Response Surface Estimation

Peter W. Glynn

Stanford University

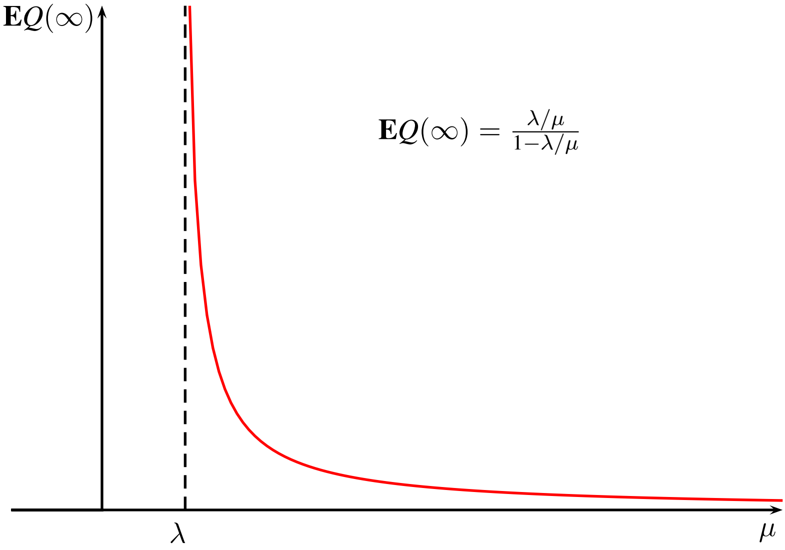
2012 NSF Workshop on Simulation Methodology, Shanghai, July 23, 2012

The Basic Problem:

View the simulation model as a function that sends input parameters to some output (performance) measure

Goal: Estimate this function

Example: How does the expected number-in-system behave as a function of the service rates in a queueing network?


$\mathbf{E}Q(\infty)$
$\mathbf{E}Q(\infty) = \frac{\lambda/\mu}{1-\lambda/\mu}$
λ
μ

Why the problem is of interest:

- One is dealing with a system in which one expects significant variability in the inputs to the system (e.g. arrival rates). One needs a system design that performs reasonably well across a wide range of parameter values.

- One needs a functional estimate as an input to a real-time decision tool in which the input value is not known a priori (e.g. computing the price of an option across a range of prices of the underlying asset).

- Stochastic models are often used not as predictive tools but to generate qualitative insight into a system. Understanding how a system behaves as a function of the input variables is one source of such insight.

- Response surface estimation can be a useful first step in studying various features of the system (e.g. optimization)

  Screening methods
  Many optimization methods implicitly estimate the response surface (at least locally in a neighborhood of the optimizer)

Why the problem is difficult:

The object to be computed is an "infinite-dimensional" (or, at least, high-dimensional) quantity.

Outline of this Talk:

- Parametric Modeling
- Nonparametric Methods
    - Common Random Numbers
    - Change-of-measure
    - Orthogonal function approximations
    - Bayesian methods
    - Shape-constrained estimation

Parametric Modeling:

Goal: Compute $\alpha(\theta) = \mathbf{E}X(\theta), \quad \theta \in \Lambda \subseteq \mathbb{R}^d$

Method:

0. Choose a finite-dimensional approximation to $\alpha(\cdot)$
   e.g. $\alpha(\beta, \theta) = \beta_0 + \sum_{i=1}^{d} \beta_i \theta_i + \sum_{i,j=1}^{d} \beta_{ij} \theta_i \theta_j$

1. Choose $\theta_1, \theta_2, \cdots, \theta_m \in \Lambda$

2. Estimate $\alpha(\theta_1), \alpha(\theta_2), \cdots, \alpha(\theta_m)$ via $\overline{X}_n(\theta_1), \ldots, \overline{X}_m(\theta_m)$, where
$$\overline{X}_n(\theta_i) = \frac{1}{n} \sum_{j=1}^{n} X_j(\theta_i)$$

3. Solve the least squares problem
$$\min_{\beta} \sum_{i=1}^{m} \left(\overline{X}_n(\theta_i) - \alpha(\beta, \theta_i)\right)^2;$$
   call the minimizer $\widehat{\beta}_{n,m}$.

4. Approximate $\alpha(\cdot)$ via $\alpha(\widehat{\beta}_{n,m}, \cdot)$.

Remark:

Suppose that

$$n^{1/2}\left(\overline{X}_n(\theta_i) - \alpha(\beta^*, \theta_i) : 1 \le i \le m\right) \Rightarrow N(0, C)$$

as $n \to \infty$. The "correct least squares problem" is:

$$\min_{\beta}\left(\overline{X}_n - \alpha(\beta, \cdot)\right)^T C^{-1}\left(\overline{X}_n - \alpha(\beta, \cdot)\right)$$

Nonparametric Methods

- Note that the natural "model-free" estimator for $\alpha(\theta)$ is

$$\overline{X}_n(\theta),$$

  where $(X_j(\theta) : j \geq 1)$ is iid.

- But, in the Monte Carlo setting, we have the freedom to choose the joint distribution $(X_j(\theta) : \theta \in \Lambda)$ to our advantage.

- A natural joint distribution is to simulate $(X_j(\theta) : \theta \in \Lambda)$ using *common random numbers* across θ

Common Random Numbers

Feed the system with common input sequences

e.g. Markov chains / stochastic recursions

$$\begin{aligned} Y_{l+1}(\theta) &= \tilde{r}(Y_l(\theta), Z_{l+1}(\theta)) \\ &= r(\theta, Y_l(\theta), Z_{l+1}) \end{aligned}$$

$$X(\theta) = f(Y_l(\theta) : 0 \le l \le t)$$

Single-server queue waiting time sequence:

$$\begin{aligned} W_{l+1}(\theta) &= [W_l(\theta) + F_V^{-1}(\theta, \widetilde{U}_l) - \chi_{l+1}]^+ \\ &= [W_l(\theta) + \theta V_l - \chi_{l+1}]^+ \end{aligned}$$

As a function of θ:

- $W_l(\cdot)$ is convex and non-decreasing
- $W_l(\cdot)$ is (typically) differentiable and

$$\frac{d}{d\theta}\mathbf{E}W_l(\theta) = \mathbf{E}\frac{d}{d\theta}W_l(\theta) \qquad \text{(IPA)}$$

Why is the use of CRN advantageous?

$$\begin{aligned}
&\mathrm{var}[X(\theta+h) - X(\theta)] \\
&= \mathrm{var}X(\theta+h) + \mathrm{var}X(\theta) - 2\underbrace{\mathrm{cov}(X(\theta), X(\theta+h))}_{\text{depends on joint distribution}} \\
&\leq \mathrm{var}X(\theta+h) + \mathrm{var}X(\theta)
\end{aligned}$$

if $\mathrm{cov}(X(\theta), X(\theta+h)) \geq 0$.

This follows if:

- $X(\theta)$ is non-decreasing in the inputs (e.g. the Z_i's in the Markov chain setting) for each θ
  Caveat: Rarely holds in the "exact" sense
- $X(\cdot)$ is continuous in probability and h is small:

  $$X(\theta+h) \xrightarrow{P} X(\theta) \qquad \text{as} \quad h \downarrow 0$$

  implies that

  $$\mathrm{cov}(X(\theta), X(\theta+h)) \to \mathrm{var}X(\theta) \geq 0 \qquad \text{as} \quad h \downarrow 0$$

  This holds in great generality; so we can expect reasonable "local behavior"

More on smoothness of $X(\theta)$:

Unless we apply CRN really poorly, we can almost always expect that $X(\cdot)$ is continuous in probability.

Can we expect more?

- In some (limited) settings:
  - $X(\cdot)$ is a.s. monotone or convex
- In other settings, $X(\cdot)$ is a.s. defined in θ and

$$\mathbf{E}X'(\theta) = \frac{d}{d\theta}\mathbf{E}X(\theta)$$

When this occurs,

$$X(\theta + h) - X(\theta) \approx hX'(\theta)$$

and (typically)

$$\mathrm{var}[X(\theta + h) - X(\theta)] = O(h^2)$$

- For a Poisson process,

$$\mathrm{var}[N(\theta + h) - N(\theta)] = \lambda h = O(h)$$

For M/M/1 number-in-system process:

$$\mathrm{var}[X(\theta + h) - X(\theta)] = O(h)$$

This behavior likely holds in great generality for discrete-event simulations

Implications of Use of CRN:

CRN guarantees that the response surface is globally defined.

There are many ways to assess the quality of a response surface:

- Integrated Mean Square (L^2) Error:

$$\mathbf{E}\int_{\Lambda}(\alpha_n(\theta)-\alpha(\theta))^2 d\theta$$

- Worst Case Error:

$$\sup_{\theta\in\Lambda}|\alpha_n(\theta)-\alpha(\theta)|$$

- Implications for optimization: e.g. How close is optimizer / optimum of $\alpha_n(\cdot)$ to optimizer/optimum of $\alpha(\cdot)$?

Use of CRN's (Typical Case)

In the "typical" case,

$$\mathrm{var}[X(\theta+h)-X(\theta)] = O(h) \qquad \text{as} \quad h \downarrow 0$$

Then,

$$\alpha_n(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i(\theta)$$

satisfies

$$n^{1/2}(\alpha_n(\theta)-\alpha(\theta)) \Rightarrow Z(\theta)$$

as $n \to \infty$, where $(Z(\theta) : \theta \in \Lambda)$ is a mean zero Gaussian random field with

$$\mathrm{cov}(Z(\theta_1), Z(\theta_2)) = \mathrm{cov}(X(\theta_1), X(\theta_2)).$$

In the "typical case", $Z(\cdot)$ is a continuous field but not differentiable a.e.

Note that

$$\mathbf{E}\int_{\Lambda} |\alpha_n(\theta) - \alpha(\theta)|^2 d\theta \;\sim\; \frac{1}{n}\,\mathbf{E}\int_{\Lambda} |Z(\theta)|^2 d\theta$$

and

$$n^{1/2} \sup_{\theta\in\Lambda} |\alpha_n(\theta) - \alpha(\theta)| \;\Rightarrow\; \sup_{\theta\in\Lambda} |Z(\theta)|$$

Similar behavior occurs in "IPA" setting, because once again

$$n^{1/2}(\alpha_n(\theta) - \alpha(\theta)) \;\Rightarrow\; Z(\theta),$$

where $Z(\cdot)$ is a Gaussian random field, except that in this setting $Z(\cdot)$ is an a.s. differentiable random field

Benchmark Analysis in Optimization Setting: What happens without use of CRN's?

- Generate m iid points $\theta_1, \theta_2, \ldots, \theta_m$ from a positive continuous density g
- Perform n independent simulations at each of the m points $(X_1(\theta_i), \ldots, X_n(\theta_i))$
- Estimate $\min_{\theta} \alpha(\theta)$ via $\min_{1 \leq i \leq m} \overline{X}_n(\theta_i)$

- Must have $\log m/n \to 0$ in order that

$$\min_{1 \le i \le m} \overline{X}_n(\theta_i) \to \min_{\theta} \alpha(\theta)$$

as $n \to \infty$ (Devroye (1978), Ensor and G (1997))

- What is an optimal choice of m and n?

For a given computer budget c:

$$m \sim rc^{d/(d+4)}$$

$$n \sim r^{-1} c^{4/(d+4)}$$

- Then, (Chia and G (2012))

$$c^{\frac{2}{d+4}}(\min_{1\leq i\leq m} \overline{X}_n(\theta_i) - \min_{\theta} \alpha(\theta)) \Rightarrow \beta$$

  where

$$\mathbf{P}(\beta \leq x) = \exp\left(-\frac{2r^{\frac{d+4}{4}} g(\theta^*)\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})\sqrt{|\det H(\theta^*)|}} \int_0^\infty \overline{\Phi}\left(\frac{2x+y}{2\sigma(\theta^*)}\right) y^{\frac{2}{d}-1} dy\right)$$

  where $\overline{\Phi}(x) = \mathbf{P}(\mathcal{N}(0,1) > x)$

- If $\mathrm{var}X(\theta) = 0$, rate = $c^{-2/d}$

Analysis of Typical CRN Setting:

Note that

$$\epsilon^{-1/2}(Z(\theta^* + \epsilon\theta) - Z(\theta^*)) \ \Rightarrow \ R(\theta)$$

where $R(\cdot)$ is a Gaussian random field

Result: If θ_n is the minimizer of $\alpha_n(\cdot)$, then

$$n^{1/3}(\theta_n - \theta^*) \ \Rightarrow \ \arg\min_{\theta\in\Lambda}[\theta^T(H(\theta^*)/2)\theta + R(\theta)]$$

$$n^{2/3}(\alpha_n(\theta_n) - \alpha_n(\theta^*)) \ \Rightarrow \ \min_{\theta\in\Lambda}[\theta^T(H(\theta^*)/2)\theta + R(\theta)]$$

$$n^{1/2}(\alpha_n(\theta_n) - \alpha(\theta^*)) \ \Rightarrow \ Z(\theta^*)$$

But we only evaluate $\alpha_n(\cdot)$ at points $\theta_1, \theta_2, \ldots, \theta_m$:

When we optimally trade-off n versus m,

$$c^{2/(d+4)}(\alpha_n(\theta_m) - \alpha(\theta^*)) \;\Rightarrow\; \Gamma$$

Note that:

> The "discretized" minimum has the same convergence rate as in the independent case

But

> The "continuous" minimum converges (much) faster and at a dimension-independent rate

Use of CRNs ("IPA" case)

In the IPA setting where $Z(.)$ is differentiable,

$$\alpha_n(\theta) \approx \alpha(\theta^*)+Z(\theta^*)/n^{1/2}+(\theta-\theta^*)^T H(\theta^*)/2(\theta-\theta^*)+\nabla Z(\theta^*)^T(\theta-\theta^*)/n^{1/2}$$

Result:

$$\begin{aligned} n^{1/2}(\theta_n - \theta^*) &\Rightarrow H(\theta^*)^{-1}\nabla Z(\theta^*)^T \\ n(\alpha_n(\theta_n) - \alpha_n(\theta^*)) &\Rightarrow \nabla Z(\theta^*)H(\theta^*)/2\nabla Z(\theta^*)^T \\ n^{1/2}(\alpha_n(\theta_n) - \alpha(\theta^*)) &\Rightarrow Z(\theta^*) \end{aligned}$$

Evaluating at $\theta_1, \ldots, \theta_m$ leads to the same $c^{-2/(d+4)}$ rate as before....

But "continuous" minimum converges faster and (even) faster than in "typical" CRN setting

Change-of-Measure

Assume that $\alpha(\theta) = \mathbf{E}_\theta X$ where

$$P_\theta(d\omega) = L(\theta, \omega) P(d\omega)$$

Then,

$$\alpha(\theta) = \mathbf{E}_\theta X = \mathbf{E} X L(\theta)$$

so the response surface can be estimated via

$$\alpha_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} X_i L_i(\theta)$$

Rarely preserves monotonicity, convexity, etc.

- For exponential families:

$$L(\theta) = \exp\left(\theta \sum_{j=0}^{t-1} Z_i - t\psi(\theta)\right)$$

so $\alpha_n(\cdot)$ is very cheap to evaluate at many θ's in this case (as opposed to CRN's).

- variance in $L(\theta)$ tends to blow up "exponentially" in t and θ
- $\alpha_n(\cdot)$ is rarely a good global approximation to $\alpha(\cdot)$

Orthogonal Function Approximations

e.g.
$$\alpha(\theta) = \sum_{i=0}^{\infty} < \alpha, \phi_i > \phi_i(\theta)$$

where
$$< \alpha, \phi_i > = \int_{\Lambda} \alpha(\theta)\phi_i(\theta)w(\theta)d\theta$$
$$= \mathbf{E}X(\theta)\phi_i(\theta)\frac{w(\theta)}{h(\theta)}$$

Estimate $< \alpha, \phi_i >$ via Monte Carlo:

$$\alpha_n(\theta) = \sum_{i=0}^{m_n} \frac{1}{n} \sum_{j=1}^{n} X_j(\theta_j)\phi_i(\theta_j)\frac{w(\theta_j)}{h(\theta_j)} \qquad (G89)$$

For Fourier basis and $\Lambda = [0, 2\pi]$, rate of convergence is $n^{-\frac{1}{2}+\frac{1}{2p}}$, when $\alpha \in C^p$.

Bayesian Methods:

Approach: Put a Gaussian prior on space of functions with domain Λ.

i.e. impose a probability P on $C(\Lambda), C^1(\Lambda), C^2(\Lambda)$, etc.

- Then, model $\alpha(\cdot)$ as a realization of such a Gaussian random field.
- Compute posterior

$$\mathbf{P}\left(\alpha \in \cdot \mid \overline{X}_n(\theta_i) : 1 \le i \le m\right)$$

  where

$$\overline{X}_n(\theta_i) \overset{\mathcal{D}}{\approx} \alpha(\theta_i) + \frac{Z(\theta_i)}{\sqrt{n}}$$

- Computationally expensive calculation
- Can also compute posterior

$$\mathbf{P}(\alpha \in \cdot \mid \overline{X}_n(\theta_i), \nabla\overline{X}_n(\theta_i) : 1 \le i \le m)$$

  when sample gradients are present

Shape-constrained Estimation:

- Observe $X_1, X_2, \ldots, X_m$ at locations $\theta_1, \theta_2, \ldots, \theta_m$
- Assume

$$X_i = \alpha(\theta_i) + \nu_i$$

for $1 \le i \le m$, where $\alpha(\cdot)$ is convex and the ν_i's here satisfy $\mathbf{E}\nu_i = 0$.

- Goal: Compute a (global) estimator for $\alpha(\cdot)$

The Estimator

- Let $\mathscr{C} = \{g : \mathbb{R}^d \to \mathbb{R}$ such that g is convex$\}$
- Given a "weight function" $w(\cdot)$, estimate α via the minimizer $\hat{g}_n$ of

$$\varphi_n(g) = \frac{1}{n} \sum_{i=1}^{n} (X_i - g(\theta_i))^2 w(\theta_i)$$

s/t $\quad g \in \mathscr{C}$

The Quadratic Program

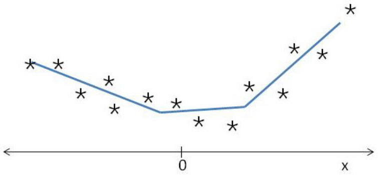
$$\min_{g_i,\xi_i} \quad \frac{1}{n}\sum_{i=1}^{n}(X_i - g_i)^2 w(\theta_i)$$

$$\text{s/t} \quad g_j \geq g_i + \xi_i^T(\theta_j - \theta_i), \quad 1 \leq i,j \leq n$$

- $(\hat{g}_1, \ldots, \hat{g}_n)$ is unique
- But the subgradients $\hat{\xi}_1, \ldots, \hat{\xi}_n$ are not unique
- There are many convex functions $\hat{g}_n$ that simultaneously minimize $\varphi_n(g)$ for $g \in \mathscr{C}$

To uniquely define $\hat{g}_n$, set

$$\hat{g}_n(x) = \sup\{g(x) : g \in \mathscr{C}, g(\theta_i) = \hat{g}_i, 1 \leq i \leq n\}$$



- $\hat{g}_n(x)$ is finite-valued on $\mathrm{conv}(\theta_1, \ldots, \theta_n)$ (∞ outside $\mathrm{conv}(\theta_1, \ldots, \theta_n)$)
- $\hat{g}_n(\cdot)$ is a "non-local" estimator (every point influences $\hat{g}_n(x)$

$\hat{g}_n(x)$ can be computed as the optimal value $\hat{y}$ to the linear program:

$$
\begin{aligned}
\max \quad & y \\
\text{s/t} \quad & \hat{g}_j \geq \hat{g}_i + \xi_i^T(\theta_j - \theta_i), \quad 1 \leq i,j \leq n \\
& y \geq \hat{g}_i + \xi_i^T(y - \theta_i), \quad 1 \leq i \leq n \\
& \hat{g}_j \geq y + \tilde{\xi}_i^T(\theta_j - y), \quad 1 \leq j \leq n
\end{aligned}
$$

Let $L^2(\Lambda) = \{g : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbf{E}g^2(\theta)w(\theta) < \infty\}$ and set

$$< g_1, g_2 >= \mathbf{E}g_1(\theta)g_2(\theta)w(\theta)$$

so $\|g\| = \sqrt{< g, g >}$.

Proposition

$\mathscr{C}^2 = \mathscr{C} \bigcap L^2(\Lambda)$ is a closed convex cone in $L^2(\Lambda)$

The minimizer g_* of

$$\min_{g \in \mathscr{C}^2} \|\alpha - g\|$$

is unique and is characterized as the function g_* for which

$$< \alpha - g_*, g - g_* > \leq 0$$

for all $g \in \mathscr{C}^2$.

Our main result...

Theorem (Lim and G (2012))

For each $c \geq 0$*,*

$$\sup_{\|x\| \leq c} |\hat{g}_n(x) - g_*(x)| \to 0 \text{ a.s.}$$

as $n \to \infty$

- previous results only for $d = 1$ (Hanson and Pledger (1976); Groeneboom, Jongbloed, Wellner (2001))
- first result on shape-constrained regression that deals with model mis-specification

Extensions

- Domain of g can be a convex subset of $\mathbb{R}^d$ (conclusion is "uniform convergence on compact subsets of interior")
- Generalizes to setting where $\mathscr{C} = \{g : \mathbb{R}^d \to \mathbb{R}$ is convex and non-decreasing$\}$

Outline of Proof

- By definition,

$$\varphi_n(\hat{g}_n) \le \varphi_n(g_*)$$

- So,

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(\theta_i) - g_*(\theta_i))^2 w(\theta_i)$$

$$\le \frac{2}{n}\sum_{i=1}^{n}(X_i - g_*(\theta_i))(\hat{g}_n(\theta_i) - g_*(\theta_i))w(\theta_i)$$

- If $\hat{g}_n(\cdot)$ were a fixed convex function in $L^2(\Lambda)$, SLLN would guarantee convergence of RHS to

$$\mathbf{E}(\alpha(\theta) - g_*(\theta))(\hat{g}_n(\theta) - g_*(\theta))w(\theta) =< \alpha - g_*, \hat{g}_n - g_* >\le 0$$

- Two problems:

$\hat{g}_n$ not fixed

$\hat{g}_n \notin L^2(\Lambda)$

So...

- Show that $(\varphi_n(\hat{g}_n) : n \geq 1)$ is a.s. a bounded sequence
- Use this to show that $(\hat{g}_n(x) : n \geq 1, \|x\| \leq c)$ is a.s. bounded
- This implies that $\hat{g}_n$ is uniformly (in n) a.s. Lipschitz over $\{x : \|x\| \leq c\}$
- Can form a *finite* ϵ-net $h_1, h_2, \ldots, h_l$ that provides a uniform cover for class of Lipschitz convex functions on $\{x : \|x\| \leq c\}$
- Each such h_j can be convexly extended to $\mathbb{R}^d$ so that $h_j \in L^2(\Lambda)$
- So,

$$\frac{2}{n}\sum_{i=1}^{n}(X_i - g_*(\theta_i))(h_j(\theta_i) - g_*(\theta_i))w(\theta_i) \to < \alpha - g_*, h_j - g_* > \leq 0$$

- But $\hat{g}_n$ is ϵ-close to one of the h_j's over $\{x : \|x\| \leq c\}$
- So, $\overline{\lim_{n\to\infty}} \frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(\theta_i) - g_*(\theta_i))^2 w(\theta_i) \leq 0$
- Because $\hat{g}_n$ is uniformly Lipschitz on $\{x : \|x\| \leq c\}$, this implies uniform convergence on $\{x : \|x\| \leq c\}$

Shape-based methods can be extended to Lipschitz constraints on the response function.

Open problem: Rates of convergence, particularly when the sampling employs common random numbers

Conclusions:

Response surface estimation is a challenging area for which many approaches are possible:

- Use of common random numbers is a central theme, and the connection to Guassian random fields and the degree of smoothness in the sample surface plays a key role
- Shape-constrained estimation is an interesting means of dealing with the infinite-dimensional aspect
- The cost of evaluating the response surface at a "new point" can be substantial, and a good choice of "interpolant" can be important
- Many open problems remain