Proceedings of the 2006 Winter Simulation Conference
L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds.

SIMULATION-BASED RESPONSE SURFACE COMPUTATION IN THE PRESENCE OF MONOTONICITY

Eunji Lim
Peter W. Glynn

Department of Management Science and Engineering
Stanford University
Stanford, CA 94305, U.S.A.

ABSTRACT

In many stochastic models, it is known that the response surface corresponding to a particular performance measure is monotone in the underlying parameter. For example, the steady-state mean waiting time for customers in a single server queue is known to be monotone in the service rate. In other contexts, the simulator may believe, on the basis of intuition, that the response surface is monotone. This paper describes an appropriate methodology for incorporating such monotonicity constraints into one's response surface estimator.

1 INTRODUCTION

One of the main challenges presented to simulation-based computation is that of computing response surfaces. In particular, suppose that we wish to compute a function (or "response surface") $\alpha(\cdot)$ via simulation. Specifically, we assume that for each θ in the domain of $\alpha(\cdot)$, we can compute the function value $\alpha(\theta)$ via simulation. The goal of a response surface computation is to accurately compute a global approximation to the function $\alpha(\cdot)$.

One widely used approach is to assume that $\alpha(\cdot)$ can be well approximated by a low order polynomial, and to use regression-based methods to estimate the coefficients of the polynomial; see, for example, Chapter 12.4 of Law and Kelton (2000). In this paper, we propose a different approach that is applicable to response surface that are known or believed to be monotone in the parameter θ. It should be noted that a large number of different stochastic models are provably known to be monotone in various parameters; see, for example, Weber (1983), Van Oyen (1997), Chang et al. (1990, 1991), Bauerle (1997), and Shaked and Shanthikumar (1988).

In the presence of monotonicity, we suggest a general framework for enforcing the monotonicity constraints on our estimator. Our approach leads to a new estimator, called the isotonic regression estimator, for computing the response surface. The isotonic regression estimator involves solving a quadratic program (with a positive definite objective function and linear inequality constraints) with as many decision variables as there are points at which function values have been computed.

Isotonic regression has been previously proposed within the statistics literature; see, for example, Brunk (1958). The simulation environment requires several generalizations relative to the statistical setting, including the possibility of correlation across the points (due to use of common random numbers, common control variates, etc.) and greater interest in the case where the number m of points is small relative to the overall computational budget.

In Section 2 of this paper, we discuss a general framework for analysis of isotonic regression that is appropriate to the simulation setting, and describes the isotonic regression estimator. Section 3 deals with the particular case in which the function evaluation simulations are independent across the different points, and argues that the isotonic regression estimator can generally be expected to dominate the conventional Monte Carlo estimator in terms of asymptotic performance. In Section 4, we discuss the general isotonic estimator in the case of two points, taking advantage of the explicit solution of the quadratic program that is available in the context. Section 5 surveys some of the limit theory that is available when the number m of points is large, while Section 6 discusses some numerical computation that compares the performance of the proposed estimators against the conventional alternatives.

2 PROBLEM FORMULATION AND BASIC APPROACH

Our goal is to compute a function $\alpha : \Theta \to R$ via Monte Carlo (stochastic) simulation. We assume that Θ is a subset of R^d. If $\theta_1 = (\theta_{11}, \ldots, \theta_{1d})^T$, $\theta_2 = (\theta_{21}, \ldots, \theta_{2d})^T$ with θ_1, $\theta_2 \in \Theta$, we say that θ_1 is less than or equal to θ_2 (and write $\theta_1 \preccurlyeq \theta_2$) if $\theta_{1i} \leq \theta_{2i}$ for $1 \leq i \leq d$. The relation $\preccurlyeq$ is a partial order on R^d; see Shen and Vereshchagin (2002) for the definition of a partial order. We assume throughout this paper that α is non-decreasing with respect to the partial order $\preccurlyeq$ over Θ, so that $\alpha(\theta_1) \leq \alpha(\theta_2)$ whenever $\theta_1, \theta_2 \in \Theta$ and $\theta_1 \preccurlyeq \theta_2$.

The function α is assumed to be computable via simulation. In particular, given a computational budget c (measured in terms of total computer time expended) and points $\theta_1, \theta_2, \ldots, \theta_m \in \Theta$, we assume that there exist estimators $\alpha_c(\theta_1), \ldots, \alpha_c(\theta_m)$ such that

$$\alpha_c(\theta_i) \Rightarrow \alpha(\theta_i) \tag{1}$$

as $c \to \infty$ for $1 \leq i \leq m$ where $\Rightarrow$ denotes "weak convergence" or, equivalently, "convergence in distribution". In fact, we impose the stronger requirement that $(\alpha_c(\theta_1), \ldots, \alpha_c(\theta_m))$ satisfy a joint central limit theorem (CLT):

A1. $$c^{1/2}(\alpha_c(\theta_1) - \alpha(\theta_1), \ldots, \alpha_c(\theta_m) - \alpha(\theta_m))^T \Rightarrow N(\overrightarrow{0}, \Gamma)$$

as $c \to \infty$, where $N(\overrightarrow{0}, \Gamma)$ is an m-dimensional normal random vector with mean $\overrightarrow{0}$ and covariance matrix Γ with (i, j)'th entry given by $\Gamma(\theta_i, \theta_j)$ for $1 \leq i, j \leq m$.

Of course, A1 clearly implies (1). Note that if the simulations at the points $\theta_1, \ldots, \theta_m$ are independently generated, then Γ is a diagonal matrix. On the other hand, if the m simulations are correlated via use of common random numbers, Γ will have non-zero off-diagonal entries.

Assumption A1 covers both the case where $\alpha_c(\theta_i)$ is a sample mean obtained by averaging the results of multiple independent and identically distributed (iid) simulations at the point θ_i (as occurs when $\alpha(\theta_i)$ can be computed via a terminating simulation) and the case where $\alpha_c(\theta_i)$ is a time-average of a single replication of a stochastic model (as is the case when $\alpha(\theta_i)$ is a steady-state expectation).

Our second assumption asserts that the covariance matrix Γ can be consistently estimated from the simulated data associated with computational budget c.

A2. There exists an $m \times m$ matrix Γ_c (with (i, j)'th entry given by $\Gamma_c(\theta_i, \theta_j)$) such that

$$\Gamma_c \Rightarrow \Gamma$$

as $c \to \infty$.

In addition, we assume that the simulations at the points $\theta_1, \theta_2, \ldots, \theta_m$ are not trivially dependent.

A3. The covariance matrix Γ is positive definite.

We now proceed to describe our basic approach to computing $\overrightarrow{\alpha}(\theta) \triangleq (\alpha(\theta_1), \ldots, \alpha(\theta_m))^T$ when $\alpha(\cdot)$ is known to be non-decreasing in the partial order $\preccurlyeq$. Let $\mathcal{R} = \{(\theta_i, \theta_j) : \theta_i \preccurlyeq \theta_j\}$ be the set of pairs of points that are partially ordered under $\preccurlyeq$. Assumption A1 asserts that the random vector $\alpha_c(\theta) \triangleq (\alpha_c(\theta_1), \ldots, \alpha_c(\theta_m))^T$ is approximately Gaussian with mean $\overrightarrow{\alpha}(\theta)$ and covariance matrix $c^{-1}\Gamma$. This suggests that we estimate $\overrightarrow{\alpha}(\theta)$ as the maximizer of the likelihood

$$(2\pi)^{-m/2}|\det\Gamma|^{-1/2} \times \exp\left(-(\alpha_c(\theta) - z)^T \Gamma^{-1}/2(\alpha_c(\theta) - z)\right) \tag{2}$$

over $z = (z_1, \ldots, z_m)^T \in R^m$. If no monotonicity constraints are imposed, then the maximum likelihood estimator is, of course, just $\widehat{\alpha}_c(\theta) = \alpha_c(\theta)$. However, when $\alpha(\cdot)$ is assumed a priori to be monotone, the likelihood (2) should be maximized subject to the monotonicity constraints

$$z_i \leq z_j \qquad \text{for } (\theta_i, \theta_j) \in \mathcal{R}.$$

Because the logarithm of the likelihood is maximized at the same point as the likelihood itself, this leads to the constrained minimization problem

$$\begin{aligned} \min_{z=(z_1,\ldots,z_m)^T} \quad & (\alpha_c(\theta) - z)^T \Gamma^{-1}/2(\alpha_c(\theta) - z) \\ s/t \quad & z_i \leq z_j,\ (\theta_i, \theta_j) \in \mathcal{R}. \end{aligned} \tag{3}$$

This optimization problem involves minimizing a strictly convex quadratic objective function, subject to linear inequality constraints, and can be effectively solved numerically; see, for example, Gill et al. (2006). Of course, in practice, the simulationist does not know Γ; only the estimate Γ_c for Γ is generally available. Hence, our general approach for computing a monotone function $\alpha(\cdot)$ via simulation involves solving the quadratic program

$$\begin{aligned} \min_{z=(z_1,\ldots,z_m)^T} \quad & (\alpha_c(\theta) - z)^T \Gamma_c^{-1}(\alpha_c(\theta) - z) \\ s/t \quad & z_i \leq z_j,\ (\theta_i, \theta_j) \in \mathcal{R}. \end{aligned} \tag{4}$$

The minimizer of (4) is known to exist uniquely; see Frank and Wolfe (1956). The unique minimizer, call it $\widehat{\alpha}_c(\theta)$, is then our recommended estimator of $\overrightarrow{\alpha}(\theta)$. In contrast to the estimator $\alpha_c(\theta)$ (that we shall henceforth refer to as the "conventional estimator"), $\widehat{\alpha}_c(\theta)$ is guaranteed to possess the desired monotonicity property. Related estimators have been previously considered in the statistics literature (specifically, when Γ is diagonal), under the term "isotonic regression". We will follow this practice, and refer to $\widehat{\alpha}_c(\theta)$ as the "isotonic (regression) estimator" for $\overrightarrow{\alpha}(\theta)$.

Before proceeding further, we establish that the estimator $\widehat{\alpha}_c(\theta)$ is consistent as an estimator of $\overrightarrow{\alpha}(\theta)$.

Proposition 1 *Under A1–A3, $\widehat{\alpha}_c(\theta) \Rightarrow \alpha(\theta)$ as $c \to \infty$.*

Proof Because $\overrightarrow{\alpha}(\theta)$ is feasible for (4), it follows that

$$\left(\alpha_c(\theta)-\widehat{\alpha}_c(\theta)\right)^T\Gamma_c^{-1}\left(\alpha_c(\theta)-\widehat{\alpha}_c(\theta)\right) \le \left(\alpha_c(\theta)-\overrightarrow{\alpha}(\theta)\right)^T\Gamma_c^{-1}\left(\alpha_c(\theta)-\overrightarrow{\alpha}(\theta)\right).$$

But

$$\begin{aligned}&\left(\alpha_c(\theta)-\overrightarrow{\alpha}(\theta)\right)^T\Gamma_c^{-1}\left(\alpha_c(\theta)-\overrightarrow{\alpha}(\theta)\right)\\ &\quad= \left(\alpha_c(\theta)-\overrightarrow{\alpha}(\theta)\right)^T\Gamma^{-1}\left(\alpha_c(\theta)-\overrightarrow{\alpha}(\theta)\right) \qquad (5)\\ &\qquad+\left(\alpha_c(\theta)-\overrightarrow{\alpha}(\theta)\right)^T\left(\Gamma_c^{-1}-\Gamma^{-1}\right)\left(\alpha_c(\theta)-\overrightarrow{\alpha}(\theta)\right).\end{aligned}$$

The first term on the right-hand side of (5) goes to zero (weakly) because $\alpha_c(\theta)\Rightarrow\overrightarrow{\alpha}(\theta)$ as $c\to\infty$ (by A1). On the other hand, $\Gamma_c^{-1}\Rightarrow\Gamma^{-1}$ as $c\to\infty$ (due to A2 and the fact that a matrix inverse is a continuous function of the matrix entries) and $(\alpha_c(\theta)-\overrightarrow{\alpha}(\theta))$ is bounded in probability as $c\to\infty$ (by virtue of A1). Consequently, the second term on the right-hand side of (5) tends to zero (weakly) as $c\to\infty$. It follows that the left-hand side of (5) goes to zero (weakly) as $c\to\infty$, thereby implying that $\alpha_c(\theta)\Rightarrow\alpha(\theta)$ as $c\to\infty$ (due to the positive definiteness of Γ ensured by A3). □

Of course, the conventional estimator is also consistent as an estimator of $\overrightarrow{\alpha}(\theta)$. This raises the question of whether the isotonic $\widehat{\alpha}_c(\theta)$ is superior to $\alpha_c(\theta)$ and, if so, in what ways. One result that suggests that the isotonic estimator will frequently outperform the conventional estimator is the following result due to Theorem 1.1 of Barlow et al. (1972).

Theorem 1 *Suppose $d=1$ and Γ is diagonal (so that the estimators $\alpha_c(\theta_i)$ for $1\le i\le m$ are independent). Then, the minimizer $\widetilde{\alpha}_c(\theta)$ of (3) satisfies*

$$\begin{aligned}&E[\left(\widetilde{\alpha}_c(\theta)-\overrightarrow{\alpha}(\theta)\right)^T\Gamma^{-1}\left(\widetilde{\alpha}_c(\theta)-\overrightarrow{\alpha}(\theta)\right)]\\ &\le E[\left(\alpha_c(\theta)-\overrightarrow{\alpha}(\theta)\right)^T\Gamma^{-1}\left(\alpha_c(\theta)-\overrightarrow{\alpha}(\theta)\right)]. \qquad (6)\end{aligned}$$

Note that

$$\begin{aligned}&E[\left(\widetilde{\alpha}_c(\theta)-\overrightarrow{\alpha}(\theta)\right)^T\Gamma^{-1}\left(\widetilde{\alpha}_c(\theta)-\overrightarrow{\alpha}(\theta)\right)]\\ &=\sum_{i=1}^m\frac{1}{\Gamma(\theta_i,\theta_i)}E[(\widetilde{\alpha}_c(\theta_i)-\alpha(\theta_i))^2];\end{aligned}$$

a similar equality holds for the right-hand side of (6). Hence, (6) establishes that $\widetilde{\alpha}_c(\theta)$ has smaller "total mean square error" than does the conventional estimator. Since we expect the minimizers of (3) and (4) to be (very) close, this suggests that $\widehat{\alpha}_c(\theta)$ will often outperform the conventional estimator (in terms of total mean square error).

In fact, a stronger result can be established.

Theorem 2 *Suppose that $d=1$ and $\alpha_c(\theta)$ is multivariate normal with Γ being diagonal. Then, the minimizer $\widetilde{\alpha}_c(\theta)$ of (3) satisfies*

$$E\left[\left(\widetilde{\alpha}_c(\theta_i)-\alpha(\theta_i)\right)^2\right]\le E\left[\left(\alpha_c(\theta_i)-\alpha(\theta_i)\right)^2\right]$$

for $1\le i\le m$.

For a proof, see Lee (1981). This result asserts that when the simulation produces (exactly) normally distributed simulation outputs that are independent across the m points, then $\widetilde{\alpha}_c(\theta_i)$ outperforms the conventional estimator at each point θ_i ($1\le i\le m$).

3 INDEPENDENT SIMULATIONS ACROSS THE POINTS

In this section, we focus on the case where the estimators $\alpha_c(\theta_i)$ ($1\le i\le m$) are independent, so that the covariance matrix Γ is known a priori to be diagonal. In this diagonal setting, there is an explicit solution to the optimization problem (4). Before proceeding further, we make some necessary definitions. Denote $\{\theta_1,\dots,\theta_m\}$ by Θ_m. A subset L of Θ_m is a lower set with respect to the partial order $\preccurlyeq$ if $\theta_1\in L$, $\theta_2\in\Theta_m$, $\theta_2\preccurlyeq\theta_1$ imply $\theta_2\in L$. A subset U of Θ_m is an upper set if $\theta_1\in U$, $\theta_2\in\Theta_m$, $\theta_2\succcurlyeq\theta_1$ imply $\theta_2\in U$. The explicit solution to (4) can be written as

$$\widehat{\alpha}_c(\theta_i)=\min_{\substack{L\\ \theta_i\in L\cap U}}\max_{U}\frac{\sum_{\theta_j\in L\cap U}\alpha_c(\theta_j)\Gamma_c^{-1}(\theta_j,\theta_j)}{\sum_{\theta_j\in L\cap U}\Gamma_c^{-1}(\theta_j,\theta_j)}, \qquad (7)$$

where L is a lower set and U is an upper set of Θ_m. See Theorem 2.8 of Barlow et al. (1972) for details. This estimator can also be computed via the Pool-Adjacent-Violators (PAV) algorithm due to Ayers et al. (1955).

Here we briefly describe one variant of the PAV algorithm, called the Minimum Lower Sets algorithm; see Barlow et al. (1972) for details. For any subset B of Θ_m, let

$$\kappa(B)\triangleq\frac{\sum_{\theta_j\in B}\alpha_c(\theta_j)\Gamma_c^{-1}(\theta_j,\theta_j)}{\sum_{\theta_j\in B}\Gamma_c^{-1}(\theta_j,\theta_j)}.$$

Minimum Lower Sets Algorithm

1. Select a lower set B_1 that achieves the minimum value of κ and let $\widehat{\alpha}_c(\theta_j)=\kappa(B_1)=\min\{\kappa(L): L \text{ is a lower set}\}$ for all $\theta_j\in B_1$. If more than one lower set has the same lowest value of κ, take their union. Now we have determined $\widehat{\alpha}_c(\cdot)$ for all $\theta_j\in B_1$.
2. Consider sets of the form $L\cap B_1^c$ where L is a lower set. Select the largest level set B_2 that achieves the minimum value of κ among those and let $\widehat{\alpha}_c(\theta_j)=\kappa(B_2)=\min\{\kappa(L\cap B_1^c): L \text{ is a lower set}\}$ for all $\theta_j\in B_2$.

3. Go to 2 and continue this process until $\widehat{\alpha}_c(\cdot)$ is determined for all elements in Θ_m.

Suppose that the function $\alpha(\cdot)$ is strictly increasing in the partial order $\preccurlyeq$ (so that $\alpha(\theta_i) < \alpha(\theta_j)$ whenever $(\theta_i, \theta_j \in \mathcal{R})$). We then expect the monotonicity constraints in the quadratic program (4) to not be binding for large values of the computer budget c. As a consequence, the solution to (4), for large c, should be identical to that associated with the unconstrained version of (4), namely the estimator $\alpha_c(\theta)$.

Proposition 2 *Suppose that $\alpha(\cdot)$ is strictly increasing. Under A1,*

$$P(\widehat{\alpha}_c(\theta) = \alpha_c(\theta)) \to 1$$

as $c \to \infty$.

Note that Proposition 2 holds even when Γ is non-diagonal (as occurs when common random numbers are used across the m points). A simple consequence of Proposition 2 is that if $\alpha(\cdot)$ is strictly increasing, then

$$c^{1/2}\left(\widehat{\alpha}_c(\theta) - \overrightarrow{\alpha}(\theta)\right) \Rightarrow N(0, \Gamma)$$

as $c \to \infty$, so that $\widehat{\alpha}_c(\theta)$ enjoys the same asymptotic behavior as does the estimator $\alpha_c(\theta)$ in this setting.

Proof Note that if $\alpha_c(\theta)$ is feasible for (4), then $\widehat{\alpha}_c(\theta) = \alpha_c(\theta)$. Hence, the result follows if $P(\alpha_c(\theta_i) \le \alpha_c(\theta_j)) \to 1$ for each $(\theta_i, \theta_j) \in \mathcal{R}$. But this is a consequence of the fact that A1 implies that $\alpha_c(\theta_i) \Rightarrow \alpha(\theta_i)$ as $c \to \infty$ for $1 \le i \le m$. □

Of course, if the number m of points is large, the computer budget c required for $\alpha_c(\theta)$ to become feasible for (4) may be (extremely) large. Thus, perhaps the most interest in our proposed isotonic regression estimator arises when m is large (so that $\widehat{\alpha}_c(\theta)$ does not coincide with the standard estimator $\alpha_c(\theta)$). When m is large and $\alpha(\theta)$ is differentiable, the difference $\alpha(\theta_i) - \alpha(\theta_j)$ will tend to be small for (θ_i, θ_j) pairs in $\mathcal{R}$ for which $\|\theta_i - \theta_j\|$ is small. As a means of understanding the behavior of the isotone estimator in such settings, we now consider the behavior of $\widehat{\alpha}_c(\theta)$ when $\alpha(\theta_i) = \alpha(\theta_j)$ for some pairs $(\theta_i, \theta_j) \in \mathcal{R}$.

Theorem 3 *Assume A1–A3 and suppose that $\{\alpha_c(\theta_i) : 1 \le i \le m\}$ is a collection of independent rv's. Then, for $1 \le i \le m$,*

$$c^{1/2}\left(\widehat{\alpha}_c(\theta_i) - \alpha(\theta_i)\right) \Rightarrow W(\theta_i)$$

as $c \to \infty$. Furthermore, if $d = 1$,

$$EW^2(\theta_i) \le \Gamma(\theta_i, \theta_i).$$

The above result shows that, when θ is real-valued, the isotonic estimator has a smaller asymptotic mean square error than does the estimator $\alpha_c(\theta_i)$ (in the independent setting).

Proof It follows from (7) that

$$c^{1/2}\left(\widehat{\alpha}_c(\theta_i) - \alpha(\theta_i)\right) = \min_{\substack{L \\ \theta_i \in L \cap U}} \max_{U} \frac{\sum_{\theta_j \in L \cap U} c^{1/2}(\alpha_c(\theta_j) - \alpha(\theta_i))\Gamma_c^{-1}(\theta_j, \theta_j)}{\sum_{\theta_j \in L \cap U} \Gamma_c^{-1}(\theta_j, \theta_j)}.$$

Note that if $\theta_j \succcurlyeq \theta_i$ with $\alpha(\theta_j) > \alpha(\theta_i)$, then $c^{1/2}(\alpha_c(\theta_j) - \alpha(\theta_i)) \Rightarrow +\infty$ as $c \to \infty$. Similarly, if $\theta_j \preccurlyeq \theta_i$ with $\alpha(\theta_j) < \alpha(\theta_i)$, then $c^{1/2}(\alpha_c(\theta_j) - \alpha(\theta_i)) \Rightarrow -\infty$ as $c \to \infty$. As a consequence,

$$c^{1/2}(\widehat{\alpha}_c(\theta_i) - \alpha(\theta_i)) = \min_{\substack{L \\ \theta_i \in L \cap U}} \max_{U} \left\{ \frac{\sum_{\theta_j \in L \cap U \cap \mathcal{A}_i} c^{1/2}(\alpha_c(\theta_j) - \alpha(\theta_i))\Gamma_c^{-1}(\theta_j, \theta_j)}{\sum_{\theta_j \in L \cap U \cap \mathcal{A}_i} \Gamma_c^{-1}(\theta_j, \theta_j)} \right\} \tag{8}$$

with probability converging to one as $c \to \infty$, where $\mathcal{A}_i = \{\theta_j : \alpha(\theta_j) = \alpha(\theta_i),\ 1 \le j \le m\}$. The right-hand side of (8) converges weakly (by the continuous mapping principle) to

$$\min_{\substack{L \\ \theta_i \in L \cap U}} \max_{U} \frac{\sum_{\theta_j \in L \cap U \cap \mathcal{A}_i} Z(\theta_j)\Gamma^{-1}(\theta_j, \theta_j)}{\sum_{\theta_j \in L \cap U \cap \mathcal{A}_i} \Gamma^{-1}(\theta_j, \theta_j)} \tag{9}$$

as $c \to \infty$, where the $Z(\theta_j)$'s are independent normal rv's with mean zero and variance $\Gamma(\theta_i, \theta_i)$.

Call the right-hand side of (9) $W(\theta_i)$. Note that any lower set L_i in $\mathcal{A}_i$ can be written as $L \cap \mathcal{A}_i$ where L is the smallest lower set of Θ_m containing L_i. Conversely, any set of the form $L \cap \mathcal{A}_i$ where L is a lower set of Θ_m is a lower set of $\mathcal{A}_i$. Hence $\{L_i : L_i \text{ is a lower set of } \mathcal{A}_i\} = \{L \cap \mathcal{A}_i : \text{L is a lower set of } \Theta_m\}$. A similar argument applies to upper sets of $\mathcal{A}_i$. Hence, $W(\theta_i)$ is precisely the isotone mapping (7) as applied to the collection of independent mean zero Gaussian rv's $(Z(\theta_j) : \theta_j \in \mathcal{A}_i)$. When $d = 1$, we can now apply Lee (1981) to conclude that $EW^2(\theta_i) \le \Gamma(\theta_i, \theta_i)$. □

4 ISOTONE REGRESSION FOR TWO POINTS

In this setting, we consider isotonic regression in the special setting of two points θ_1 and θ_2 for which $\theta_1 \preccurlyeq \theta_2$. In this two point context, we can explicitly write down the solution to (4) when Γ is non-diagonal (thereby permitting us to analyze the case in which correlation across the points is induced by use of some variance reduction method).

When $m = 2$, we write Γ_c in the form

$$\Gamma_c = \begin{pmatrix} \sigma_1^2(c) & \rho(c)\sigma_1(c)\sigma_2(c) \\ \rho(c)\sigma_1(c)\sigma_2(c) & \sigma_2^2(c) \end{pmatrix},$$

so that $\sigma_i^2(c)$ is an estimator for the (asymptotic) variance of $\alpha_c(\theta_i)$ for $i=1,2$ and $\rho(c)$ is an estimator of the (asymptotic) coefficient of correlation between $\alpha_c(\theta_1)$ and $\alpha_c(\theta_2)$.

The optimization problem (4) then can be re-written as the minimization of a quadratic function subject to a single inequality constraint:

$$\min_{z_1,z_2}\left[\frac{(\alpha_c(\theta_1)-z_1)^2}{\sigma_1^2(c)}+\frac{(\alpha_c(\theta_2)-z_2)^2}{\sigma_2^2(c)}\right. \left.-2(\alpha_c(\theta_1)-z_1)(\alpha_c(\theta_2)-z_2)\frac{\rho(c)}{\sigma_1(c)\sigma_2(c)}\right]\frac{1}{1-\rho(c)^2} \quad s/t \quad z_1\le z_2. \tag{10}$$

The solution can be explicitly written down. In particular,

$$\widehat{\alpha}_c(\theta_1)=\begin{cases}\alpha_c(\theta_1) & ,\ \text{if } \alpha_c(\theta_1)\le\alpha_c(\theta_2)\\ \widehat{a}\alpha_c(\theta_1)+(1-\widehat{a})\alpha_c(\theta_2) & ,\ \text{otherwise}\end{cases}$$

and

$$\widehat{\alpha}_c(\theta_2)=\begin{cases}\alpha_c(\theta_2) & ,\ \text{if } \alpha_c(\theta_1)\le\alpha_c(\theta_2)\\ \widehat{a}\alpha_c(\theta_1)+(1-\widehat{a})\alpha_c(\theta_2) & ,\ \text{otherwise}\end{cases}$$

where

$$\widehat{a} = \left(\frac{1}{\sigma_1^2(c)}-\frac{\rho(c)}{\sigma_1(c)\sigma_2(c)}\right)\cdot \left(\frac{1}{\sigma_1^2(c)}-\frac{2\rho(c)}{\sigma_1(c)\sigma_2(c)}+\frac{1}{\sigma_2^2(c)}\right)^{-1}.$$

As noted in Section 3, $(\widehat{\alpha}_c(\theta_1),\widehat{\alpha}_c(\theta_2))$ has identical asymptotic behavior to that of $(\alpha_c(\theta_1),\alpha_c(\theta_2))$ when $\alpha(\theta_1)<\alpha(\theta_2)$ (even when $\rho\neq 0$). We therefore focus on the special case in which $\alpha(\theta_1)=\alpha(\theta_2)$.

Theorem 4 *Suppose that $\alpha(\theta_1)=\alpha(\theta_2)$. If A1–A3 hold, then*

$$c^{1/2}(\widehat{\alpha}_c(\theta_i)-\alpha(\theta_i))\Rightarrow W(\theta_i)$$

for $i=1,2$ as $c\to\infty$, where

$$EW^2(\theta_i)\le\Gamma(\theta_i,\theta_i)\ (=\sigma_i^2)$$

for $i=1,2$.

In other words, the above result establishes that the isotonic estimator reduces mean square error relative to $\alpha_c(\theta)$, even in the presence of correlation between the estimators at the two points.

Proof Set $\beta_c(\theta_i)=c^{1/2}(\alpha_c(\theta_i)-\alpha(\theta_i))$ and note that

$$\begin{aligned}&c^{1/2}(\widehat{\alpha}_c(\theta_1)-\alpha(\theta_1))\\ &\quad= \beta_c(\theta_1)I(\beta_c(\theta_1)\le\beta_c(\theta_2))\\ &\qquad+(\widehat{a}\beta_c(\theta_1)+(1-\widehat{a})\beta_c(\theta_2))I(\beta_c(\theta_1)>\beta_c(\theta_2)).\end{aligned}$$

Then,

$$\begin{aligned}&c^{1/2}(\widehat{\alpha}_c(\theta_1)-\alpha(\theta_1))\\ &\quad\Rightarrow Z(\theta_1)I(Z(\theta_1)\le Z(\theta_2))\\ &\qquad+(aZ(\theta_1)+(1-a)Z(\theta_2))I(Z(\theta_1)>Z(\theta_2))\\ &\qquad\triangleq W(\theta_1)\end{aligned}$$

where

$$a=\frac{(1-\rho\sigma_1/\sigma_2)/\sigma_1^2}{(1-\rho\sigma_1/\sigma_2)/\sigma_1^2+(1-\rho\sigma_2/\sigma_1)/\sigma_2^2},$$

σ_i^2 is the (asymptotic) variance of $\alpha_c(\theta_i)$ for $i=1,2$, ρ is the (asymptotic) coefficient of correlation between $\alpha_c(\theta_1)$ and $\alpha_c(\theta_2)$, and $(Z(\theta_1),Z(\theta_2))$ has a bivariate $N(0,\Gamma)$ distribution. Because $I(Z(\theta_1)\le Z(\theta_2))$ and $I(Z(\theta_1)>Z(\theta_2))$ can not simultaneously be non-zero,

$$\begin{aligned}EW^2(\theta_1)&=EZ^2(\theta_1)I(Z(\theta_1)-Z(\theta_2)\le 0)\\ &\quad+E(aZ(\theta_1)+(1-a)Z(\theta_2))^2I(Z(\theta_1)-Z(\theta_2)>0).\end{aligned}$$

Let (Z_1,Z_2) be a bivariate mean zero Gaussian random vector. Because $(-Z_1,-Z_2)$ has the same mean and covariance structure as does (Z_1,Z_2), it follows that $(Z_1,Z_2)\stackrel{\mathcal{D}}{=}(-Z_1,-Z_2)$, where $\stackrel{\mathcal{D}}{=}$ denotes equality in distribution. Consequently,

$$\begin{aligned}&EZ_1^2I(Z_2\le 0)\\ &\quad=E(-Z_1)^2I(-Z_2\le 0)=EZ_1^2I(Z_2\ge 0)\end{aligned}$$

and so

$$EZ_1^2I(Z_2\le 0)=\frac{1}{2}EZ_1^2,$$

regardless of the correlation structure between Z_1 and Z_2. Hence,

$$EZ^2(\theta_1)I(Z(\theta_1)-Z(\theta_2)\le 0)=\frac{1}{2}EZ^2(\theta_1)=\frac{\sigma_1^2}{2}$$

and

$$E\left(aZ(\theta_1)+(1-a)Z(\theta_2)\right)^2 I\left(Z(\theta_1)-Z(\theta_2)>0\right)$$
$$=\frac{1}{2}a^2EZ^2(\theta_1)+\frac{1}{2}(1-a)^2EZ^2(\theta_2)$$
$$+a(1-a)E(Z(\theta_1)Z(\theta_2))$$
$$=\frac{1}{2}a^2\sigma_1^2+\frac{1}{2}(1-a)^2\sigma_2^2+a(1-a)\rho\sigma_1\sigma_2.$$

Let

$$g(\rho)\triangleq 2\left(a^2+(1-a)^2\right)\left(\Gamma(\theta_1,\theta_1)-EW^2(\theta_1)\right).$$

We will show $g(\rho)\geq 0$ for all $-1\leq\rho\leq 1$. Note that

$$g(\rho)=\left(\frac{1}{\sigma_2^2}-\frac{\rho}{\sigma_1\sigma_2}\right)^2(\sigma_1^2-\sigma_2^2)$$
$$+2\left(\frac{1}{\sigma_1^2}-\frac{\rho}{\sigma_1\sigma_2}\right)\left(\frac{1}{\sigma_2^2}-\frac{\rho}{\sigma_1\sigma_2}\right)(\sigma_1^2-\rho\sigma_1\sigma_2)$$

and $g'(\rho)$ vanishes at $\rho_1=\sigma_1/\sigma_2$ and $\rho_2=(2\sigma_1)/(3\sigma_2)+\sigma_2/(3\sigma)$. If $\sigma_1\leq\sigma_2$, then $0\leq\rho_1,\rho_2\leq 1$, $g(\rho_1)=0$, $g(\rho_2)\geq 0$, and $g(1)\geq 0$. Hence $g(\rho)\geq 0$ for all $-1\leq\rho\leq 1$. If $\sigma_1>\sigma_2$, then $\rho_1,\rho_2>1$, hence from $g(1)\geq 0$ it follows that $g(\rho)\geq 0$ for all $-1\leq\rho\leq 1$.

A similar argument establishes the result for $c^{1/2}(\widehat{\alpha}_c(\theta_2)-\alpha(\theta_2))$. □

This result depends crucially on the fact that the asymptotic distribution for $c^{1/2}(\alpha_c(\theta)-\overrightarrow{\alpha}(\theta))$ is Gaussian. The example below shows that $\widetilde{\alpha}_c(\theta)$ can have larger mean square error than $\alpha_c(\theta)$ if the distribution of $\alpha_c(\theta)$ is appropriately chosen. This suggests that $\widehat{\alpha}_c(\theta)$ can have worse mean square error than $\alpha_c(\theta)$ for small values of c.

Example 1 Suppose that $(\alpha_c(\theta_1),\alpha_c(\theta_2))$ takes values $(4,3)$, $(2,4)$, $(-4,-4)$, and $(-2,-3)$ with probability 1/4 each. Then $\sigma_1^2=10$, $\sigma_2^2=25/2$, $\rho=21/2$, and

$$E\left(\widetilde{\alpha}_c(\theta_1)-\alpha(\theta_1)\right)^2>E\left(\alpha_c(\theta_1)-\alpha(\theta_1)\right)^2. \;\square$$

We conclude this section by noting that $\widehat{\alpha}_c(\theta_1)$ can have unusually large bias when $\alpha(\theta_1)=\alpha(\theta_2)$. Note that

$$W(\theta_1)=Z(\theta_1)$$
$$+(1-a)\left(Z(\theta_2)-Z(\theta_1)\right)I\left(Z(\theta_1)>Z(\theta_2)\right)$$

so

$$EW(\theta_1)=\frac{(a-1)}{2}E|Z(\theta_2)-Z(\theta_1)|.$$

Hence, if $(c^{1/2}(\widehat{\alpha}_c(\theta_1)-\alpha(\theta_1)):c\geq 1)$ is uniformly integrable, evidently

$$E\widehat{\alpha}_c(\theta_1)=\alpha(\theta_1)$$
$$+c^{-1/2}\tfrac{(a-1)}{2}E|Z(\theta_2)-Z(\theta_1)|+o(c^{-1/2})$$

as $c\to\infty$, where $o(c^{-1/2})$ is a term tending to zero faster than $c^{-1/2}$ as $c\to\infty$. Thus, the isotonic estimator has a bias of order $c^{-1/2}$ for c large. This is to be contrasted with the bias of order c^{-1} that is more typical of Monte Carlo estimators; see, for example, Asmussen and Glynn (2007). However, despite the large bias, Theorem 3 shows that the isotonic estimator reduces asymptotic mean square error.

5 ISOTONIC REGRESSION FOR LARGE NUMBERS OF POINTS

As mentioned in Section 3, our main motivation for suggesting isotonic regression as an alternative means of computing a response surface arises when the number m of points is large. Of course, when functional estimation is a principal goal, the number m of points will often be large. Consequently, the asymptotic regime in which m is sent to infinity is of clear practical interest.

Here, we consider only the case in which the simulations at $\theta_1,\ldots,\theta_m$ are generated independently. We assume, throughout the remainder of this section, that $d=1$, $\Theta=[0,1]$, and c is fixed. We also assume that, for any $\theta\in\Theta$, $\alpha_c(\theta)$ is unbiased for $\alpha(\theta)$, so that $\mathrm{E}\alpha_c(\theta)=\alpha(\theta)$. Consistency and asymptotic behavior of the minimizer $\widetilde{\alpha}_c(\theta)$ of (3) were investigated by Brunk (1970) and presented below. Define $\widetilde{\alpha}_c(\cdot)$ by the linear interpolation of $\widetilde{\alpha}_c(\theta_i)$, $i=1,\ldots,m$, between two adjacent points.

Theorem 5 *Assume that $\alpha(\cdot)$ is continuous and $\sigma^2(\cdot)\triangleq var(\alpha_c(\cdot))$ is bounded. Then, for any $0<a<b<1$,*

$$P\left(\lim_{m\to\infty}\sup_{\lfloor am\rfloor\leq j\leq\lfloor bm\rfloor}|\widetilde{\alpha}_c(j/m)-\alpha(j/m)|=0\right)=1.$$

Theorem 6 *Let $\theta_j=j/m$, $j=1,\ldots,m$ and θ_0 be fixed. Suppose that $(\alpha_c^2(\theta):\theta\in[0,1])$ is a uniformly integrable family of rv's. Suppose that $\sigma^2(\cdot)$ and the derivative $\alpha'(\cdot)$ of $\alpha(\cdot)$ are continuous in a neighborhood of θ_0. Then*

$$\left(\frac{2m}{\sigma^2(\theta_0)\alpha'(\theta_0)}\right)^{\frac{1}{3}}(\widetilde{\alpha}_c(\theta_0)-\alpha(\theta_0))$$

converges in distribution to the slope at 0 of the greatest convex minorant of $W(t)-t^2$ where W is a standard two-sided Brownian motion originating from 0.

6 NUMERICAL RESULTS

In this section, we investigate the performance of our proposed isotonic estimator in the setting of a simple queueing model. Specifically, we let $\alpha(\theta) = EW_\infty(\theta)$, where $W_\infty(\theta)$ is the steady-state waiting time (exclusive of service) in M/M/1 queue with first in/first out (FIFO) queue discipline, unit arrival rate, and service rate equal to θ. For this model,

$$\alpha(\theta) = \frac{1}{\theta(\theta-1)};$$

clearly, $\alpha(\cdot)$ is strictly decreasing in θ. We consider two different computational strategies. In the first, a total of r customers is simulated across the m points, with r/m customers per point. With an equal number of customers per point, we can easily apply common random numbers in this setting (by using the same sequence of inter-arrival times across all values of θ, and by using a scaled version of a single service time sequence, scaled so that the mean service time at θ is $1/\theta$). For this simple model, the use of common random numbers forces $\alpha(\cdot)$ to be automatically monotone, so that the isotonic estimator, as applied to the common random numbers simulation output, coincides with the common random numbers estimator. So, let estimator 1 be the estimator of $\alpha(\cdot)$ when common random numbers are applied with r/m customers per point. Estimator 2 is the conventional estimator of $\alpha(\cdot)$ when independent simulations are conducted at each point (with r/m customers per point). Estimator 3 is the isotonic estimator of $\alpha(\cdot)$ obtained from estimator 2. The diagonal sample covariance matrix Γ_c of A2 is obtained by using the regenerative variance estimator at each θ value; see Bratley et al. (1987) for details.

The second computational strategy recognizes that the computation of $\alpha(\theta)$ increases in difficulty as θ decreases to 1, and allocates the computational budget accordingly. Because of the unequal sample sizes across the differing θ values, common random numbers can no longer be applied. To determine the appropriate number $n(\theta)$ of customers to simulate at service rate θ, we note that Whitt (1989) suggests choosing $n(\theta)$ in proportion to $\theta^2(1-1/\theta)^{-2}$. Hence, we select

$$n(\theta_i) = \frac{r\theta_i^2(1-1/\theta_i)^{-2}}{\sum_{j=1}^{m}\theta_j^2(1-1/\theta_j)^{-2}}.$$

Estimator 4 is the conventional estimator in which $n(\theta_i)$ customers are independently simulated at each θ_i, $1 \le i \le m$. Estimator 5 is the isotonic version of estimator 4 (Γ_c obtained using the regenerative variance estimator).

To measure the accuracy of the above estimators, we compute the normalized root mean square error (NRMSE) for each estimator at each θ value. The NRMSE was obtained by generating 600 independent replications of our five estimators, each computed with a total of $r = 10843$ customers suitably allocated across the ten different θ values $1.1, 1.2, \ldots, 2.0$. The NRMSE for a particular estimator at a given θ value is its sample root mean square error RMSE divided by $\alpha(\theta)$. Table 1 lists the NRMSE values for the five estimators at each of the ten different θ values. We include the maximum NRMSE value across θ (MNRMSE) as a summary statistic that characterizes each estimator's performance.

The table shows the superior performance of the isotonic versions (estimators 3 and 5) of estimators 2 and 4 across a wide range of θ values. Also, estimator 5 has the smallest MNRMSE. Finally, it should be noted that estimators 1, 2, and 3 exhibit a significantly broader range of NMRSE values than do estimators 4 and 5, in large part because the equal allocation of computational effort across θ undersamples those values of θ corresponding to the M/M/1 queue in "heavy traffic".

Table 1: NRMSE and MNRMSE for Estimator 1–5

	NRMSE										MNRMSE
θ	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	
Estimator 1	0.521	0.377	0.301	0.257	0.225	0.204	0.189	0.179	0.171	0.165	0.521
Estimator 2	0.465	0.369	0.288	0.271	0.222	0.185	0.185	0.169	0.171	0.163	0.465
Estimator 3	0.461	0.304	0.229	0.194	0.177	0.150	0.147	0.138	0.128	0.132	0.461
Estimator 4	0.286	0.308	0.301	0.335	0.303	0.255	0.268	0.246	0.260	0.228	0.335
Estimator 5	0.285	0.275	0.248	0.232	0.210	0.188	0.191	0.177	0.179	0.179	0.285

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis.* Springer-Verlag, to appear.

Ayer, M., H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman. 1955. An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics* 26:641–647.

Barlow, R. E., D. J. Bartholomew, J. M. Bremmer, and H. D. Brunk. 1972. *Statistical inference under order restrictions.* New York:Wiley.

Bauerle, N. 1997. Monotonicity results for MR/GI/1 queues. *Journal of Applied Probability* 34(2):514–524.

Bauerle, N., and T. Rolski. 1998. A monotonicity result for the workload in Markov-modulated queues. *Journal of Applied Probability* 35:741–747.

Berger, A. W., and W. Whitt. 1992. Comparison of multi-server queues with finite waiting rooms. *Stochastic Models* 8(4):719–732.

Bondi, A. B., and W. Whitt. 1986. The influence of service-time variability in a closed network of queues. *Performance Evaluation* 6:219–234.

Bratley, P., B. L. Fox, and L. E. Schrage. 1983. *A guide to simulation.* New York: Springer Verlag.

Brunk, H. D. 1958. On the estimation of parameters restricted by inequalities. *Annals of Mathematical Statistics.* 29:437–454.

Brunk, H. D. 1970. Estimation of isotonic regression. *Nonparametric techniques in statistical inference*. Cambridge, 177–195.

Budka, K. C., and D. D. Yao. 1990. Monotonicity and convexity properties of rate control throttles. In *Proceedings of 29th IEEE conference on Decision and Control*, 883–884.

Chang, C. S., X. Chao, and M. Pinedo. 1990. Monotonicity results for queues with doubly stochastic Poisson arrivals: Ross' conjecture. *Advances in Applied Probability* 23:210-228.

Chang, C., X. Chao, M. Pinedo, and J. Shanthikumar. 1991. Stochastic convexity for multidimensional processes and its applications. *IEEE Transactions on Automatic Control* 36(12):1347–1355.

Frank, M., and P. Wolfe. 1956. An algorithm for quadratic programming. *Naval Research Logistics Quarterly* 3:95–110.

Gill, P. E., W. Murray, and M. A. Saunders. 2006. User's guide for SQOPT version 7: software for large-scale linear and quadratic programming. Available online via `<http://www.cam.ucsd.edu/~peg/sqdoc7.pdf>` [accessed August 7, 2006].

Jean-Marie, A., and L. Gun. 1993. Parallel queues with resequencing. *Journal of the ACM* 40(6):1188–1208.

Law, A. M., and W. D. Kelton. 2000. *Simulation modelling and analysis.* McGraw-Hill, Inc.

Lee, C. C. 1981. The Quadratic loss of isotonic regression under normality. *Annals of Statistics* 9(3):686–688.

Shaked, M., and J. G. Shanthikumar. 1988. Stochastic convexity and its applications. *Advances in Applied Probability* 20:427–446.

Shen, A., and N. K. Vereshchagin. 2002. *Basic set theory.* American Mathematical Society. Student mathematical library, Vol. 17.

Van Oyen, M. 1997. Monotonicity of optimal performance measures for polling systems. *Probability in Engineering and Information Science* 11(2):219–228.

Weber, R. R. 1983. A note on waiting times in single server queues. *Operations Research* 31:950–951.

Whitt, W. 1983. Comparison conjectures about the M/G/s queue. *Operations Research Letters* 2(5): 203–208.

Whitt, W. 1989. Planning queueing simulations. *Management Science* 35(11):1341–1366.

Ziya, S., H. Ayhan, and R. D. Foley. 2004. A monotonicity result for the blocking probability in a G/GI/c/m queueing system. Available online via `<http://www2.isye.gatech.edu/people/faculty/Hayriye_Ayhan/block-note.pdf>` [accessed August 7, 2006].

AUTHOR BIOGRAPHIES

EUNJI LIM is a Ph.D. student in the Department of Management Science and Engineering at Stanford University. Her research interests include statistical inference under shape restrictions, stochastic optimization, simulation and applied probability.

PETER W. GLYNN received his Ph.D in Operations Research from Stanford University in 1982. He then joined the faculty of the University of Wisconsin at Madison, where he held a joint appointment between the Industrial Engineering Department and Mathematics Research Center, and courtesy appointments in Computer Science and Mathematics. In 1987, he returned to Stanford, where he is now the Thomas Ford Professor of Engineering in the Department of Management Science and Engineering. He also has a courtesy appointment in the Department of Electrical Engineering. He is a Fellow of the Institute of Mathematical Statistics and he has research interests in computational probability, queueing theory, statistical inference for stochastic processes, and stochastic modeling.