New Approaches for Enhancing Simulation Metamodeling

Xiaowei Zhang

HKUST

Joint work with Haihui Shen (CityU), L. Jeff Hong (CityU), and Liang Ding (HKUST)

Outline

Introduction

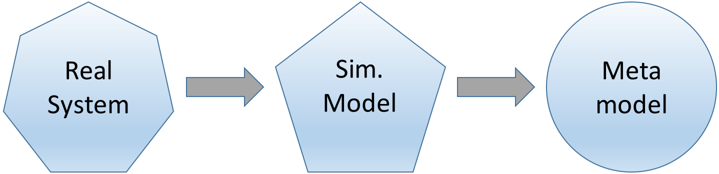
- Simulation is a popular tool for studying complex stochastic systems
  - e.g., supply chains, call centers, manufacturing systems

- We can optimize system performance by varying input parameters or design points of the simulation model

- However, simulation is often computationally expensive
  - e.g., large-scale queueing networks

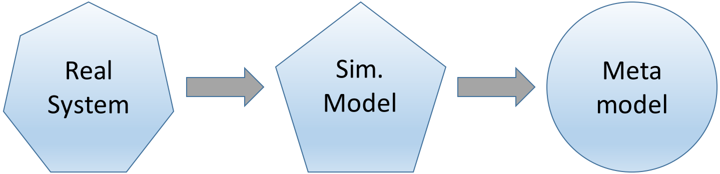
Metamodeling

- "Model of the simulation model"
- Run simulation at a small number of design points
- Use the simulation outputs at the selected design points to *predict* the simulation outputs at others *without* doing simulation



Metamodeling

- "Model of the simulation model"
- Run simulation at a small number of design points
- Use the simulation outputs at the selected design points to *predict* the simulation outputs at others *without* doing simulation



- $\mathbf{x} = (x_1, \ldots, x_d)^\intercal$: vector of decision variables of a real system
  - arrival rate to a critical care facility, service rates at different care units, and the routing probabilities among the units
- $\eta(\mathbf{x})$: mean performance of the simulation model

A Popular Metamodel: Stochastic Kriging (SK)

- Kriging was originally used in geostatistics (Matheron, 1963) and later in the design and analysis of computer experiments (Sacks et al., 1989)
- SK was proposed by Ankenman et al. (2010)

A Popular Metamodel: Stochastic Kriging (SK)

- Kriging was originally used in geostatistics (Matheron, 1963) and later in the design and analysis of computer experiments (Sacks et al., 1989)
- SK was proposed by Ankenman et al. (2010)

- SK metamodel: represent $\eta(\mathbf{x})$ by

$$\eta(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\intercal \boldsymbol{\beta} + M(\mathbf{x})$$

  - $\mathbf{f}(\mathbf{x})$: vector of known functions
  - $\boldsymbol{\beta}$: vector of unknown parameters
  - $M(\cdot)$: zero-mean Gaussian random field
- Simulation output:

$$Y_j(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\intercal \boldsymbol{\beta} + M(\mathbf{x}) + \epsilon_j(\mathbf{x})$$

  - $\epsilon_1(\mathbf{x}), \epsilon_2(\mathbf{x}), \ldots$ are the simulation errors

SK (cont'd)

- Slow: simulate $\eta(\mathbf{x})$ at $(\mathbf{x}_i : i = 1, \ldots, k)$
  - $Y_j(\mathbf{x}_i)$: simulation output on replication j at location $\mathbf{x}_i$

$$\overline{Y}(\mathbf{x}_i) := \frac{1}{n_i} \sum_{j=1}^{n_i} Y_j(\mathbf{x}_i)$$

- Fast: predict $\hat{\eta}(\mathbf{x_0})$ with $(\overline{Y}(\mathbf{x}_i) : i = 1, \ldots, k)$ for any $\mathbf{x_0}$

$$\hat{\eta}(\mathbf{x_0}) = \mathbf{f}(\mathbf{x_0})^\intercal \boldsymbol{\beta} + \boldsymbol{\Sigma}_M(\mathbf{x}_0, \cdot)^\intercal (\boldsymbol{\Sigma}_M + \boldsymbol{\Sigma}_\epsilon)^{-1} (\overline{\mathbf{Y}} - \mathbf{F}\boldsymbol{\beta})$$

Parameter Estimation via Maximum Likelihood

- Covariance function of M: $\mathrm{Cov}[M(\mathbf{x}, \mathbf{x}')] = \tau^2 R(\mathbf{x} - \mathbf{x}'; \boldsymbol{\theta})$
  - Typical example: $R(\mathbf{x} - \mathbf{x}'; \boldsymbol{\theta}) = \exp[-\theta \sum_i (x_i - x_i')^2]$

- Log-likelihood

$$\ell(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}) = -\ln |\boldsymbol{K}(\tau^2, \boldsymbol{\theta})| - (\overline{\mathbf{Y}} - \mathbf{F}\boldsymbol{\beta})^\intercal \boldsymbol{K}(\tau^2, \boldsymbol{\theta})^{-1} (\overline{\mathbf{Y}} - \mathbf{F}\boldsymbol{\beta}),$$

where $K(\tau^2, \boldsymbol{\theta}) = \boldsymbol{\Sigma}_M(\tau^2, \boldsymbol{\theta}) + \boldsymbol{\Sigma}_\epsilon$

Potential Issues

1. Specification the "trend term" $\mathbf{f}(\mathbf{x})$
   - *Stylized-model enhanced SK*: if a rough analytical approximation is available
   - *Regularized SK*: otherwise, apply statistical learning methods to automatically select from a large collection of candidates

Potential Issues

1. Specification the “trend term” $\mathbf{f}(\mathbf{x})$
   - *Stylized-model enhanced SK*: if a rough analytical approximation is available
   - *Regularized SK*: otherwise, apply statistical learning methods to automatically select from a large collection of candidates

2. Computation of the inverse covariance matrix $\boldsymbol{K}^{-1}$
   - Complexity is $\mathcal{O}(k^3)$
   - Prone to numerical instability: $\boldsymbol{K}$ becomes close to being singular if there are two design points that “close” to each other
   - *Markovian SK*: model K^{-1} directly and introduce sparsity by imposing Markovian structure

Literature on Enhancing Metamodels

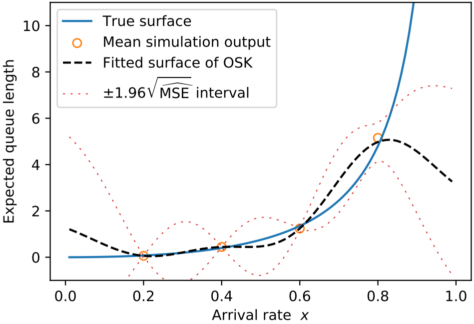
- Incorporating gradient information
  - Morris et al. (1993), Mitchell et al. (1994) in DACE
  - Chen et al. (2013), Qu and Fu (2014) in stochastic simulation

- Leveraging another coarser but faster simulation model
  - Kennedy and O'Hagan (2000), Forrester et al. (2007)

- Let $f(x) = p(x)/(1-x)^n$ with $p(x)$ being a polynomial
  - Cheng and Kleijnen (1999), Yang et al. (2007)
  - hard to generalize if $\mathbf{x}$ is multidimensional

Ordinary Stochastic Kriging (OSK)

- Despite its general form, in applications $\mathbf{f}(\mathbf{x})$ is mostly taken as a constant, i.e., $\mathbf{f}(\mathbf{x})^\intercal \boldsymbol{\beta} \equiv \beta_0$
  - Assuming no prior knowledge about $\eta(\mathbf{x})$
- Hard to capture highly nonlinear response surfaces



Stylized-model Enhanced Stochastic Kriging (SESK)

- A bulk of simulation models in practice are queueing networks
- Assuming no prior knowledge seems overly simplified given the long history of queueing theory

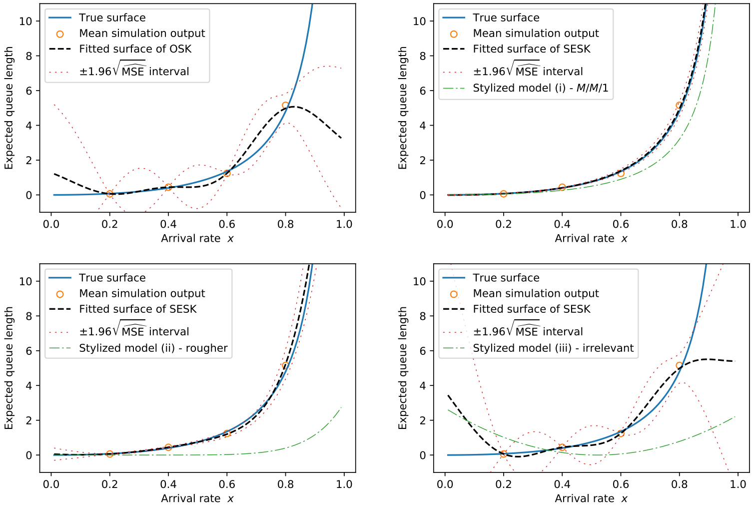
Stylized-model Enhanced Stochastic Kriging (SESK)

- A bulk of simulation models in practice are queueing networks
- Assuming no prior knowledge seems overly simplified given the long history of queueing theory

- SESK: add a stylized model with a closed-form solution to the trend term, $\mathbf{f}(\mathbf{x}) = (1, q(\mathbf{x}))$
  - e.g., $q(\mathbf{x})$ is the mean queue length of the Jackson network
- We do not expect much *quantitative* accuracy from $q(\mathbf{x})$ but merely a rough prediction of the *qualitative* behavior of $\eta(\mathbf{x})$.

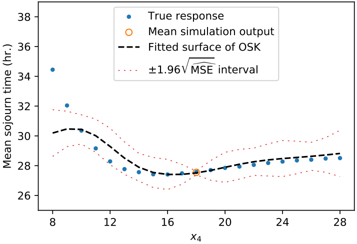
Example: $M/G/1$ Queue

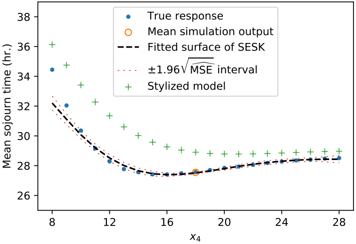
- True surface $\eta(x) = 1.5x^2/(1-x)$
- $q^{(1)}(x) = x^2/(1-x)$, $q^{(2)}(x) = 3x^9$, $q^{(3)}(x) = 10(x-0.52)^2$.



Example: Patient Flow in a Hospital

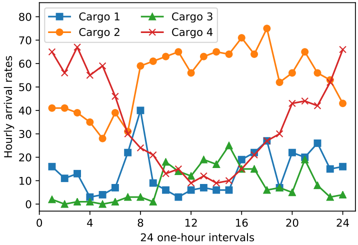
- An open queueing network with 9 servers (medical units)
- Each server has a finite capacity and patients may be blocked
- Stylized model: treat each server as an isolated $M/M/s/c$ queue
  - Adjust arrival rate and service rate via a system of heuristic equations





Example: Dock Allocation at an Air Cargo Terminal

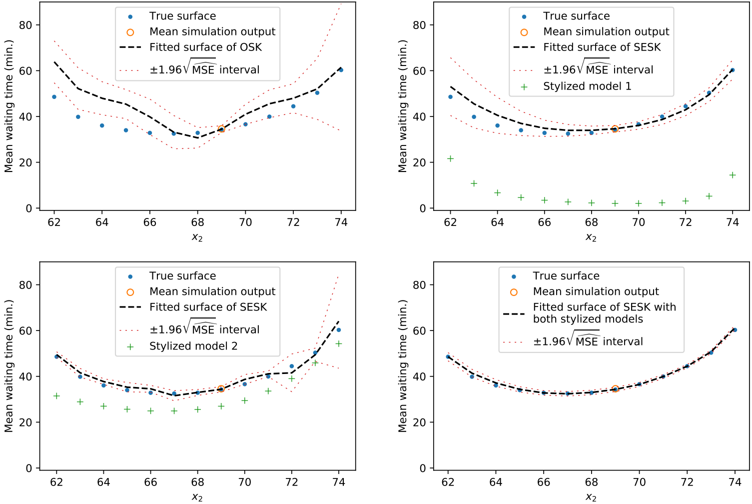
- Four multi-server queues with time-varying arrivals: $M_t/G/s$ queues
- Find the optimal scheme for allocating servers to the four queues to minimize the mean waiting time



Cargo Type	Service Time Distribution (min.)	Number of Docks
1	WEIB(21.8, 1.3)	x_1
2	7 + WEIB(67.6, 1.5)	x_2
3	7 + GAMM(25.7, 0.9)	x_3
4	7 + GAMM(9.4, 3.0)	x_4

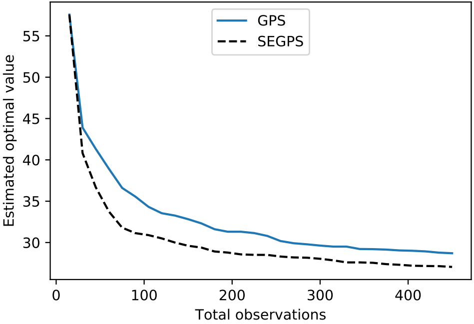
Dock Allocation (cont'd)

- Consider two distinct stylized models
  - Stationary approximation for each queue: $M/M/s$
  - Fluid approximation of $M_t/M/s$



Dock Allocation (cont'd)

- There are 111 servers in total
- Decision variable
  $\{x \in \mathbb{N}_+^4 : \sum_i x_i \leq 111, x_1 \geq 5, x_2 \geq 61, x_3 \geq 5, x_4 \geq 21\}$
  - 8,855 possible values in total
- Apply the Gaussian process-based search (GPS) algorithm (Sun et al., 2014) for optimization
  - The original version uses OSK
  - Replace it with SESK



Regularized Stochastic Kriging (RSK)

- What if an analytical approximation is not easy to find or implement?

- $\mathbf{f}(\mathbf{x})$ is analogous to basis functions in nonparametric regression
  - Nontrivial to select (form, number of terms, etc.) *manually*

Regularized Stochastic Kriging (RSK)

- What if an analytical approximation is not easy to find or implement?

- $\mathbf{f}(\mathbf{x})$ is analogous to basis functions in nonparametric regression
  - Nontrivial to select (form, number of terms, etc.) *manually*
- Treat it as a feature selection problem in statistical learning
- Use the regularization technique to *automatically* select proper basis functions from a large collection
  - Penalize the magnitude of $\boldsymbol{\beta}$ properly in its estimation
  - L_1 penalty drives the estimated coefficients of the *insignificant* functions to zero
  - Different from LASSO regression because of the correlated noise

Penalized Maximum Likelihood Estimation

- Log-likelihood of SK:

$$\ell(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}) = -\ln |\boldsymbol{K}(\tau^2, \boldsymbol{\theta})| - (\overline{\mathbf{Y}} - \mathbf{F}\boldsymbol{\beta})^\intercal \boldsymbol{K}(\tau^2, \boldsymbol{\theta})^{-1}(\overline{\mathbf{Y}} - \mathbf{F}\boldsymbol{\beta}),$$

  where $K(\tau^2, \boldsymbol{\theta}) = \boldsymbol{\Sigma}_M(\tau^2, \boldsymbol{\theta}) + \boldsymbol{\Sigma}_\epsilon$

- Penalized log-likelihood of RSK:

$$\tilde{\ell}(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}) = \ell(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}) - p(\boldsymbol{\beta}),$$

  where $p(\cdot)$ is a penalty function
  - L_1 penalty: $p(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_1$
  - Elastic net penalty: $p(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2$

Penalized Maximum Likelihood Estimation

- Log-likelihood of SK:

$$\ell(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}) = -\ln|\boldsymbol{K}(\tau^2, \boldsymbol{\theta})| - (\overline{\mathbf{Y}} - \mathbf{F}\boldsymbol{\beta})^\intercal \boldsymbol{K}(\tau^2, \boldsymbol{\theta})^{-1}(\overline{\mathbf{Y}} - \mathbf{F}\boldsymbol{\beta}),$$

  where $K(\tau^2, \boldsymbol{\theta}) = \boldsymbol{\Sigma}_M(\tau^2, \boldsymbol{\theta}) + \boldsymbol{\Sigma}_\epsilon$
- Penalized log-likelihood of RSK:

$$\tilde{\ell}(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}) = \ell(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}) - p(\boldsymbol{\beta}),$$

  where $p(\cdot)$ is a penalty function
  - L_1 penalty: $p(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$
  - Elastic net penalty: $p(\boldsymbol{\beta}) = \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2$
- Use the block-coordinate descent method for numerical optimization
  - Alternately maximize over one of $\boldsymbol{\beta}$, τ^2, $\boldsymbol{\theta}$ by fixing the other two

Penalized Maximum Likelihood Estimation

- Log-likelihood of SK:

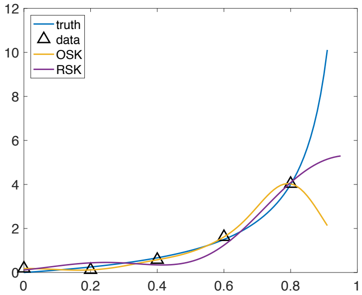
$$\ell(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}) = -\ln|\boldsymbol{K}(\tau^2, \boldsymbol{\theta})| - (\overline{\mathbf{Y}} - \mathbf{F}\boldsymbol{\beta})^\intercal \boldsymbol{K}(\tau^2, \boldsymbol{\theta})^{-1}(\overline{\mathbf{Y}} - \mathbf{F}\boldsymbol{\beta}),$$

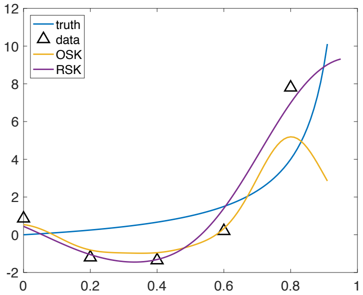
  where $K(\tau^2, \boldsymbol{\theta}) = \boldsymbol{\Sigma}_M(\tau^2, \boldsymbol{\theta}) + \boldsymbol{\Sigma}_\epsilon$
- Penalized log-likelihood of RSK:

$$\tilde{\ell}(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}) = \ell(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}) - p(\boldsymbol{\beta}),$$

  where $p(\cdot)$ is a penalty function
  - L_1 penalty: $p(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$
  - Elastic net penalty: $p(\boldsymbol{\beta}) = \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2$
- Use the block-coordinate descent method for numerical optimization
  - Alternately maximize over one of $\boldsymbol{\beta}$, τ^2, $\boldsymbol{\theta}$ by fixing the other two
- Nontrivial to prove the "oracle" property
  - Sparsity: estimated coefficients of the insignificant basis functions become zero asymptotically
  - Asymptotic optimality: estimated coefficients follow a multivariate normal distribution asymptotically

Example: $M/M/1$ Queue





- True surface $\eta(x) = x/(1-x)$
- Improvement relative to OSK is significant but not as much as SESK
  - SESK is better *if* a good stylized model is available
  - RSK is more widely applicable

Numerical Issues

- MLE requires repeated computation of $\mathbf{K}(\tau^2, \boldsymbol{\theta})^{-1}$
- Computational complexity is $\mathcal{O}(k^3)$
  - k becomes large easily if $\mathbf{x}$ is multidimensional
- A more serious numerical issue is $\mathbf{K}$ becomes near-singular easily
- $\boldsymbol{\theta}$ is hard to estimate (Li and Sudjianto, 2005)
  - $\boldsymbol{\theta}$ controls the correlation: $\mathrm{Corr}[M(\mathbf{x}), M(\mathbf{x}')] = \exp[-\theta \sum_i (x_i - x_i')^2]$
  - Log-likelihood function is "flat" near the optimum of $\boldsymbol{\theta}$

Numerical Issues

- MLE requires repeated computation of $\mathbf{K}(\tau^2, \boldsymbol{\theta})^{-1}$
- Computational complexity is $\mathcal{O}(k^3)$
  - k becomes large easily if $\mathbf{x}$ is multidimensional
- A more serious numerical issue is $\mathbf{K}$ becomes near-singular easily
- $\boldsymbol{\theta}$ is hard to estimate (Li and Sudjianto, 2005)
  - $\boldsymbol{\theta}$ controls the correlation: $\mathrm{Corr}[M(\mathbf{x}), M(\mathbf{x}')] = \exp[-\theta \sum_i (x_i - x_i')^2]$
  - Log-likelihood function is "flat" near the optimum of $\boldsymbol{\theta}$

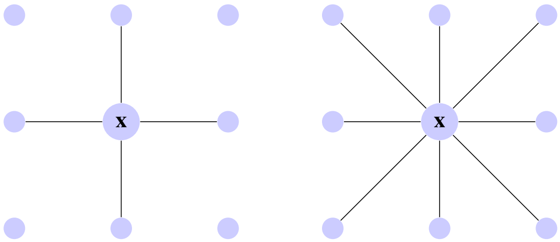
- Solution: model $\mathbf{Q} = \mathbf{K}^{-1}$ directly
  - Other approaches: tampering, low-rank approximation, etc.

Markovian Structure and Sparsity

- Crucial property: $\mathbf{Q}_{i,j} = 0$ if $\mathbf{x}_i \perp \mathbf{x}_j$ conditional on the others
  - $\{M(\mathbf{x}_i) : i = 1, \ldots, k\}$ forms a Markov chain

Markovian Structure and Sparsity

- Crucial property: $\mathbf{Q}_{i,j} = 0$ if $\mathbf{x}_i \perp \mathbf{x}_j$ conditional on the others
  - $\{M(\mathbf{x}_i) : i = 1, \ldots, k\}$ forms a Markov chain
- $M(\mathbf{x}_i)$ and $M(\mathbf{x}_j)$ are independent unless they are "neighbors"
- "Neighborhood" is defined by a user-specified graph, so $\mathbf{Q}$ can be made sparse
  - Accelerate the related matrix computation dramatically
  - Solve the near-singularity issue



(a) (b)

- If $\mathbf{x}$ is discrete, such $M(\mathbf{x})$ is a Gaussian Markov random field

Markovian Stochastic Kriging (MSK)

- With $\mathbf{x}$ being continuous, we assume $M(\mathbf{x})$ is a Gaussian free field
- The domain of $\mathbf{x}$ must be specified
- $G(\mathbf{x}, \mathbf{y}) := \mathrm{Cov}[M(\mathbf{x}), M(\mathbf{y})]$ is the solution to a PDE (heat equation with Dirichlet boundary)

Markovian Stochastic Kriging (MSK)

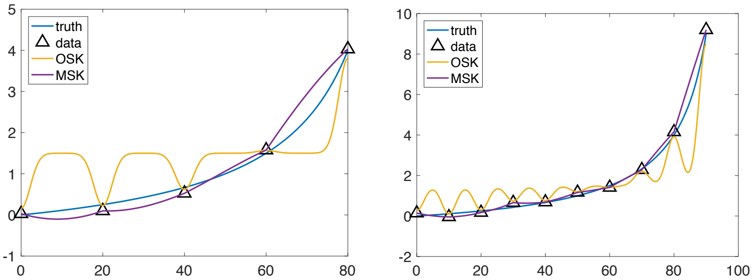
- With $\mathbf{x}$ being continuous, we assume $M(\mathbf{x})$ is a Gaussian free field
- The domain of $\mathbf{x}$ must be specified
- $G(\mathbf{x}, \mathbf{y}) := \mathrm{Cov}[M(\mathbf{x}), M(\mathbf{y})]$ is the solution to a PDE (heat equation with Dirichlet boundary)
  - E.g., if the domain is $[0, L]$, then $G(x, y)$ can computed analytically and $\mathbf{Q}$ is a tridiagonal matrix

$$\mathbf{Q} = \begin{pmatrix} b & -a & 0 & \cdots & 0 \\ -a & b & -a & \cdots & 0 \\ 0 & -a & b & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & b \end{pmatrix},$$

  - Log-likelihood becomes

$$\ell(\boldsymbol{\beta}, a, b) = -\ln|\mathbf{Q}(a, b)^{-1}| - (\overline{\mathbf{Y}} - \mathbf{F}\boldsymbol{\beta})^\intercal \mathbf{Q}(a, b)(\overline{\mathbf{Y}} - \mathbf{F}\boldsymbol{\beta})$$

Example: $M/M/1$ Queue



- True surface $\eta(x) = x/(100 - x)$
- Set $\mathbf{f}(\mathbf{x}) \equiv 1$ for MSK
- OSK does not perform well because the correlation parameter θ is hard to estimate (close to 0)

Conclusions

- Discussed several issues of the SK metamodel
  - Hard to specify the trend term $\mathbf{f}(\mathbf{x})$
  - High computational complexity
  - Numerical instability

- Proposed three approached for enhancing SK
  - They address different issues but can be used in a combined way
- SESK performs very well if a good stylized model is available
- RSK provides moderate enhancement but its applicability is higher than SESK
- MSK is very promising but the shape of the domain must be chosen carefully so that the PDE can be solved analytically

References I

B. Ankenman, B. L. Nelson, and J. Staum. Stochastic kriging for simulation metamodeling. *Oper. Res.*, 58(2):371–382, 2010.

X. Chen, B. Ankenman, and B. L. Nelson. Enhancing stochastic kriging metamodels with gradient estimators. *Oper. Res.*, 61(2):512–528, 2013.

R. C. Cheng and J. P. Kleijnen. Improved design of queueing simulation experiments with highly heteroscedastic responses. *Oper. Res.*, 47(5): 762–777, 1999.

A. I. Forrester, A. Sóbester, and A. J. Keane. Multi-fidelity optimization via surrogate modelling. In *Proc. R. Soc. A*, volume 463, pages 3251–3269, 2007.

M. C. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87 (1):1–13, 2000.

R. Li and A. Sudjianto. Analysis of computer experiments using penalized likelihood in Gaussian kriging models. *Technometrics*, 47(2):111–120, 2005.

G. Matheron. Principles of geostatistics. *Econ. Geol.*, 58(8):1246–1266, 1963.

References II

T. Mitchell, M. Morris, and D. Ylvisaker. Asymptotically optimum experimental designs for prediction of deterministic functions given derivative information. *J. Stat. Plann. Infer.*, 41(3):377–389, 1994.

M. D. Morris, T. J. Mitchell, and D. Ylvisaker. Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction. *Technometrics*, 35(3):243–255, 1993.

H. Qu and M. C. Fu. Gradient extrapolated stochastic kriging. *ACM Trans. Model. Comput. Simul.*, 24(4):23:1–23:25, 2014.

J. Sacks, S. B. Schiller, and W. J. Welch. Designs for computer experiments. *Technometrics*, 31(1):41–47, 1989.

L. Sun, L. J. Hong, and Z. Hu. Balancing exploitation and exploration in discrete optimization via simulation through a Gaussian process-based search. *Oper. Res.*, 62(6):1416–1438, 2014.

F. Yang, B. Ankenman, and B. L. Nelson. Efficient generation of cycle time-throughput curves through simulation and metamodeling. *Naval Res. Logist.*, 54(1):78–93, 2007.