

Designing Computer Experiments to Find Constrained Optima

Thomas Santner

SAMSI Workshop

October 2006

Joint with Brian Williams and William Notz

With the cooperation of the Cornell/HSS Biomechanics Program

Outline

I. Introduction

II. The Model and a Specialization

III. The VIPER Algorithm

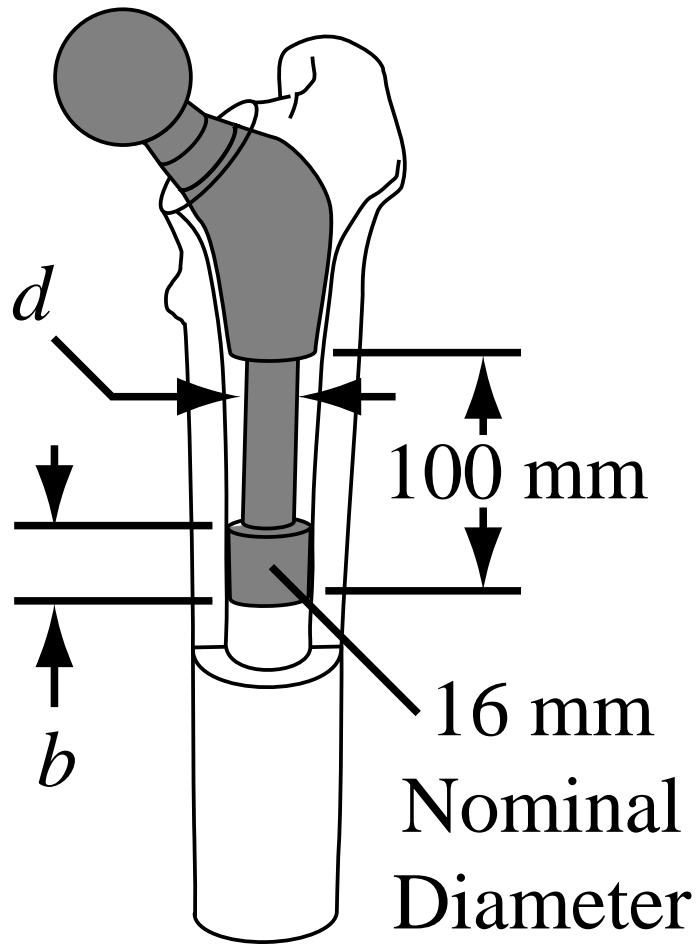
IV. Examples

V. Discussion and Take Home Points

I. Introduction

Overview This talk introduces an algorithm for **constrained optimization** of the **mean** of a computer output $y_1(\boldsymbol{x})$ (with respect to *known* distribution of field variables) when the constraint is defined by the mean of a second compute output $y_2(\boldsymbol{x})$. The algorithm is particularly useful when observations are *expensive* to compute. The $y_1(\boldsymbol{x})$ and $y_2(\boldsymbol{x})$ outputs can come from the same or different black box codes.

Motivating Example — Designing a hip implant

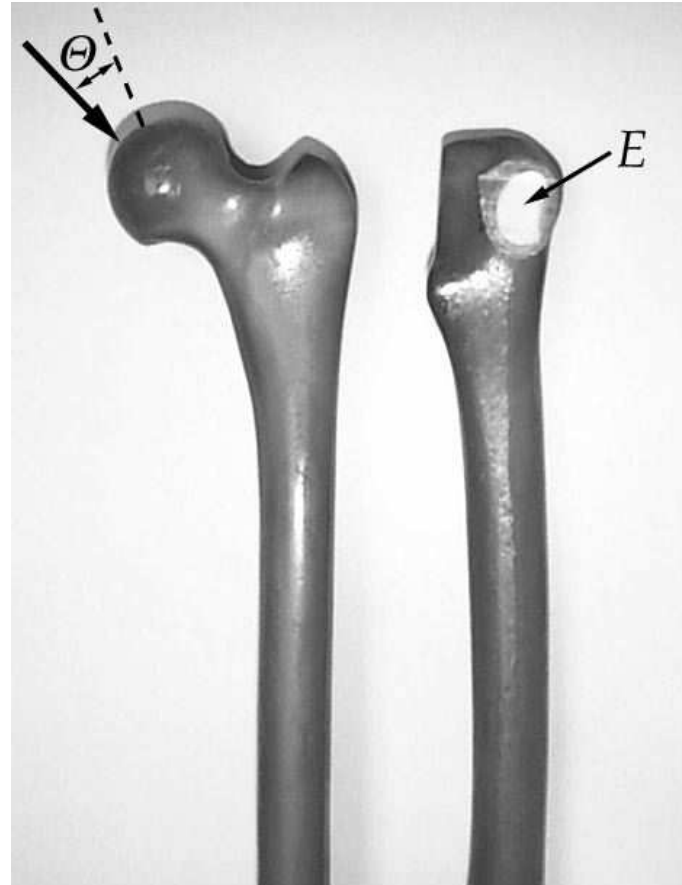


Goal: To determine the hip implant design, i.e., (b, d) , that **minimizes** femoral stress shielding while providing **adequate resistance** to implant toggling.

- b = bullet-tip length
- d = midstem diameter

- **Environmental Variables**

- E = elastic modulus of trabecular (aka cancellous) bone; a “spongy” bone structure found in the inner part of the bone
- Θ = joint force angle
- μ = interface friction ($\mu = 0$ is no friction)



Summary

- **Engineering Design Variables** = (b, d)
- **Environmental Variables** = (E, Θ, μ)
- Known joint distribution of Environmental Variables observed in the “Field”
 1. $(E, \Theta) \perp \mu$
 2. $[\mu]$ 10 point discrete uniform distribution on $[0.0, 0.42]$
 3. $[E, \Theta]$ 12 point discrete distribution on
 $\{60, 200, 400\} \times \{-10, -5, +5, +10\}$

Outputs (competing objectives)

1. $S = S(b, d, E, \Theta, \mu)$ = normalized measure of bone stress shielding for engineering design (b, d) in environment (E, Θ, μ)
2. $T = T(b, d, E, \Theta, \mu)$ = normalized measure of implant toggling

Biomechanically Increasing prosthesis stiffness, decreases implant toggling but increases bone stress shielding

Goal (*Formulation 1*) Minimize

$$y(b, d) = \omega S(b, d, E_0, \Theta_0, \mu_0) + (1 - \omega)T(b, d, E_0, \Theta_0, \mu_0)$$

where (E_0, Θ_0, μ_0) is a typical value, or (better)

$$y(b, d) = \omega E_{(E, \Theta, \mu)}\{S\} + (1 - \omega)E_{(E, \Theta, \mu)}\{T\}$$

where ω measures the relative importance of the two objectives.

Goal (*Formulation 2*) Minimize

$$\mu_S(b, d) = E_{(E, \Theta, \mu)} \{S(b, d, E, \Theta, \mu)\}$$

subject to

$$\mu_T(b, d) = E_{(E, \Theta, \mu)} \{T(b, d, E, \Theta, \mu)\} \leq B$$

where B is a specified, biomechanically meaningful bound

• Can replace $E_{(E, \Theta, \mu)} \{ \cdot \}$ in stochastic constraint that characterizations “smallness,” eg,

1. $P_{(E, \Theta, \mu)} \{T(b, d, E, \Theta, \mu) < \text{target}\} \geq 1 - \alpha$

2. $E_{(E, \Theta, \mu)} \{T(b, d, E, \Theta, \mu)\} + 2 \times \text{SE}_{(E, \Theta, \mu)} \{T(b, d, E, \Theta, \mu)\} \leq B^*$

Setup

- $\boldsymbol{x} = (\boldsymbol{x}_c, \boldsymbol{x}_e)$
 - \boldsymbol{x}_c = vector of control (manufacturing, engineering) variables
 - \boldsymbol{x}_e = vector of field (environmental) variables;
 $\boldsymbol{X}_e \sim F_e$, a **known** distribution F_e
- Bivariate output (signal) is $(y_1(\boldsymbol{x}), y_2(\boldsymbol{x}))$
- $\mu_i(\boldsymbol{x}_c) = E_{F_e} \{y_i(\boldsymbol{x}_c, \boldsymbol{X}_e)\}, i = 1, 2$
- **Goal** Find

$$\boldsymbol{x}_c^* = \operatorname{argmin}_{\boldsymbol{x}_c \in \mathcal{X}_c} \mu_1(\boldsymbol{x}_c) \quad \text{subject to} \quad \mu_2(\boldsymbol{x}_c^*) \leq U$$

- Approach is “Bayesian”

II. The Model and a Specialization

- For $i = 1, 2$, $y_i(\cdot)$ is viewed as a draw from the process

$$Y_i(\mathbf{x}) = \beta_i^\top \mathbf{f}_i(\mathbf{x}) + W_i(\mathbf{x}) + \epsilon_i(\mathbf{x})$$

where β_i is an **unknown** vector of regressors, $\mathbf{f}_i(\cdot)$ is a **known** vector of regression functions, $W_i(\cdot)$ is a **stationary** GSP $(0, \tau_i^2, R_i(\cdot | \xi_i))$ with **unknown** τ_i^2 and ξ_i and cross-correlation function $R_{12}(\cdot | \xi_{12})$, and $\epsilon_i(\mathbf{x})$ is a scaled WN process with variance σ_i^2

- $Cov(Y_i(\mathbf{x}_1), Y_i(\mathbf{x}_2)) = \tau_i^2 R_i(\mathbf{x}_1 - \mathbf{x}_2 | \xi_i)$ for $i = 1, 2$
- $Cov(Y_1(\mathbf{x}_1), Y_2(\mathbf{x}_2)) = \tau_1 \tau_2 R_{12}(\mathbf{x}_1 - \mathbf{x}_2 | \xi_{12})$
- $(R_1(\cdot | \xi_1), R_2(\cdot | \xi_2), R_{12}(\cdot | \xi_{12}))$ must be valid for all choices of $\xi = (\xi_1, \xi_2, \xi_{12})$

- **Covariance parameters**

$$(\tau_1^2, \eta \equiv \tau_2^2/\tau_1^2, \rho_1 \equiv \sigma_1^2/\tau_1^2, \rho_2 \equiv \sigma_2^2/\tau_2^2 = \sigma_2^2/(\eta\tau_1^2), \boldsymbol{\xi})$$

In the case of computer experiments, $\rho_1 = 0 = \rho_2$ so that the covariance parameters are $(\tau_1^2, \eta, \boldsymbol{\xi})$

- Assume F_e is **discrete** so that

$$\mu_i(\mathbf{x}_c) = E_{F_e} \{y_i(\mathbf{x}_c, \mathbf{X}_e)\} = \sum_{j=1}^{n_e} w_j y_i(\mathbf{x}_c, \mathbf{x}_{e,j}^{sp}).$$

- **Warning** Throughout, we assume that $y_1(\cdot)$ and $y_2(\cdot)$ are “expensive” to compute, so that neither $\mu_1(\mathbf{x}_c)$ nor $\mu_2(\mathbf{x}_c)$ would typically be computable at any \mathbf{x}_c

Specialization (Spatial autoregressive process à la Kenney and O'Hagan (2000))

$$Y_i(\mathbf{x}) = \beta_i + W_i(\mathbf{x}), \quad i = 1, 2$$

where $W_1(\mathbf{x}) \sim GSP(0, \tau_1^2, R_1(\cdot))$,

$$W_2(\mathbf{x}) = rW_1(\mathbf{x}) + W_\delta(\mathbf{x}),$$

$$W_1(\cdot) \perp W_\delta(\cdot),$$

$$W_\delta(\mathbf{x}) \sim GSP(0, \tau_\delta^2, R_\delta(\cdot))$$

1. If $r = 0$, $y_1(\mathbf{x})$ and $y_2(\mathbf{x})$ are modelled as independent processes (suggested by Schonlau et al. (1997) for constrained optimization)
2. If $y_1(\mathbf{x})$ and $y_2(\mathbf{x})$ are positively or negatively associated and on the same scale, then there can be identifiability issues—similar fits can be obtained with multiple r and τ_δ^2

Note

$$W_1(\cdot) \sim GSP(0, \tau_1^2, R_1(\cdot))$$

$$W_2(\mathbf{x}) = rW_1(\mathbf{x}) + W_\delta(\mathbf{x}),$$

$$W_1(\cdot) \perp W_\delta(\cdot)$$

\implies

$$\tau_2^2 \equiv \text{Var}(W_2(\mathbf{x})) = r^2\tau_1^2 + \tau_\delta^2$$

$$\eta \equiv \tau_2^2/\tau_1^2 = r^2 + \tau_\delta^2/\tau_1^2$$

$$R_2(\mathbf{x}_1 - \mathbf{x}_2) \equiv \text{Cor}(Y_2(\mathbf{x}_1), Y_2(\mathbf{x}_2))$$

$$= [r^2 R_1(\mathbf{x}_1 - \mathbf{x}_2) + (\eta - r^2) R_\delta(\mathbf{x}_1 - \mathbf{x}_2)] / \eta$$

$$R_{12}(\mathbf{x}_1 - \mathbf{x}_2) \equiv \text{Cor}(Y_1(\mathbf{x}_1), Y_2(\mathbf{x}_2)) = r R_1(\mathbf{x}_1 - \mathbf{x}_2) / \sqrt{\eta}$$

III. A biVariate constraiNed exPectEd impRovement

Algorithm

- **Recall** Find $\mathbf{x}_c^* = \arg \min \mu_1(\mathbf{x}_c)$ subject to $\mu_2(\mathbf{x}_c^*) \leq U$
- **Overview**
 1. Compute $y_1(\cdot)$ and $y_2(\cdot)$ at the points in an initial (space-filling) design.
 2. Use the information from the initial runs to select the next point according to a bivariate expected improvement criterion.
 3. Continue selecting points using the information from all of the previous runs until a stopping criterion is met.

- **Training Data** $S_n = \{\mathbf{x}_1^t, \dots, \mathbf{x}_n^t\}$ where $\mathbf{x}_j^t = (\mathbf{x}_{c,j}^t, \mathbf{x}_{e,j}^t)$
- **Control Variable Portions of S_n** $S_n^c = \{\mathbf{x}_{c,1}^t, \dots, \mathbf{x}_{c,n}^t\}$.
- $\mu_i(\mathbf{x}_c) = E_{F_e} \{y_i(\mathbf{x}_c, \mathbf{X}_e)\}, i = 1, 2$
- $\mu_1^{\min} = \min\{\mu_1(\mathbf{x}_{c,i}^t) : \mathbf{x}_{c,i}^t \in S_n^c, \mu_2(\mathbf{x}_{c,i}^t) \leq U\},$
- **Improvement** For arbitrary \mathbf{x}_c ,

$$i_n(\mathbf{x}_c) = \begin{cases} \mu_1^{\min} - \mu_1(\mathbf{x}_c), & \mu_1^{\min} > \mu_1(\mathbf{x}_c) \text{ and } \mu_2(\mathbf{x}_c) \leq U \\ 0, & \text{o.w.} \end{cases}$$

$$= \max\{0, \mu_1^{\min} - \mu_1(\mathbf{x}_c)\} \times I[\mu_2(\mathbf{x}_c) \leq U],$$

where

$I(A)$ is 1 if A occurs and is 0 otherwise

- **Prior for $\mu_i(\mathbf{x}_c)$** $M_i(\mathbf{x}_c) = \sum_{j=1}^{n_e} w_j Y_i(\mathbf{x}_c, \mathbf{x}_{e,j})$

- Take the distribution of

$$I_n(\mathbf{x}_c, \gamma) = \max\{0, M_1^{\min} - M_1(\mathbf{x}_c)\} \times I [M_2(\mathbf{x}_c) \leq U]$$

as a “prior” for $i_n(\cdot)$ where M_1^{\min} is defined to be

$$\min \left\{ M_1(\mathbf{x}_{c,i}^t) : E\{M_2(\mathbf{x}_{c,i}^t) \mid \mathbf{Y}_1^n, \mathbf{Y}_2^n, \gamma\} \right. \\ \left. - t_{2n-2, .95} \sqrt{\text{Var}\{M_2(\mathbf{x}_{c,i}^t) \mid \mathbf{Y}_1^n, \mathbf{Y}_2^n, \gamma\}} \leq U \right\}$$

- Randomness left in the definition of $I_n(\cdot, \gamma)$ involves $M_1(\mathbf{x}_c)$, $M_2(\mathbf{x}_c)$, and (some of) $(M_1(\mathbf{x}_{c,1}^t), \dots, M_1(\mathbf{x}_{c,n}^t))$

The VIPER Algorithm

- S0*: Choose the **initial** set of design points $S_n = \{\mathbf{x}_1^t, \dots, \mathbf{x}_n^t\}$
- S1*: **Estimate** the correlation parameter vector $\boldsymbol{\gamma} \equiv (\eta, \boldsymbol{\xi})$ by the mode of the posterior density of $\boldsymbol{\gamma}$ given $(\mathbf{Y}_1^n, \mathbf{Y}_2^n)$, say $\hat{\boldsymbol{\gamma}}_n$
- S2*: Choose the $(n + 1)$ -st **control** variable $\mathbf{x}_{c,n+1}^t$ to maximize the *posterior expected improvement* given the data and $\hat{\boldsymbol{\gamma}}_n$, i.e.,

$$E \{ I_n(\mathbf{x}_c, \hat{\boldsymbol{\gamma}}_n) \mid \mathbf{Y}_1^n, \mathbf{Y}_2^n, \hat{\boldsymbol{\gamma}}_n \} , \quad (1)$$

S3: Choose the **environmental** variable $\mathbf{x}_{e,n+1}^t$ corresponding to $\mathbf{x}_{c,n+1}^t$ to minimize the posterior MSPE given the data and $\hat{\gamma}_n$, i.e.,

$$E \left\{ \left[\widehat{M}_1^{n+1}(\mathbf{x}_{c,n+1}^t) - M_1(\mathbf{x}_{c,n+1}^t) \right]^2 \mid \mathbf{Y}_1^n, \mathbf{Y}_2^n, \hat{\gamma}_n \right\}$$

Here $\widehat{M}_1^{n+1}(\cdot)$ is the posterior mean of $M_1(\cdot)$, based on the data

$$S_n \cup \{Y_1(\mathbf{x}_{c,n+1}^t, \mathbf{x}_e), Y_2(\mathbf{x}_{c,n+1}^t, \mathbf{x}_e)\}$$

so that the expectation above is a function of \mathbf{x}_e

S4: If the stopping criterion **is not met** , set

$$S_{n+1} = S_n \cup \{(\mathbf{x}_{c,n+1}^t, \mathbf{x}_{e,n+1}^t)\} ,$$

calculate $y_1(\mathbf{x}_{c,n+1}^t, \mathbf{x}_{e,n+1}^t)$ and $y_2(\mathbf{x}_{c,n+1}^t, \mathbf{x}_{e,n+1}^t)$,
increment n to $(n + 1)$, and go to *S1*.

If the criterion **is met** estimate the global minimizer to be
the minimizer of the EBLUP of $M_1(\cdot)$ subject to the
EBLUP of $M_2(\cdot)$ satisfying the constraint

Implementation Details and Choices

- **Step S0** Space-filling designs maximin LHDs or, perhaps better, maximin orthogonal array LHDs work well (initial sample size n ? 5/input dimension?)
- **Step S1** Assuming the noninformative prior distribution,

$$[\boldsymbol{\beta}, \tau_1^2] \propto \frac{1}{\tau_1^2},$$

$(\mathbf{Y}_1^n, \mathbf{Y}_2^n)$ given $\boldsymbol{\gamma}$ has a multivariate t -distribution. So

$$\hat{\boldsymbol{\gamma}}_n = \operatorname{argmax}[\boldsymbol{\gamma}] \times [\mathbf{Y}_1^n, \mathbf{Y}_2^n | \boldsymbol{\gamma}]$$

(there are other options to estimate $\boldsymbol{\gamma}$)

- **Step S2** (Calculating the Posterior Expected Improvement).

Let $\mathbf{Z}_c^n = (\mathbf{M}_1^n, \mathbf{Y}_1^n, \mathbf{Y}_2^n)$ where

$\mathbf{M}_1^n = (M_1(\mathbf{x}_{c,1}^t), \dots, M_1(\mathbf{x}_{c,n}^t))$ is the vector of $M_1(\cdot)$ values evaluated on S_n^c . Then we observe

$$E \{ I_n(\mathbf{x}_c) \mid \mathbf{Y}_1^n, \mathbf{Y}_2^n, \gamma \} = E_{\mathbf{M}_1^n \mid \mathbf{Y}_1^n, \mathbf{Y}_2^n, \gamma} \{ E \{ I_n(\mathbf{x}_c) \mid \mathbf{Z}_c^n, \gamma \} \} .$$

The inner expectation can be calculated explicitly because

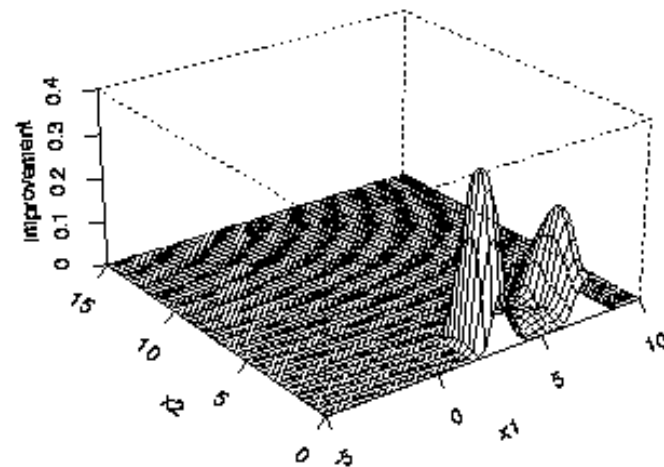
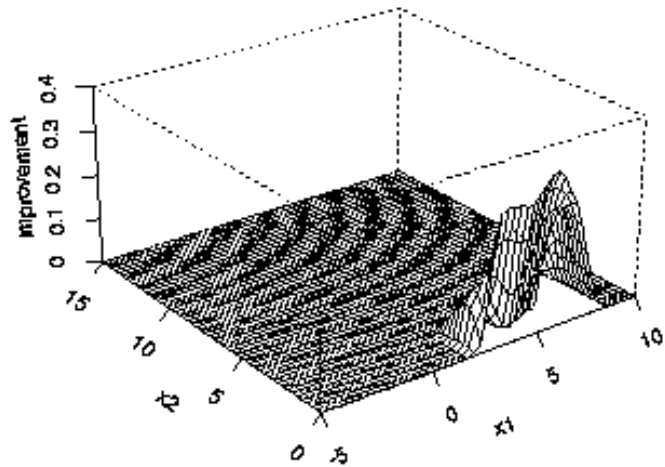
$[(M_1(\mathbf{x}_c), M_2(\mathbf{x}_c)) \mid \mathbf{Z}_c^n, \gamma]$ is a shifted, bivariate t which gives the formula for $E \{ I_n(\mathbf{x}_c) \mid \mathbf{Z}_c^n, \gamma \}$

$$\begin{aligned} & \sqrt{\widehat{\tau}_{1,c}^2 \mathbf{R}_{c,11}} \times \left[U_1 T_{2,\widehat{r}}(U_1, U_2, 3n - k) \right. \\ & \left. + C(U_1) T_{3n-1-k} \left(\frac{U_2 - \widehat{r}U_1}{\zeta_{\widehat{r}}(U_1)} \right) + \widehat{r} C(U_2) T_{3n-1-k} \left(\frac{U_1 - \widehat{r}U_2}{\zeta_{\widehat{r}}(U_2)} \right) \right] \end{aligned}$$

where all the terms are well-defined.

- **S2 (continued)** The outer expectation is calculated by Monte-Carlo draws from $[M_1^n | Y_1^n, Y_2^n, \gamma]$. (The number of draws-our experience, at least 1,000 MC samples)

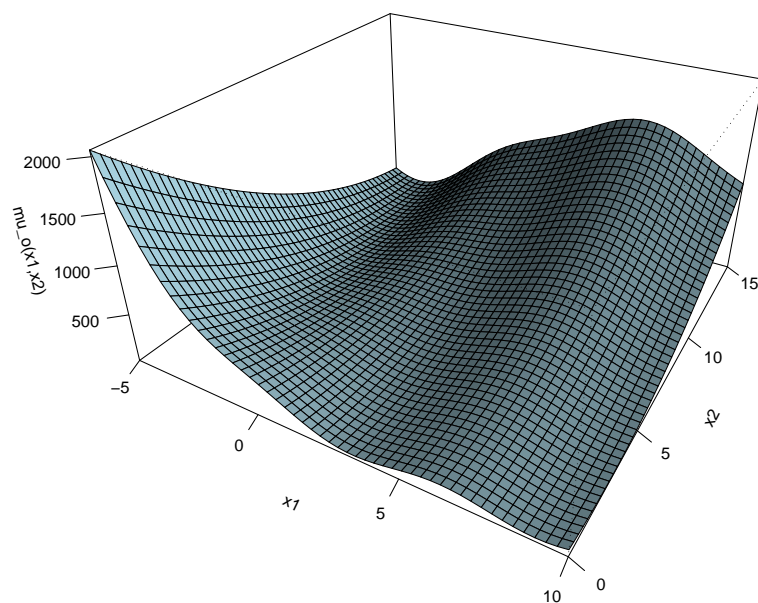
Typical Expected Improvement Surface



- **Step S3** There is a closed-form expression for posterior MSPE which can be minimized to obtain $\boldsymbol{x}_{e,n+1}$
- **Step S4** We stop only when both a moving **average** and a moving **range** of the expected improvements become “small”
- **Standardizing Output** We **always** center $y_1(\cdot)$ and $y_2(\cdot)$; based on our computational experience, we recommend that users also divide each of the centered $y_1(\cdot)$ and $y_2(\cdot)$ values by their sample standard deviation (\equiv “standardization”)

IV. An Example

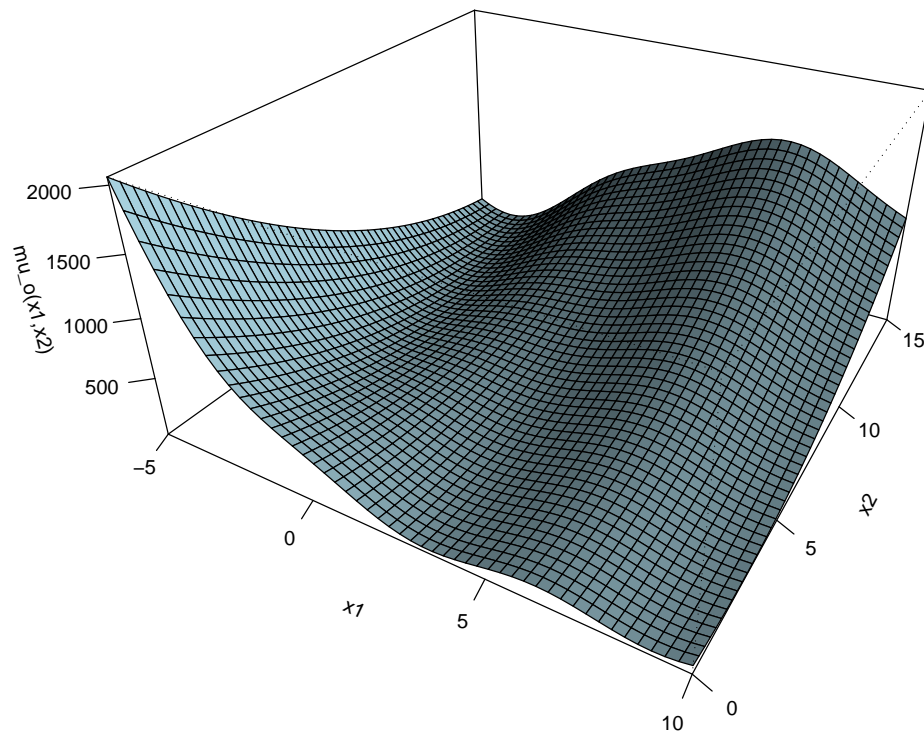
- **Six inputs** with $\mathbf{x}_c = (x_1, x_2) \in [-5, 10] \times [0, 15]$ and $\mathbf{x}_e = (x_3, x_4, x_5, x_6) \in [0, 1]^4$, uniform \mathbf{X}_e distribution on 81 support points
- **Objective Function** $\mu_1(x_1, x_2) = E_{F_e} \{y_1(\mathbf{x}_c, \mathbf{X}_e)\}$



- $\mu_1(\cdot)$ **ranges** from 1.5 to 2000

- **Three global minima**

$$\mu_1(\pi, 2.275) = \mu_1(3\pi, 2.475) = \mu_1(-\pi, 12.275) = 1.52680$$



- **Constraint Function** $\mu_2(x_1, x_2) = E_{F_e} \{y_2(\mathbf{x}_c, \mathbf{X}_e)\}$ (not shown) is based on same \mathbf{X}_e distribution as for $y_1(\mathbf{x}_c, \mathbf{x}_e)$
 1. $\mu_2(-\pi, 12.275) = -307.2373$, $\mu_2(\pi, 2.275) = -201.3822$,
and $\mu_2(3\pi, 2.475) = -104.8448$,
 2. If $\mathbf{X} = (X_1, \dots, X_6)$ has independent uniformly distributed components, then $\text{Cor}(y_1(\mathbf{X}), y_2(\mathbf{X})) \approx -0.51$ which suggests that **VIPER** based on r -arbitrary model may work better than **VIPER** or other optimization algorithms that assume $r \equiv 0$ (independent models)
- Taking $\mu_2(x_1, x_2) \leq -250 \implies \mathbf{x}_c^* = (-\pi, 12.2750)$ is the constrained minimizer of $\mu_1(x_1, x_2)$

- Some summary results for **VIPER** runs with implementation choices $\{r = 0, r \text{ arbitrary}\} \times \{\text{sequential, fixed}\}$

Algorithm	\hat{x}_1	\hat{x}_2	$\hat{x}_1 - x_1^{true}$	$\hat{x}_2 - x_2^{true}$
VIPER /BIV/45→70	-3.1400	12.2342	2.6100E-6	1.6660E-3
(One-stage)/BIV/70	-3.1639	10.4076	4.9786E-4	3.4873
VIPER /IND/45→70	-3.1489	12.3804	5.3652E-5	1.1109E-2
(One-stage /IND)/70	-2.9466	10.2731	3.8016E-2	4.0078
Truth	3.1416	12.275		

- Sequential design (45→70) dominates the fixed design that takes 70 observations according to a MmLHD
- **VIPER** with the bivariate model/predictor has a slight, but measurable advantage, over **VIPER** with independent objective and constraint function predictors

V. Discussion and Take Home Points

- **VIPER** is a useful tool for constrained optimization when observations are **expensive** to compute
- **VIPER** extends the idea of **expected improvement** introduced in Schonlau (1997)
- **VIPER** can have measureable improvement in finding x_c^* over the independent, fixed design of Jones, Schonlau, and Welch (1998)
- **VIPER** can also be used to optimize means of outputs that are measured with error
- **VIPER** can be used when different environmental variable distributions apply to the objective and constraint functions

- **VIPER** uses traditional minimization methods, appropriate for rapidly computed functions
 - to maximize the REML (to estimate γ)
 - to maximize the posterior expected improvement (to compute $\mathbf{x}_{c,n+1}$)
 - to minimize the MSPE (to compute $\mathbf{x}_{e,n+1}$)

Starting points for each optimization were a 200 point MmLHD in parameter space with the best 5 points selected as the starting points for Nelder-Mead simplex iterations followed by a Quasi-Newton iteration.

Extensions

1. The environmental variable distribution may be only **partially known** . This suggests that a **robust** (set of) control variable be selected
2. Feasible regions with **multiple** (≥ 2) **constraints** ?
3. Find the **Pareto frontier** of the x_c -space for the multiple objective functions.
4. There are numerous **alternative ways of choosing** $x_{e,n+1}$.
Any selection of $x_{e,n+1}$ that is space-filling in a neighborhood of $x_{c,n+1}$ may be reasonable (and easier to implement than the MSPE-minimization method of Step *S4*)

4. The $(W_1(\cdot), W_2(\cdot))$ model does not allow for the possibility that the strongest correlation occur between \mathbf{x}_1 and \mathbf{x}_2 that shifted by an unknown spatial parameter. One simple model to accommodate such a possibility is to take

$W_2(\mathbf{x}) = rW_1(\mathbf{x} - \Delta) + W_\delta(\mathbf{x})$ where Δ is a spatial shift parameter. The modified model \implies

$R_{12}(\mathbf{x}_1 - \mathbf{x}_2) = rR_1(\mathbf{x}_1 - \mathbf{x}_2 + \Delta)/\sqrt{\eta}$, so the dependence between $W_1(\mathbf{x}_1)$ and $W_2(\mathbf{x}_2)$ is strongest when $\mathbf{x}_1 - \mathbf{x}_2 = -\Delta$.