right\} + 0.5 \Pr\left\{\chi^2_{k+1} \geq u_1\right\} \\
\alpha &= \sup_{\widetilde{\mathbf{\Sigma}} > \mathbf{0}} \Pr\left\{\widehat{W} \geq u_2 | \ \widehat{\mathbf{\Sigma}}\right\} \\
&= 0.5 \Pr\left\{\chi^2_{k+|A(\underline{d})|-1} \geq u_2\right\} + 0.5 \Pr\left\{\chi^2_{k+|A(\underline{d})|} \geq u_2\right\}.
\end{aligned}$$

To test (4.36), $\widetilde{W}$, which is a realization of $\widehat{W}$, is compared to u_1 and u_2: if $\widetilde{W} \leq u_1$, then $H_0^{(2)}$ in (4.36) is not rejected. However, if $\widetilde{W} \geq u_2$, then $H_0^{(2)}$ in (4.36) is rejected. The problem with Kodde and Palm (1986)'s approach is that the test is inconclusive whenever $u_1 < \widetilde{W} < u_2$. In such a case, the weights w_t in (4.40) must be computed.

Chapter 5

Conclusions and Further Research

In this thesis, we focused on a specific black box simulation optimization method, namely Response Surface Methodology (RSM) - where black box means that the gradient estimates through simulation are not available using either perturbation analysis or likelihood ratio score function. Other black box simulation optimization methods are the stochastic approximation method and metaheuristics (e.g., genetic algorithms, tabu search, and simulated annealing).

In this study, we dealt with several disadvantages of RSM. Myers and Montgomery (2002) states that classic RSM suffers from two well-known disadvantages: (i) the classic RSM's search direction is scale dependent; (ii) the step size along the search direction is selected intuitively. Therefore, we derived a novel, scale independent search direction - which we called adapted steepest ascent - by adjusting the estimated first-order factor effects through their estimated covariance matrix. In most of our numerical experiments, the novel search direction performed better than the classic RSM's search direction. We also derived and explored possible solutions for the step size.

Another disadvantage of RSM is that classic RSM was derived for unconstrained optimization problems. Myers (1999) also points out the need for extending classic RSM to cope with multiple random responses. Hence, we extended classic RSM to handle stochastic constraints on the outputs, as well as deterministic box constraints on the inputs. This was achieved through the generalization of the search direction, using ideas from interior point methods. Furthermore, we proved that the proposed search direction is scale independent. We also provided a heuristic procedure that uses the proposed search direction iteratively. That procedure was primarily meant for time-consuming simulation optimization problems with a tight computational budget. Our numerical results were encouraging; that is, our procedure reached the desired neighborhood of the true optimum in a relatively small number of simulation runs.

The final disadvantage of RSM that we handled in this thesis is the lack of a statistical stopping rule when there are (possibly stochastic) constraints. Thus, we derived a statistical stopping rule that assumes large samples for black box simulation optimization methods. That rule tests the first-order necessary optimality conditions - originally derived for deterministic optimization - at a feasible point. To obtain that rule, we used the Delta method, which proves that under certain conditions nonlinear statistics are asymptotically multivariate normally distributed, and Kodde and Palm (1986)'s statistical procedure, which is based on Wald's statistic. The results of our numerical experiments were encouraging; that is, the null hypothesis (i.e., the first-order necessary optimality conditions hold) was "accepted" at the true optimum very often. Furthermore, the estimated power increased as the points moved away from the true optimum.

There are more problems of RSM that we have not tackled in this thesis. For example, RSM does not necessarily exploit the stochastic structure of the simulated system. As a result, RSM may be computationally more expensive than white box methods, which do take the stochastic structure into account. Therefore, in cases where the system's stochastic structure is available, white box methods will be computationally better alternatives. Furthermore, RSM is not a global optimization method, and it is not suitable for discrete optimization. In those situations, metaheuristics may be better alternatives. Finally, the convergence properties of RSM have not yet been studied, whereas the convergence properties of the stochastic approximation method - which is an alternative to RSM - have been well-studied. These problems are left for further research.

Samenvatting

Dit proefschrift behandelt het optimaliseren van gesimuleerde systemen. Het proefschrift gebruikt "Response Surface Methodology" (RSM), die het simulatiemodel als een "black box" beschouwt. Die beschouwing houdt in dat de gradiënten (wiskundige afgeleiden) niet door middel van simulatie worden geschat, zoals dat wel gebeurt in "perturbation analysis" en "likelihood ratio/score function". Oorspronkelijk is RSM door Box en Wilson (1951) geïntroduceerd. Myers en Montgomery (2002) geven - in hun monografie over RSM - de volgende globale omschrijving, uitgaande van een minimalisatie-probleem.

In de eerste stap van RSM wordt een statistische proefopzet gebruikt om een eerste-orde polynoom te schatten dat lokaal goed aansluit bij de geobserveerde waarden van de stochastische (toevallige) simulatierespons. Die schatting maakt gebruik van lineaire regressieanalyse. Vervolgens wordt - via dat eerste-orde polynoom - de zoekrichting gezocht met de "steepest ascent". Stappen in die richting worden een aantal malen herhaald - totdat er geen significante verbetering van de response gevonden wordt. De gehele procedure wordt herhaald, totdat het eerste-orde polynoom een slechte benadering wordt, wat wil zeggen dat de gradiënt niet significant afwijkt van nul.

In de tweede stap van RSM wordt een tweede-orde polynoom geschat, die wordt geminimaliseerd. Verder worden canonieke en "ridge" analyses toegepast om de eigenschappen van de gevonden doelfunctie te bepalen (concaaf, convex, of onbepaald) en van het geschatte optimum (uniek optimum of meerdere optima). Deze tweede stap van RSM wordt in dit proefschrift niet verder behandeld.

Klassieke RSM neemt aan dat er een enkele, stochastische respons is. In Hoofdstuk 2 bestuderen we zo'n probleem. In tegenstelling tot de klassieke benadering die de geschatte stochastische respons maximaliseert, maximaliseren wij de ondergrens van die respons. Onze benadering is robuuster (pessimistischer). Wij concentreren ons op de eerste stappen, waarbij RSM met lineaire regressie een lokaal eerste-orde polynoom zoekt dat past bij de respons. Dat gebeurt middels de gebruikelijke kleinste-

kwadratenmethode. Om deze schatting mogelijk te maken, kan RSM een proefopzet van "resolution-3" gebruiken, omdat zo'n proefopzet zuivere schatters oplevert voor de regressiecoëfficiënten - met een klein aantal simulatieruns, vooropgesteld dat eerste-orde polynomen adequate benaderingen zijn (Kleijnen 1998).

Myers en Montgomery (2002) stellen dat de Steepest Ascent zoekrichting twee problemen kent:

- De Steepest Ascent zoekrichting is schaalafhankelijk.
- De stapgrootte in die richting moet intuïtief bepaald worden.

Simulatieanalisten gebruiken hun intuïtie om een stapgrootte te keizen. Wanneer vervolgens deze stapgrootte tot een slechtere respons leidt, wordt hij verkleind; zo niet, dan wordt nog een stap in dezelfde richting gezet. Een voorbeeld is de praktijkstudie in Kleijnen (1993), waarin een stapgrootte wordt gebruikt die de belangrijkste invoervariabele met een factor twee laat toenemen.

In Hoofdstuk 2 leiden we een nieuwe zoekrichting af, die we "adaptive Steepest Ascent" noemen. We vinden deze richting door de geschatte eerste-orde effecten van de invoervariabelen aan te passen via hun geschatte covarianties. Verder bewijzen we dat deze adaptatie schaalonafhankelijk is. In de meeste numerieke experimenten presteerde de nieuwe zoekrichting beter dan de klassieke zoekrichting. We leiden ook nieuwe oplossingen voor de stapgrootte af, en exploreren hun toepassing.

In Hoofdstuk 3 behandelen we een probleem met een enkele stochastische doelfunctie en meerdere voorwaarden voor de overige stochastische uitvoervariabelen, naast deterministische randvoorwaarden voor de invoervariabelen. Ons probleemtype is dus algemener dan het klassieke RSM probleemtype. Tot nu toe heeft ons probleemtype niet veel aandacht gekregen in de simulatieoptimalisatie, zie Fu (2002, p.6). Ook Myers (1999) wijst op het belang van het uitbreiden van klassieke RSM, teneinde om te kunnen gaan met meervoudige stochastische responsies.

Er zijn verschillende benaderingen om RSM te generaliseren, zodat RSM overweg kan met stochastische randvoorwaarden. Voorbeelden zijn de "desirability function" (Harrington 1965, Derringer en Suich 1980), de "generalized distance" (Khuri en Conlon 1981), de duale response (Myers en Carter 1973, Vining en Myers 1990, Del Castillo en Montgomery 1993, en Fan en Del Castillo (1999)), en "prediction interval constrained goal programming" (Wei, Olson, en White 1990). Bij al deze benaderingen wordt het optimalisatieprobleem met nevenvoorwaarden geherformuleerd door de nevenvoorwaarden te combineren met de oorspronkelijke doelfunctie: een

nieuwe, enkelvoudige doelfunctie wordt geformuleerd via geschikte transformaties. Vervolgens wordt het resulterende probleem (zonder nevenvoorwaarden) opgelost met een gebruikelijk niet-lineair programmeringalgoritme. Een groot nadeel van deze benaderingen is de arbitraire keuze van de transformaties.

Een van de bijdragen van Hoofdstuk 3 is dus de uitbreiding van klassieke RSM dusdanig dat RSM kan omgaan met zowel stochastische beperkingen voor de uitvoer als deterministische beperkingen voor de invoervariabelen. Dit wordt bereikt door de generalisatie - van de geschatte Steepest Ascent zoekrichting - via ideeën uit "inwendige punt" methoden, specifiek het "affine scaling"-algoritme. Onze zoekrichting is de geschaalde en geprojecteerde geschatte Steepest Ascent (estimated affine scaling search direction). Verder bewijzen we dat onze zoekrichting twee belangrijke eigenschappen heeft: zij is inderdaad een richting die leidt tot lagere waarden in minimalisatieproblemen, en zij is schaalonafhankelijk.

Een andere bijdrage van Hoofdstuk 3 is een heuristiek voor de eerste stap in RSM, die de voorgestelde nieuwe zoekrichting iteratief volgt. Deze heuristiek gebruikt een aangepaste binaire zoekmethode om de stapgrootte in de zoekrichting te bepalen. Ook past onze methode statistische toetsing van hypotheses toe, om de volgende iteratie in de voorgestelde zoekrichting te bepalen. Deze procedure is voornamelijk bedoeld voor problemen - in de optimalisatie van gesimuleerde systemen - die veel rekentijd kosten en een strak budget voor computertijd hebben. Met andere woorden, onze procedure convergeert snel naar een acceptabele benadering van het optimum. Dit is noodzakelijk bij simulatiestudies waarbij iedere computerrun vele uren of dagen kost maar het totale budget zo beperkt is dat er maar een beperkt aantal runs doorgerekend kunnen worden. Tenslotte laten we zien, hoe de originele heuristiek voor een stochastisch probleem kan worden vereenvoudigd tot die voor een deterministisch probleem.

Gebruik makend van ideeën uit "inwendige punt " methoden (dat wil zeggen, naar de optimale oplossing toewerken door volledig binnen de toegelaten oplossingsruimte te blijven) heeft de volgende twee voordelen:

- Onze heuristiek vermijdt dat hij langs de rand van de oplossingsruimte kruipt.
- Sommige simulatieprogrammatuur kan falen of verkeerde resultaten geven buiten de toegelaten oplossingsruimte - maar onze heuristiek komt daar niet!

Hoofdstuk 4 gaat weer over problemen met een stochastische doelfunctie en stochastische nevenvoorwaarden, maar we concentreren ons nu op de laatste fase van

een “black box” methode voor de optimalisatie van gesimuleerde systemen. In concreto, leiden we een statistische stopregel af die grote steekproeven veronderstelt. Deze regel toetst de Karush-Kuhn-Tucker (KKT) eerste-orde noodzakelijke optimaliteitscondities voor een toelaatbare oplossing die wel bekend zijn uit de deterministische optimalisatie. Omdat ons probleem stochastisch van aard is, gebruiken we de statistische procedure van Kodde en Palm (1986), die gebaseerd is op Wald's toets. Deze procedure maakt het ons mogelijk om samengestelde (meervoudige) hypothesen te toetsen De resultaten van onze numerieke experimenten zijn bemoedigend: de nulhypothese (namelijk, de KKT-condities gelden) wordt heel vaak in het optimum geaccepteerd. Ook nemen de fouten van type-II (oftewel de bètafouten) af naar mate de getoetste oplossingen dichter bij het werkelijke optimum liggen, wat een erg gewenste eigenschap is. Opgemerkt zij dat recentelijk Bettonvil en Kleijnen (2004) een nieuwe procedure afgeleid hebben die de KKT-condities toetst voor rekenintensieve simulaties (zonder grote steekproeven), gebruikmakend van een “bootstrap”-procedure.

Wanneer er (stochastische of deterministische) nevenvoorwaarden zijn, hoeven we wellicht niet tot de tweede stap van RSM over te gaan. De geschatte gradiënt van de gevonden oplossing kan namelijk significant afwijken van nul - in de buurt van het werkelijke optimum (per definitie eindigt de eerste stap van RSM wanneer de geschatte gradiënt niet significant meer van nul afwijkt). Opgemerkt zij dat - voor zover wij weten - de KKT-condities niet genoemd worden in de klassieke RSM-literatuur, omdat in deze literatuur nog niet uitgebreid naar nevenvoorwaarden is gekeken; een uitzondering is Fan en Del Castillo (1999).

Verder presenteren we twee alternatieve statistische toetsen, namelijk Roy's toets en de klassieke F-toets gecombineerd met Bonferroni's ongelijkheid. We gebruiken deze toetsen om te kijken naar de benaderingsfout (lack of fit) van de geschatte polynomen van de eerste orde. We laten ook zien dat Roy's toets de klassieke F-toets generaliseert voor multipele responsies.

References

Alrefaei, M. H., S. Andradóttir. 1999. A simulated annealing algorithm with constant temperature for discrete stochastic optimization. *Management Science* **45** (5), 748-764.

Anandalingam, G. 2001. Simulated annealing. In *Encyclopedia of Operations Research and Management Science*, 2nd ed., edited by S. Gass and C. Harris, 748-751. Kluwer Academic Publishers, Boston.

Andradóttir, S. 1998. A review of simulation optimization techniques. In *Proceedings of the 1998 Winter Simulation Conference*, edited by D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, 151-158. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Andradóttir, S., J. S. Carson, M. C. Fu, F. Glover, C. R. Harrell, Y. C. Ho, J. P. Kelly, S. M. Robinson. 2000. Integrating optimization and simulation: research and practice. In *Proceedings of the 2000 Winter Simulation Conference*, edited by J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, 610-616. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Angün, E., J. P. C. Kleijnen. 2004. An asymptotic stopping rule for simulation optimization. Working paper, Tilburg University.

Angün, E., G. Gürkan, D. den Hertog, J. P. C. Kleijnen. 2002. Response surface methodology revisited. In *Proceedings of the 2002 Winter Simulation Conference*, edited by E. Yücesan, C. H. Chen, J. L. Snowdon, and J. M. Charnes, 377-383. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Angün, E., G. Gürkan, D. den Hertog, J. P. C. Kleijnen. 2003. Response surface methodology with stochastic constraints for expensive simulation. Working paper, Tilburg University.

April, J., F. Glover, J. P. Kelly. 2003. Optfolio - a simulation optimization system for project portfolio planning. In *Proceedings of the 2003 Winter Simulation Conference*, edited by S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 301-309. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Azadivar, F. 1999. Simulation optimization methodologies. In *Proceedings of the 1999 Winter Simulation Conference*, edited by P. A. Farrington, H. B. Nembhard, D. T. Stuffock, and G. W. Evans, 93-100. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Azadivar, F., T. H. Truong. 2003. Simulation based optimization for supply chain configuration design. In *Proceedings of the 2003 Winter Simulation Conference*, edited by S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 1268-1275. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Barnes, E. R. 1986. A variation on Karmarkar's algorithm for solving linear programming problems. *Mathematical Programming* **36**, 174-182.

Ben-Gal, I., J. Bukchin. 2002. The ergonomic design of workstations using virtual manufacturing and response surface methodology. *IIE Transactions* **34**, 375-391.

Bettonvil, B., J. P. C. Kleijnen. 1996. Searching for important factors in simulation models with many factors: sequential bifurcation. *European Journal of Operational Research* **96**, 180-194.

Bettonvil, B., J. P. C. Kleijnen. 2004. Statistical testing of optimality conditions in simulation-based optimization. Working paper, Tilburg University.

Boesel, J., B. L. Nelson, S. H. Kim. 2003. Using ranking and selection to "clean up" after simulation optimization. *Operations Research* **51** (5), 814-825.

Boesel, J., R. O. Bowden, F. Glover, J. P. Kelly, E. Westwig. 2001. Future of simulation optimization. In *Proceedings of the 2001 Winter Simulation Conference*, edited by B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 1466-1469. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Booker, A. J., J. E. Dennis, Jr., P. D. Frank, D. B. Serafini, V. Torczon, M. W. Trosset. 1999. A rigorous framework for optimization of expensive functions by surrogates. *Structural Optimization* **17** (1), 1-13.

Box, G. E. P. 1999. Statistics as a catalyst to learning by scientific method, part II - a discussion. *Journal of Quality Technology* **31** (1), 16-29.

Box, G. E. P., K. B. Wilson. 1951. On the experimental attainment of optimum conditions. *Journal Royal Statistical Society*, Series B, **13** (1), 1-38.

Brekelmans, R., L. Driessen, H. Hamers, D. den Hertog. 2003. Gradient estimation schemes for noisy functions. *Journal of Optimization Theory and Applications*, to appear.

Campolongo, F., J. P. C. Kleijnen, T. Andres. 2000. Screening methods in sensitivity analysis. In *Mathematical and Statistical Methods for Sensitivity Analysis of Model Output*, edited by A. Saltelli and K. Chan. Wiley, Chichester.

Chen, C., V. Kumar, Y. Luo. 1998. Motion planning of walking robots using ordinal optimization. *IEEE Robotics and Automation Magazine* **5**, 22-32.

Cheng, R. C. H. 1997. Searching for important factors: sequential bifurcation under uncertainty. In *Proceedings of the 1997 Winter Simulation Conference*, edited by S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson, 275-280. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Conn, A. R., N. Gould, Ph. L. Toint. 2000. *Trust-Region Methods.* SIAM, Philadelphia.

Dai, L. 1996. Convergence properties of ordinal comparison in the simulation of discrete event dynamic systems. *Journal of Optimization Theory and Applications* **91**, 363-388.

Dai, L., C. Chen. 1997. Rate of convergence for ordinal comparison of dependent simulations in discrete event dynamic systems. *Journal of Optimization Theory and Applications* **94**, 29-54.

Dean, A. M., S. M. Lewis, eds. 2004. *Screening.* Springer-Verlag, New York.

Del Castillo, E., D. C. Montgomery. 1993. Nonlinear programming solution to the dual response problem. *Journal of Quality Technology* **25** (3), 199-204.

Derringer, G., R. Suich. 1980. Simultaneous optimization of several response variables. *Journal of Quality Technology* **12**, 214-219.

Donohue, J. M., E. C. Houck, R. H. Myers. 1993. Simulation design and correlation induction for reducing second-order bias in first-order response surfaces. *Operations Research* **41** (5), 880-902.

Donohue, J. M., E. C. Houck, R. H. Myers. 1995. Simulation designs for the estimation of quadratic response surface gradients in the presence of model misspecification. *Management Science* **41** (2), 244-262.

Draper, N. R., F. Pukelsheim. 2000. Ridge analysis of mixture response surfaces. *Statistics and Probability Letters* **48**, 131-140.

Driessen, L., R. Brekelmans, H. Hamers, D. den Hertog. 2001. On D-optimality based trust regions for black-box optimization problems. Center Discussion Paper No. 2001-69, Tilburg University.

Fan, S. K., E. Del Castillo. 1999. Calculation of an optimal region of operation for dual response systems fitted from experimental data. *Journal of the Operational Research Society* **50** (8), 826-836.

Fouskakis, D., D. Draper. 2002. Stochastic optimization: a review. *International Statistical Review* **70** (3), 315-349.

Fu, M. C. 2002. Optimization for simulation: theory vs. practice. *INFORMS Journal on Computing* **14** (3), 192-215.

Fu, M. C., S. D. Hill. 1997. Optimization of discrete event systems via simultaneous perturbation stochastic approximation. *IIE Transactions* **29**, 233-243.

Fu, M. C., J. Q. Hu. 1997. *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*. Kluwer Academic, Boston.

Gass, S. I., C. M. Harris, eds. 2000. *Encyclopedia of Operations Research and Management Science*, 2nd ed. Kluwer Academic, Boston.

Gerencsér, L., S. D. Hill, Z. Vágó. 1999. Optimization over discrete sets via SPSA. In *Proceedings of the 1999 Winter Simulation Conference*, edited by P. A. Farrington, H. B. Nembhard, D. T. Stuffock, and G. W. Evans, 466-470. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Gill, P. E., W. Murray, M. H. Wright. 2000. *Practical Optimization*, 12th ed. Academic Press, London.

Glasserman, P. 1991. *Gradient Estimation via Perturbation Analysis*. Kluwer Academic, Boston.

Glover, F. 2001. Tabu search. In *Encyclopedia of Operations Research and Management Science*, 2nd ed., edited by S. Gass and C. Harris, 821-827. Kluwer Academic Publishers, Boston.

Glover, F., M. Laguna. 1997. Tabu search. Kluwer Academic, Boston.

Glynn, P. W. 1989. Optimization of stochastic systems via simulation. In *Proceedings of the 1989 Winter Simulation Conference*, edited by P. Heidelberger, E. A. MacNair, and K. J. Musselman, 90-105. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Goldsman, D., B. L. Nelson. 1994. Ranking, selection, and multiple comparison in computer simulation. In *Proceedings of the 1994 Winter Simulation Conference*, edited by J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila, 192-199. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Goldsman, D., B. L. Nelson, T. Opicka, A. A. B. Pritsker. 1999. A ranking and selection project: experiences from a university-industry collaboration. In *Proceedings of the 1999 Winter Simulation Conference*, edited by P. A. Farrington, H. B. Nembhard, D. T. Stuffock, and G. W. Evans, 83-92. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Gouriéroux, C., A. Holly, A. Monfort. 1982. Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. *Econometrica* **50**, 63-80.

Gutjahr, W. J., G. Ch. Pflug. 1996. Simulated annealing for noisy cost functions. *Journal of Global Optimization* **8**, 1-13.

Gürkan, G., A. Y. Özge, S. M. Robinson. 1994. Sample-path optimization in simulation. In *Proceedings of the 1994 Winter Simulation Conference*, edited by J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila, 247-254. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Gürkan, G., A. Y. Özge, S. M. Robinson. 1999. Solving stochastic optimization problems with stochastic constraints: an application in network design. In *Proceedings of the 1999 Winter Simulation Conference*, edited by P. A. Farrington,

H. B. Nembhard, D. T. Stuffock, and G. W. Evans, 471-478. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Harrington, E. C. 1965. The desirability function. *Industrial Quality Control* **21**, 494-498.

Helton, J. C., D. R. Anderson, M. G. Marietta, R. P. Rechard. 1997. Performance assessment for the waste isolation pilot plant: from regulation to calculation for 40 CFR 191.13. *Operations Research* **45** (2), 157-177.

den Hertog, D., H. P. Stehouwer. 2002. Optimizing color picture tubes by high-cost nonlinear programming. *European Journal of Operational Research* **140** (2), 197-211.

Ho, Y. C., X. R. Cao. 1991. *Perturbation Analysis of Discrete Event Dynamic Systems*. Kluwer Academic, Boston.

Ho, Y. C., M. E. Larson. 1995. Ordinal optimization approach to rare event probability problems. *Journal of Discrete Event Dynamic Systems* **5**, 281-301.

Ho, Y. C., R. S. Sreenivas, P. Vakili. 1992. Ordinal optimization of DEDS. *Journal of Discrete Event Dynamic Systems* **2**, 61-88.

Ho, Y. C., C. G. Cassandras, C. H. Chen, L. Dai. 2000. Ordinal optimisation and simulation. *Journal of the Operational Research Society* **51**, 490-500.

Hood, S. J., P. D. Welch. 1993. Response surface methodology and its application in simulation. In *Proceedings of the 1993 Winter Simulation Conference*, edited by G. W. Evans, M. Mollaghasemi, E. C. Russell, and W. E. Biles, 115-122. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Hussey, J. R., R. H. Myers, E. C. Houck. 1987. Correlated simulation experiments in first-order response surface design. *Operations Research* **35** (5), 744-758.

Hutchison, D. W., S. D. Hill. 2001. Simulation optimization of airline delay with constraints. In *Proceedings of the 2001 Winter Simulation Conference*, edited by B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 1017-1022. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Irizarry, M., J. R. Wilson, J. Trevino. 2001. A flexible simulation tool for manufacturing-cell design, II: response surface analysis and case study. *IIE Transactions* **33**, 837-846.

Irizarry, M., M. E. Kuhl, E. K. Lada, S. Subramanian, J. R. Wilson. 2003. Analyzing transformation-based simulation metamodels. *IIE Transactions* **35**, 271-283.

Jones, D. R., M. Schonlau, W. J. Welch. 1998. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization* **13**, 455-492.

Khuri, A. I. 1985. A test for lack of fit of a linear multiresponse model. *Technometrics* **27**, 213-218.

Khuri, A. I. 1996. Multiresponse surface methodology. In *Handbook of Statistics* **13**, edited by S. Ghosh and C. R. Rao. Elsevier, Amsterdam.

Khuri, A. I., M. Conlon. 1981. Simultaneous optimization of multiple responses represented by polynomial regression functions. *Technometrics* **23** (4), 363-375.

Khuri, A. I., J. A. Cornell. 1996. *Response Surfaces: Design and Analysis*, 2nd ed. Marcel Dekker, New York.

Kiefer, J., J. Wolfowitz. 1952. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics* **23**, 462-466.

Kirkpatrick, S., C. D. Gelatt, M. P. Vwecchi. 1981. Optimization by simulated annealing. *Science* **220**, 671-680.

Kleijnen, J. P. C. 1975. *Statistical Techniques in Simulation*, Part II. Marcel Dekker, New York.

Kleijnen, J. P. C. 1987. *Statistical Tools for Simulation Practitioners.* Marcel Dekker, New York.

Kleijnen, J. P. C. 1993. Simulation and optimization in production planning: a case study. *Decision Support Systems* **9**, 269-280.

Kleijnen, J. P. C. 1998. Experimental design for sensitivity analysis, optimization and validation of simulation models. In *Handbook of Simulation*, edited by J. Banks. Wiley, New York.

Kleijnen, J. P. C., D. den Hertog, E. Angün. 2004. Response surface methodology's steepest ascent and step size revisited. *European Journal of Operational Research*, in press.

Kleinman, N. L., J. C. Spall, D. Q. Naiman. 1999. Simulation-based optimization with stochastic approximation using common random numbers. *Management Science* **45** (11), 1570-1578.

Kodde, D. A., F. C. Palm. 1986. Wald criteria for jointly testing equality and inequality restrictions. *Econometrica* **54**, 1243-1248.

Koksoy, O. 2003. Joint optimization of mean and standard deviation using response surface methods. *Journal of Quality Technology* **35** (3), 239-252.

Law, A. M., W. D. Kelton. 2000. *Simulation Modeling and Analysis*, 3rd ed. McGraw-Hill, Boston.

L'Ecuyer, P. 1991. An overview of derivative estimation. In *Proceedings of the 1991 Winter Simulation Conference*, edited by B. L. Nelson, W. D. Kelton, and G. M. Clark, 207-217. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

McDaniel, W. R., B. E. Ankenman. 2000. A response surface test bed. *Quality and Reliability Engineering International* **16**, 363-372.

Magnus, J. R., H. Neudecker. 1988. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, Chichester.

Martin, E., M. Schouwenaar. 2003. Optimization of a telecommunication billing system. In *Proceedings of the 2003 Winter Simulation Conference*, edited by S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 1843-1847. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Maryak, J. L., D. C. Chin. 2001. Global random optimization by simultaneous perturbation stochastic approximation. In *Proceedings of the 2001 Winter Simulation Conference*, edited by B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 307-312. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Mathai, A. M., S. B. Provost. 1992. *Quadratic Forms in Random Variables: Theory and Applications*. Dekker, New York.

Michalewicz, Z., M. Schoenauer. 2001. Evolutionary algorithms. In *Encyclopedia of Operations Research and Management Science*, 2nd ed., edited by S. Gass and C. Harris, 264-269. Kluwer Academic Publishers, Boston.

Monma, C. L., A. J. Morton. 1987. Computational experience with a dual affine variant of Karmarkar's method for linear programming. *Operations Research Letters* **6** (6), 261-267.

Muirhead, R. J. 1982. *Aspects of Multivariate Statistical Theory.* Wiley, New York.

Myers, R. H. 1999. Response surface methodology - current status and future directions. *Journal of Quality Technology* **31** (1), 30-74.

Myers, R. H., W. H. Carter. 1973. Response surface techniques for dual response systems. *Technometrics* **15**, 301-317.

Myers, R. H., D. C. Montgomery. 2002. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 2nd ed. Wiley, New York.

Neddermeijer, H. G.,G. J. van Ootmarsum, N. Piersma, R. Dekker. 2000. A framework for response surface methodology for simulation optimization models. In *Proceedings of the 2000 Winter Simulation Conference*, edited by J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, 129-136. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Patsis, N. T., C. Chen, M. E. Larson. 1997. SIMD parallel discrete event dynamic system simulation. *IEEE Transactions on Control Systems Technology* **5**, 30-41.

Rao, C. R. 1967. Least squares theory using an estimated dispersion matrix and its application to measurement of signals. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, I, 355-372.

Roy, S. N., R. Gnanadesikan, J. N. Srivastava. 1971. *Analysis and Design of Certain Quantitative Multiresponse Experiments.* Pergamon Press, Oxford.

Rubinstein, R. Y., A. Shapiro. 1993. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method.* Wiley, Chichester.

Ruud, P. A. 2000. *An Introduction to Classical Econometric Theory.* Oxford University Press, New York.

Sacks, J., W. J. Welch, T. J. Mitchell, H. P. Wynn. 1989. Design and analysis of computer experiments. *Statistical Science* **4** (4), 409-435.

Safizadeh, M. H. 2002. Minimizing the bias and variance of the gradient estimate in RSM simulation studies. *European Journal of Operational Research* **136** (1), 121-135.

Schruben, L. W., B. H. Margolin. 1978. Pseudo-random number assignment in statistically designed simulation and distribution sampling experiments. *Journal of the American Statistical Association* **73**, 504-520.

Shapiro, A. 1988. Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review* **56**, 49-62.

Shapiro, A., T. Homem-de-Mello. 1998. A simulation-based approach to two-stage stochastic programming with recourse. *Mathematical Programming* **81**, 301-325.

Simpson, T. W., T. M. Mauery, J. J. Korte, F. Mistree. 2001. Kriging metamodels for global approximation in simulation-based multidisciplinary design optimization. *American Institute of Aeronautics and Astronautics Journal* **39** (12), 2233-2241.

Spall, J. C. 1992. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control* **37** (3), 332-341.

Spall, J. C. 1998. Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Transactions on Aerospace and Electronic Systems* **34** (3), 817-823.

Spall, J. C. 1999. Stochastic optimization and the simultaneous perturbation method. In *Proceedings of the 1999 Winter Simulation Conference*, edited by P. A. Farrington, H. B. Nembhard, D. T. Stuffock, and G. W. Evans, 101-109. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Spall, J. C. 2000. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Transactions on Automatic Control* **45** (10), 1839-1853.

Spall, J. C. 2003. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley, Hoboken, New Jersey.

Theil, H. 1971. *Principles of Econometrics*. Wiley, New York.

Trocine, L., L. C. Malone. 2000. Finding important independent variables through screening designs: a comparison of methods. In *Proceedings of the 2000 Winter*

Simulation Conference, edited by J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, 749-754. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Vining, G. G., R. H. Myers. 1990. Combining Taguchi and response surface philosophies: a dual response approach. *Journal of Quality Technology* **22** (1), 38-45.

Vonk Noordegraaf, A., M. Nielen, J. P. C. Kleijnen. 2003. Sensitivity analysis by experimental design and metamodelling: case study on simulation in national animal disease control. *European Journal of Operational Research* **146** (3), 433-443.

Wei, C. J., D. L. Olson, E. M. White. 1990. Simultaneous optimization in process quality control via prediction-interval constrained programming. *Journal of the Operational Research Society* **41** (12), 1161-1167.

Whitney, J. E., L. I. Solomon, S. D. Hill. 2001. Constrained optimization over discrete sets via SPSA with application to non-separable resource allocation. In *Proceedings of the 2001 Winter Simulation Conference*, edited by B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 313-317. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Yang, T., L. Tseng. 2002. Solving a multi-objective simulation model using a hybrid response surface method and lexicographical goal programming approach: a case study on integrated circuit ink-marking machines. *Journal of the Operational Research Society* **53** (2), 211-221.

Ye, Y. 1989. An extension of Karmarkar's algorithm and the trust region method for quadratic programming. *Progress in Mathematical Programming: Interior-Point and Related Methods*. Wiley, Chichester.