Metamodel-Based Optimization

NSF Workshop on Simulation Optimization
May 24, 2010

Russell Barton
Penn State

Acknowledgments and Apologies

- To Michael and Barry.
- To NSF.
- To all of you.

Overview

- Metamodel-based optimization framework.
- Common issues: many factors, non-i.i.d. variation.
- RSM-based methods and issues.
- Global metamodel methods and issues.
- Current and future research opportunities.

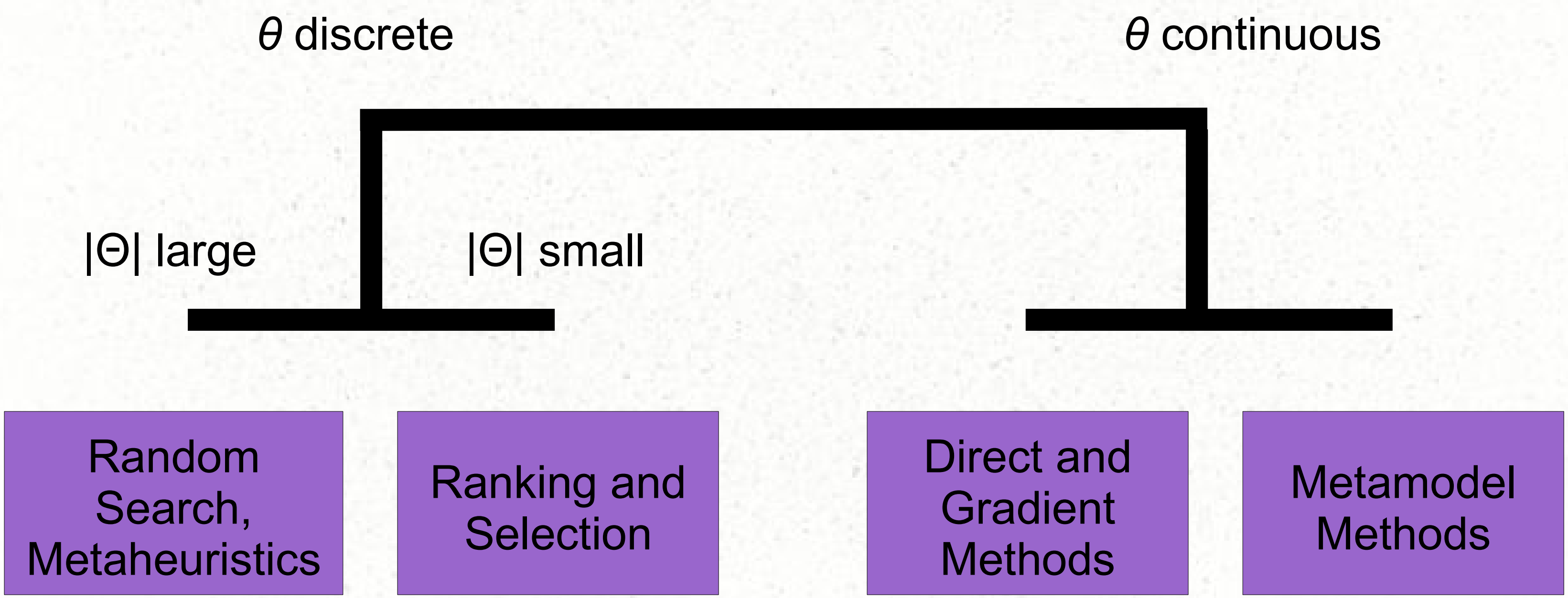
Framework: Simulation Optimization

$$min\ f(x) \equiv E(Y_0(x))$$

$$s.t.$$
$$a(x) \leq b$$
$$c(Y_0(x)) \leq d$$

- Y_0 generally random.
- One or more components of θ may be discrete.
- Constraints arising from c only implicit.
- Might optimize some other characteristic of the distribution of Y_0, e.g. a quantile (see Kleijnen, Pierreval, Zhang 2009 – working paper).
- Might have another statistical characteristic captured by c, e.g. variance, for robust parameter design – more later.

Taxonomy of Discrete-Event Simulation Optimization



Many simulation optimization ideas in Fu, Chen and Shi (2008).

Framework: Metamodel-Based (Simulation) Optimization

$$\min \hat{f}(x) \approx E(Y_0(x))$$
$$s.t.$$
$$a(x) \leq b$$
$$\hat{c}(Y_0(x)) \leq d$$
$$\hat{c}(Y_0(x)) \approx c(Y_0(x))$$

- f and c replaced by metamodels.
- Advantages:
  - Metamodels generally deterministic.
  - Metamodels can sometimes provide “insight.”
  - (Relatively) inexpensive to evaluate.
- Process is not this simple:
  - Metamodels refined as the optimization progresses.

Framework: Metamodel-Based (Simulation) Optimization

Stochastic Y_0

- Discrete-event simulation.
- Derivatives difficult or nonexistent.
- How long to run or how many replications?

Deterministic Y_0

- Finite-element and other engineering.
- Derivatives easy.
- Run length known.

Framework: Metamodel-Based Optimization

1. **Screen x s and Scale Y s.**
2. **Select initial DOE, make runs and fit initial metamodel.**
3. **Loop until done: assess fit and solution, refine/replace DOE, make additional runs and refine/replace metamodel.**

Common Issue: Many Factors

- Metamodels unreliable, DOEs large for many factors.
- Solution depends on Pareto principle: **Screen *x* s**.
- Screening has a long history, many methods, not the focus of this review. See Kleijnen et al. (2005), Kleijnen (2008a).
- Supersaturated designs
  - Fewer runs than factors.
  - Stepwise or ridge regression selection.
  - From Satterthwaite (1959) to Li and Lin (2003).
- Frequency domain
  - Jacobson, Buss and Schruben (1991).
- Likelihood ratio, IPA, SF
  - Fu and Hu (1997), Glynn and L'Ecuyer (1995), Rubinstein and Shapiro (1993).

Common Issue: Many Factors

- Sequential Bifurcation (SB)
    - Bettonvil and Kleijnen (1997).
- CSB (Wan et al. 2006; Wan et al. 2009) guarantees Pr(Type I error) < α for any effect with $|\beta_i| < \Delta_0$, and that the power of detection is greater than γ for any effect with $|\beta_i| > \Delta_1$. The 2009 modification to CSB gives fully sequential for each group without requiring $\alpha = 1 - \gamma$. CSB-X (Wan et al. 2009) screens in the presence of two-factor interactions.
- These methods assume sign of effects is known.
- FFCSB or FFCSB-X (Sanchez Wan and Lucas 2009) apply CSB/CSB-X using initial single-replication saturated FF to determine signs.

Common Issue: Heterogeneous Variance

- Discrete-event simulation plagued by heterogeneous variance of Y as a function of x.
- One solution: differential replication allocation.
- Kleijnen and van Groenendaal (1995) propose selecting the additional replicates such that the variances of average responses become a constant. Also, see Kleijnen (2008a).
- Plagued by $1/n$ property for variance reduction.
- If can tolerate higher prediction error in regions with high response means, then variance stabilizing transformations provide less costly alternative: **Scale *Y*s**.

Scaling Ys to Reduce Heterogeneous Variance

$$y(\lambda)=\begin{cases}\dfrac{y^{\lambda}-1}{\lambda} & \lambda\neq 0\\ \log(y) & \lambda=0\end{cases}$$

- Box-Cox Transformation
  - Estimate λ by maximum likelihood.
  - Or estimate λ from slope of plot of log(s.d.) vs log(mean) over all design points.
- Often a DOUBLE BENEFIT
  - Approximately equal variance
  - Better-behaved (closer to low-order polynomial) response surface!

Common Issue: Non-Independent Variance

- Discrete-event simulation provides opportunity for non-independence through common and antithetic random variates.
- Blocking strategies available for RSM only: Schruben and Margolin (1978), Tew and Wilson (1992, 1994), Donohue (1993), and GLM (WLS) see Kleijnen (1988, 2008).
- Strategies for global metamodels?

Framework: Metamodel-Based Optimization

- Local (RSM) vs. Global:
  - Different metamodel types.
  - Different experiment designs.
  - Different refine/replace strategies.

Framework: Local vs. Global

Local (RSM) metamodel strategy:

- Fit a sequence of local metamodels followed by local search. Old points discarded.
- Easier to provide good fit over a local region.
- Don't waste computational effort capturing response surface in regions that are suboptimal.

Global metamodel strategy:

- Fit a single metamodel to the entire region of interest.
- New metamodel types can provide global fidelity, given enough design points.
- Sequential local improvement in promising regions. Old points generally retained.
- Helps avoid focusing on a local optimum.

Local Metamodel Strategy: "Response Surface Methodology (RSM)"

- Metamodel type: usually low-order polynomial regression:

$$Y_0(x) = g(x)'\beta + \varepsilon,\ \varepsilon \text{ i.i.d. } N(0, \sigma^2)$$

- Box and Wilson (1951) for direct experiments. An early and extensive RSM bibliography is in Kleijnen's *Statistical Techniques in Simulation II* (1975). Current texts: Myers, Montgomery and Anderson-Cook (2009) and del Castillo (2007).
- Biles (1974) was an early application (WSC Proceedings).
- The method has only recently been formalized:
  - Donohue, Houck and Myers (ACM TOMACS 1993)
  - Neddermeijer (WSC 2000)
  - Nicolai (WSC 2004)
  - Barton and Meckesheimer (Handbook Sim. 2006)
  - Chang et al., Chang and Wan (WSC 2007, 2009).

Formal RSM Procedure

First-order Regression	Second-order Regression
L1: Determine initial local region.	
Small enough so linear approximation adequate, large enough so expected effects will be significant.	*NA*
L2: Choose a local metamodel form.	
First-order polynomial.	*Second-order polynomial.*
L3: Design local metamodel fitting experiment.	
Fractional factorial plus center point.	*Central composite, small composite or augmented fractional factorials.*
L5/L6/L7: Fit local metamodel and check fit for adequacy. Change model if necessary. Lack of fit test and tests for significance of regression coefficients.	
See Table	See Table
L8/L9: Provide a search direction or optimize the metamodel.	
Steepest ascent/descent.	*Direction based on canonical/ridge analysis.*
L10: Check the performance of the simulation at the metamodel-predicted optimum. Confirmation runs.	

L2: Local Metamodel Form

Metamodel Form	Experiment Design	Fitting; Model
First-order	Fractional-factorial (with or without CRN/ARN)	OLS, WLS, GLS; LR, GLM, Bayesian
Second-order	Central or small composite (with or without CRN/ARN)	OLS, WLS, GLS; LR, GLM, Bayesian

Formal RSM Procedure: first-order regression lack of fit test outcomes

	$\beta = 0$	$\beta \neq 0$
LOF	AUGMENT DESIGN AND FIT QUADRATIC MODEL	AUGMENT DESIGN AND FIT QUADRATIC MODEL **GO TO PHASE II**
NO LOF	CHOOSE LARGER RANGE FOR A NEW FIRST-ORDER DESIGN -----OR----- INCREASE THE NUMBER OF REPLICATIONS	LINE SEARCH IN NEGATIVE GRADIENT DIRECTION **GO TO L8/L9**

a) Model adequacy for Phase I.

Formal RSM Procedure: second-order regression lack of fit test outcomes

	$\beta = 0$	$\beta \neq 0$
LOF	UNLIKELY - IF THIS OCCURS, CHOOSE SMALLER RANGE FOR A NEW SECOND-ORDER DESIGN	CHOOSE SMALLER RANGE FOR A NEW SECOND-ORDER DESIGN
NO LOF	UNLIKELY - IF THIS OCCURS, INCREASE THE NUMBER OF REPLICATIONS	LINE SEARCH IN DIRECTION BASED ON CANONICAL ANALYSIS **GO TO L8/L9**

b) Model Adequacy for Phase II.

RSM Variations

- OLS can be replaced by WLS (see Scaling discussion above), GLS.
- Bayesian estimates of parameters (Cheng and Currie 2004).
- Linear model may be replaced by GLM (McCullagh and Nelder 1989, Staum 2009).
- Situation-specific nonlinear regression (Yang et al. 2007, Yang, Ankenman and Nelson 2007).

Recent Research: Local RSM Method

- Local search:
  - Stopping rule (del Castillo 2007).
  - Optimal budget (Peng, Lee and Ng 2007).
  - Trust region (Chang and Wan 2009).
  - Expected improvement criterion (Kleijnen et al. 2004; 2006) – maximize lower CI bound.
- Dealing with constraints and optimality:
  - Biles et al. (2007), Kleijnen (2008b), Bettonvil, del Castillo and Kleijnen (2009).
- Robust design:
  - Metamodels for mean and variance (Dellino, Kleijnen and Melloni 2010).
- Isotonic regression:
  - Lim and Glynn (2006).

Framework: Metamodel-Based Optimization

1. **Screen *x* s and Scale *Y* s.**
2. **Select initial DOE, make runs and fit initial metamodel.**
3. **Loop until done: assess fit and solution, refine/replace DOE, make additional runs and refine/replace metamodel.**

Global Metamodel Strategy

(does not match Barton and Meckesheimer 2006)

G1: Determine global region.

G2: Choose global metamodel form.

G3: Design initial global metamodel fitting experiment.

G4/5/6: Make runs, fit global metamodel, assess fit (e.g. via cross-validation).

G7: Optimize an expected improvement criterion. If expected improvement is small, stop.

G8: Conduct simulation experiment(s) at the optimum and refit the global model. Return to G7.

G2: Global Metamodel Form

Metamodel Form	Deterministic	Stochastic
Neural Networks	x	x
Radial Basis Functions	x	(x)
Spatial Correlation (Kriging)	x	x
(smoothing) Splines	x	x (low dimension)

Focus: Spatial Correlation (Kriging)

- *What is the model form*?

$Y(x) = \beta_0 + Z(x)$ or $Y(x) = x'\beta + Z(x)$, where $Z(x)$ is a stochastic process exhibiting spatial correlation:

$\mathrm{Cov}(Z(x), Z(x+\delta)) = \sigma_z^2 R(\delta)$, where R is the spatial correlation function, often $exp(-\Sigma(\theta_i\delta_i)^2)$.

- Historically, a version using a BLUP fit has been given the unfortunate name 'kriging' after one of the early developers (Krige).

Spatial Correlation Models

- Prediction (with intercept-only model):

$$\hat{Y}(x_0) = \hat{\beta}_0 + \hat{\Sigma}_Z(x_0, (X))' \hat{\Sigma}_Z^{-1}(Y(X) - \vec{1}\hat{\beta}_0)$$

Spatial Correlation Models

- Model provides interpolation
- Can lead to 'bumpy' surface for stochastic responses.
- Allen, Bernshteeyn and Kabiri-Bamoradian (JQT 2003) showed this error is often small, and benefit of better fit (less bias) over usual RSM methods.
- Kleijnen and co-authors successfully use deterministic kriging metamodels.

Spatial Correlation Models

- How can one handle stochastic responses?
- Huang et al. (2006), Ankenman Nelson and Staum (2010):

$$Y(x) = \beta_0 + Z_1(x) + \varepsilon(x)$$

exhibiting extrinsic uncertainty:

$$\mathrm{Cov}(Z_1(x), Z_1(x+\delta)) = \sigma_z^2 R(\delta)$$

and intrinsic uncertainty (ANS 2010):

$$\mathrm{Var}(\varepsilon(x)) = \sigma_\varepsilon^2 + Z_2(x).$$

Spatial Correlation Models

- Fitting (ANS 2010):

Fit spatial correlation model for Var($\varepsilon(x)$).

Estimate for Σ_ε diagonal with entries $\hat{V}(x_k)/n_k$

$$\hat{Y}(x_0) = \hat{\beta}_0 + \hat{\Sigma}_Z(x_0,(X))'[\hat{\Sigma}_Z + \hat{\Sigma}_\epsilon]^{-1}(Y(X) - \vec{1}\hat{\beta}_0)$$

G3: Initial Global Experiment Design

- Initial design typically "space filling."
  - Latin Hypercube (maximin Jones et al. 1998, Huang et al. 2006).
  - Orthogonal Array.
  - Maximin.
  - Grids problematic for kriging models.
- IMSE-optimal not attractive, since not interested in global fidelity in metamodel-based optimization.

G4/5/6: Fitting the Global Metamodel

- Fitting for kriging:
  - Originally by plotting the empirical 'variogram' which is the average squared difference in response over all x pairs in the DOE that are h units apart, plotted vs. h.
  - Instead, now typically use a known spatial correlation function and fit using MLE for β and σ^2 and then using these, for θ.
  - Kriging fit can be better with cross-validation in place of maximum likelihood (Sasena et al. 2002).

G7: Optimize Expected Improvement

- Improvement: $I = max\{0, f^n_{min} - Y\}$

  , f^n_{min} = smallest observed y

$$E(I) = \begin{cases} (f^n_{min} - \hat{y})\Phi\left(\frac{f^n_{min} - \hat{y}}{\hat{s}}\right) + \hat{s}\phi\left(\frac{f^n_{min} - \hat{y}}{\hat{s}}\right) & \text{if } s > 0 \\ 0 & \text{if } s = 0 \end{cases},$$

G7: Optimize Expected Improvement

- Booker et al. (1999) add runs during pattern optimization.
- Efficient Global Optimization (EGO). Concept presented in the deterministic response setting by Mockus, Tiesis and Zilinskas (1978). Next experiment design point (called an *infill* point) selected as on maximizing an improvement criterion.
- Generalized by Jones, Shonlau and Welch (1998) by adding an exponent g to I. Uses concept of *expected improvement*.

G7: Optimize Expected Improvement

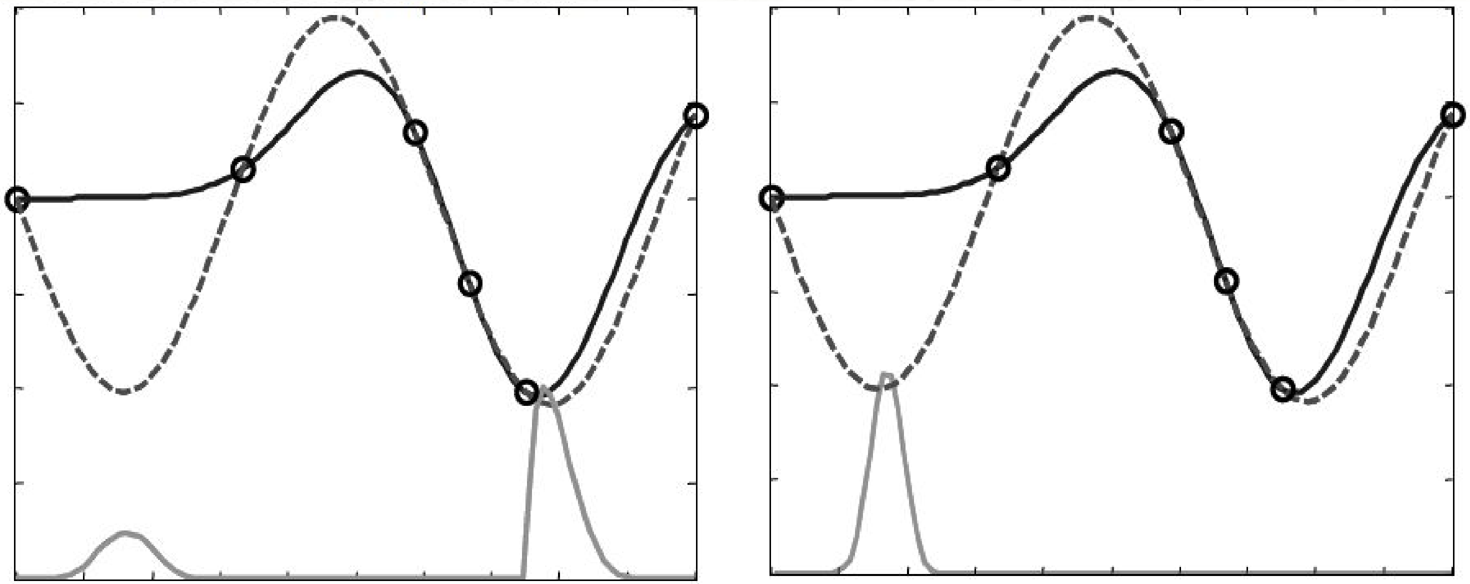


FIGURE 1 EI function for $g = 1$ (left) and $g = 5$ (right)

- Studied by Sasena, Papalambros and Goovaerts (2002). The exponent generalization did not work well on a suite of test problems.

G7: Optimize Expected Improvement

- Alternate form for Expected Improvement

$$\mathrm{WB2} = \begin{cases} \hat{y} + (f_{\min}^{n} - \hat{y})\Phi(f'^{n}_{\min}) + \hat{s}\phi(f'^{n}_{\min}) & \text{if } s > 0 \\ 0 & \text{if } s = 0 \end{cases} .$$

- This form worked well in Sasena et al. (2002).

G7: Optimize Expected Improvement

- Expected Improvement in the stochastic setting: Huang, Allen, Notz and Zeng (2006).
- Usual EI reduced by a multiple of the estimated intrinsic standard deviation.
- Called SKO.
- SKO reported less effective than SPO (Bartz-Beielstein 2006; Bartz-Beielstein & Preuss 2007) by Hutter et al. (2009).
- SPO is a stochastic version of the infill strategy discussed in (Sasena et al. 2002).

G7: Optimize Expected Improvement

- Regardless of the EI choice, global optimization has used multistart methods (e.g. Huang et al. 2006) or Lipschitzian optimization (Jones et al. 1993).

Recent Research: Global Methods

- Global metamodel-based optimization using stochastic kriging is new (2006).
- Stochastic kriging.
  - Work by Ankenman, Nelson, Huang, Allen cited earlier.
- Some interesting research with deterministic kriging.
  - Monotonic quantile fits (Kleijnen and van Beers 2009).
  - Constrained optimization via kriging and KKT (Kleijnen, van Beers and Nieuwenhuyse 2010).
  - Robust design (Dellino, Kleijnen and Melloni 2009).

Framework: Metamodel-Based Optimization

1. **Screen *x* s and Scale *Y* s.**
2. **Select initial DOE, make runs and fit initial metamodel.**
3. **Loop until done: assess fit and solution, refine/replace DOE, make additional runs and refine/replace metamodel.**

Future Directions for Metamodel-Based (Simulation) Optimization

- Sequential designs for stochastic kriging metamodeling – extending bootstrap approach of Kleijnen and van Beers (2004), examining alternative EI forms.

Future Directions for Metamodel-Based (Simulation) Optimization

- Sequential designs for stochastic kriging metamodeling – extending bootstrap approach of Kleijnen and van Beers (2004), examining alternative EI forms.

- Why are all DOE points always retained for global metamodel fits? Retaining all points increases run time and can introduce numerical instability.

Future Directions for Metamodel-Based (Simulation) Optimization

- Sequential designs for stochastic kriging metamodeling – extending bootstrap approach of Kleijnen and van Beers (2004), examining alternative EI forms.
- Why are all DOE points always retained for global metamodel fits? Retaining all points increases run time and can introduce numerical instability.
- The problem of performance testing (see Neddermeijer, Piersma, van Oortmarssen, Habbema and Dekker 1999 working paper cited in Neddermeijer et al. 2000, Pasupathy and Henderson 2006).

Future Directions for Metamodel-Based (Simulation) Optimization

- Sequential designs for stochastic kriging metamodeling – extending bootstrap approach of Kleijnen and van Beers (2004), examining alternative EI forms.
- Why are all DOE points always retained for global metamodel fits? Retaining all points increases run time and can introduce numerical instability.
- The problem of performance testing (see Neddermeijer, Piersma, van Oortmarssen, Habbema and Dekker 1999 working paper cited in Neddermeijer et al. 2000, Pasupathy and Henderson 2006).
- Combining multiple fidelity models for optimization - simulation fusion at the methodological level (Chan, Schruben, Nelson and Jacobson 2009, Kennedy and O'Hagan 2001).

Future Directions for Metamodel-Based (Simulation) Optimization

$$\begin{gathered} min\ \hat{f}(x) \approx E(Y_0(x)) \\ s.t. \\ a(x) \leq b \\ \hat{c}(Y_0(x)) \leq d \\ \hat{c}(Y_0(x)) \approx c(Y_0(x)) \end{gathered}$$

- *f* and *c* replaced by functions of metamodels characterizing the distribution of Y_0 rather than its mean (or mean and variance). May be useful for robust design and prediction intervals.

Future Directions for Metamodel-Based (Simulation) Optimization

$$\begin{gathered} \min\ \hat{f}(x) \approx E(Y_0(x)) \\ s.t. \\ a(x) \leq b \\ \hat{c}(Y_0(x)) \leq d \\ \hat{c}(Y_0(x)) \approx c(Y_0(x)) \end{gathered}$$

- *f* and *c* replaced by functions of metamodels characterizing the distribution of Y_0 rather than its mean (or mean and variance). May be useful for robust design and prediction intervals.
- In some cases, optimization is employed as a surrogate for the lack of an inverse. Metamodels can be built for inverse functions "for free." (Meckesheimer et al., 2002; Barton 2006).

Discussion