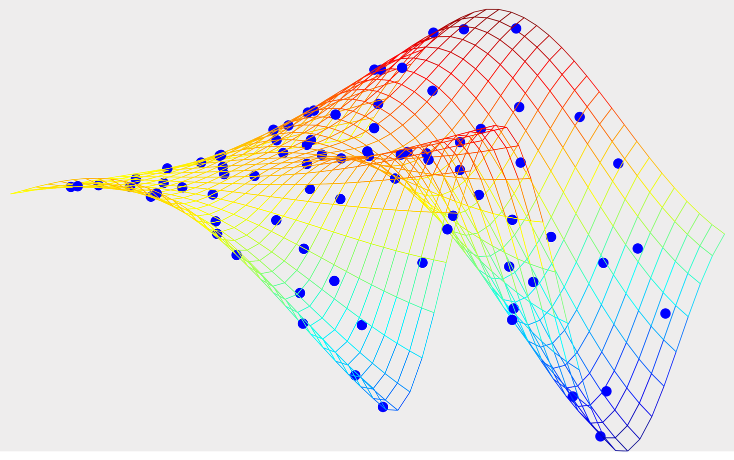


Design of Experiments when Gradient Information Is Available

Yiou Li
Department of Applied Mathematics

ILLINOIS INSTITUTE OF TECHNOLOGY
Transforming Lives.Inventing the Future.www.iit.edu

Summary Message

- For some computer experiments, evaluating a single data point can be computationally expensive, limiting the number of data that one can afford.
- Evaluating the gradient of the function using adjoint techniques, and involving gradient information in the model can substantially improve the accuracy of the prediction.
- The gradient of a d-variate function provides d more scalar pieces of information, at a cost of perhaps only several times the cost of a function value alone.
- Choosing the experimental design from two perspectives, *robustness* and *efficiency*.

Statistical Model

- The experimental region, Ω, is some measurable subset of $\Omega_1 \times \cdots \times \Omega_d$.
- $\mathcal{H}$ be some vector space of real-valued functions defined on Ω, and assumed to be a separable Hilbert space with a reproducing kernel $K : \Omega \times \Omega \to \mathbb{R}$.
- Define an operator $\mathbf{L}_{\boldsymbol{x}} : \mathcal{H} \to \mathbb{R}^{d+1}$, which when applied to a d-variate function $f \in \mathcal{H}$, returns

$$\mathbf{L}_{\boldsymbol{x}} f = \left(f(\boldsymbol{x}), \frac{\partial f}{\partial x_1}(\boldsymbol{x}), \cdots, \frac{\partial f}{\partial x_d}(\boldsymbol{x}) \right)^T.$$

- For a vector function $\boldsymbol{f} = (f_1, \ldots, f_\ell)^T : \Omega \to \mathbb{R}^\ell$, the definition of this operator is extended:

$$\mathbf{L}_{\boldsymbol{x}} \boldsymbol{f}^T = (\mathbf{L}_{\boldsymbol{x}} f_1, \ldots, \mathbf{L}_{\boldsymbol{x}} f_\ell).$$

- A linear regression model with gradient information:

$$\tilde{\boldsymbol{y}}_i = \left(\mathbf{L}_{\boldsymbol{x}_i} \boldsymbol{g}^T \right) \beta + \tilde{\varepsilon}_i, \qquad i = 1, \ldots, n,$$

$\tilde{\boldsymbol{y}}_i$: observed vector response at the design point $\boldsymbol{x}_i$
$\boldsymbol{g} = (g_1, \ldots, g_k)^T$: the vector of basis functions
β: regression coefficient to be estimated
$\tilde{\varepsilon}_i$: the error in estimating the response by the linear combination of k basis functions. It is assumed that $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n$ are i.i.d with zero mean and covariance matrix $\sigma^2 \widetilde{\boldsymbol{\Lambda}}$.

- Vector-matrix notation:

$$\boldsymbol{y} = \mathrm{G}\beta + \varepsilon,$$

where

$$\mathrm{G} = \begin{pmatrix} \mathbf{L}_{\boldsymbol{x}_1} \boldsymbol{g}^T \\ \vdots \\ \mathbf{L}_{\boldsymbol{x}_n} \boldsymbol{g}^T \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} \tilde{\boldsymbol{y}}_1 \\ \vdots \\ \tilde{\boldsymbol{y}}_n \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \tilde{\varepsilon}_1 \\ \vdots \\ \tilde{\varepsilon}_n \end{pmatrix},$$

and ε has zero mean and covariance matrix $\sigma^2 \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\widetilde{\boldsymbol{\Lambda}}, \ldots, \widetilde{\boldsymbol{\Lambda}})$.

- The weighted least squares estimate of the regression coefficient β:

$$\hat{\beta} = \mathrm{B}\boldsymbol{y} = \beta + \mathrm{B}\varepsilon, \qquad \mathrm{B} = \left(\mathrm{G}^T \boldsymbol{\Lambda}^{-1} \mathrm{G} \right)^{-1} \mathrm{G}^T \boldsymbol{\Lambda}^{-1}.$$

Scaled Integrated Mean Squared Error

- Experimental design:

$$\xi = \begin{Bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_N \\ w_1 & w_2 & \cdots & w_N \end{Bmatrix}, \text{where } N \le n.$$

- The goal of the linear regression is assumed to be the estimation of $\mathbf{T}(\boldsymbol{g}^T \beta)$, where $\mathbf{T} : \mathcal{H} \to \mathcal{L}_2^s$.
- For any vectors of functions, $\boldsymbol{u} = (u_1, \ldots, u_s)^T$ and $\boldsymbol{v} = (v_1, \ldots, v_s)^T$, let $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\mathcal{L}_2^s} = \int_\Omega \boldsymbol{u}^T(\boldsymbol{x}) \boldsymbol{v}(\boldsymbol{x}) \, \mathrm{dF_{IMSE}}(\boldsymbol{x})$.
- $\mathrm{IMSE}(\xi, \boldsymbol{g}) = \frac{n}{\sigma^2} E \left\| \mathbf{T}(\boldsymbol{g}^T \beta) - \mathbf{T}(\boldsymbol{g}^T \hat{\beta}) \right\|_{\mathcal{L}_2^s}^2$.

Proposition

$$\mathrm{IMSE}(\xi, \boldsymbol{g}) = \mathrm{tr}(\mathrm{M}^{-1}\mathrm{A}),$$

with

$$\mathrm{M} = \mathrm{M}_\xi = \frac{1}{n} \mathrm{G}^T \boldsymbol{\Lambda}^{-1} \mathrm{G}, \text{ and } \mathrm{A} = \left(\langle \mathbf{T} g_i, \mathbf{T} g_j \rangle_{\mathcal{L}_2^s} \right)_{i,j=1}^k$$

Low Discrepancy Design Bounds IMSE

Theorem

Suppose that F_T is a probability distribution function defined on Ω, which may be different from F_{IMSE} , and that $\mathcal{H}$ is a reproducing kernel Hilbert space of functions defined on Ω with reproducing kernel K. Consider the information matrix for F_T,

$$\mathrm{M}_{F_T} = \int_\Omega (\mathbf{L}_{\boldsymbol{x}} \boldsymbol{g}^T)^T \widetilde{\boldsymbol{\Lambda}}^{-1} (\mathbf{L}_{\boldsymbol{x}} \boldsymbol{g}^T) \, \mathrm{dF_T}(\boldsymbol{x}),$$

and suppose that the function $h_\alpha : \boldsymbol{x} \mapsto \boldsymbol{\alpha}^T (\mathrm{M}_{F_T})^{-\frac{1}{2}} \mathrm{M}_{\boldsymbol{x}} (\mathrm{M}_{F_T})^{-\frac{1}{2}} \boldsymbol{\alpha}$ lies in $\mathcal{H}$ for any $\boldsymbol{\alpha} \in \mathbb{R}^k$. Define a variation over the basis $\boldsymbol{g}$as

$$V_{\boldsymbol{g}, F_T} = \sup_{\|\alpha\|_2 \le 1} V(h_\alpha),$$

where V is the variation. Then it follows that the integrated mean square error is bounded above by

$$\mathrm{IMSE}(\xi, \boldsymbol{g}) \le \frac{\mathrm{tr}(\mathrm{M}_{F_T}^{-1} \mathrm{A})}{1 - D_{F_T}(\xi) V_{\boldsymbol{g}, F_T}},$$

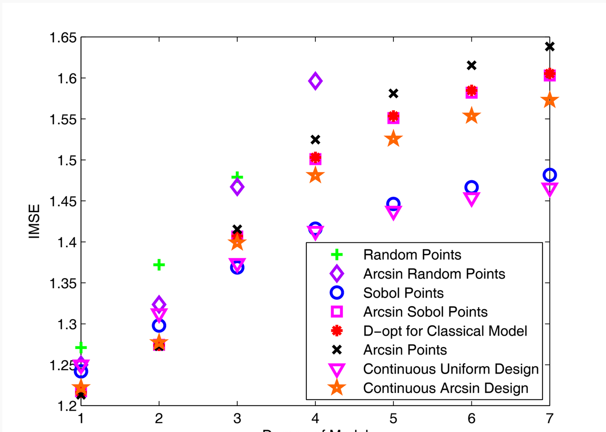
provided that $D_{F_T}(\xi) V_{\boldsymbol{g}, F_T} < 1$.

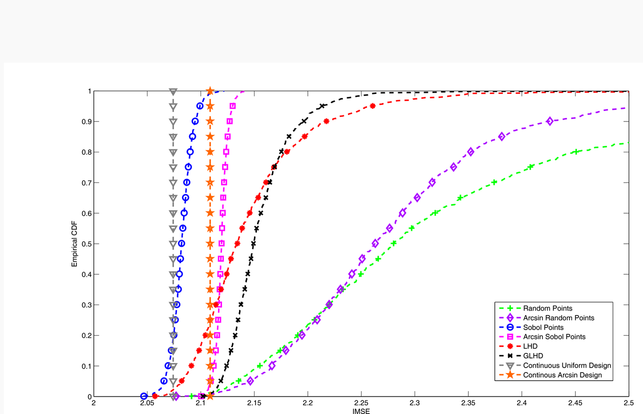
Remark

Note that, in some cases, there exists F_T, such that $\mathrm{tr}(\mathrm{M}_{F_T}^{-1}\mathrm{A}) < \mathrm{tr}(\mathrm{M}_{F_{\mathrm{IMSE}}}^{-1}\mathrm{A})$. This means that choosing the design to match the distribution F_T will yield to smaller upper bound compared to choosing the design to match distribution F_{IMSE}.

Numerical Experiments on Low Discrepancy Design

- $\Omega = [-1, 1]$, polynomial basis, 16 sample points, repeated 1000 times, and $\Omega = [-1, 1] \times [-1, 1]$, orthogonal basis up to degree 4, 32 sample points, repeated 1000 times.





Semi-definite Programming with Gradient Information

- Define $\boldsymbol{G}^T(\boldsymbol{x}) = \mathbf{L}_{\boldsymbol{x}} \boldsymbol{g}^T$, $\boldsymbol{F}(\boldsymbol{x}_j) = \boldsymbol{A}^{-\frac{1}{2}} \boldsymbol{G}(\boldsymbol{x}_j)$.
- Equivalent SDP model for I-optimal design:

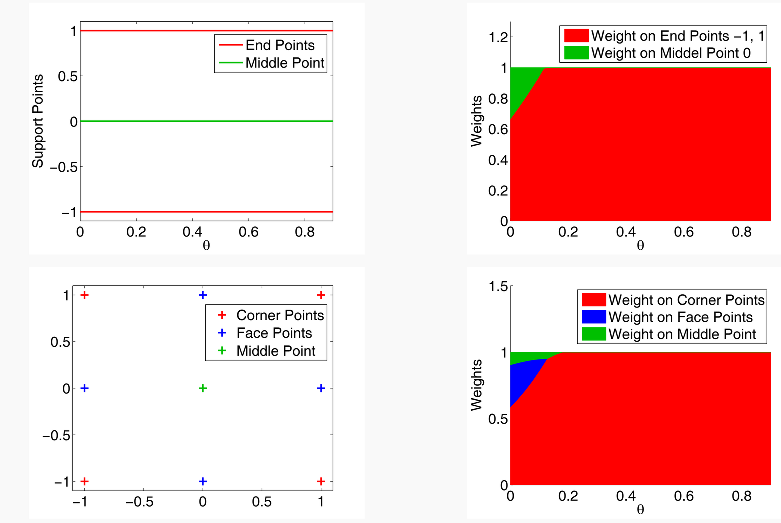
$$\begin{aligned} &\text{Minimize}_{w_j, \gamma} \ \boldsymbol{e}^T \gamma \\ &\text{Subject to: } \begin{bmatrix} \sum_{j=1}^N w_j \boldsymbol{F}(\boldsymbol{v}_j) \tilde{\boldsymbol{\Lambda}}^{-1} \boldsymbol{F}^T(\boldsymbol{v}_j) & \boldsymbol{I} \\ \boldsymbol{I} & \mathrm{diag}(\gamma) \end{bmatrix} \succeq 0, \\ &\qquad \boldsymbol{e}^T \boldsymbol{w} = 1, \\ &\qquad \boldsymbol{w} \ge 0. \end{aligned}$$

- SDP model for D-optimal design:

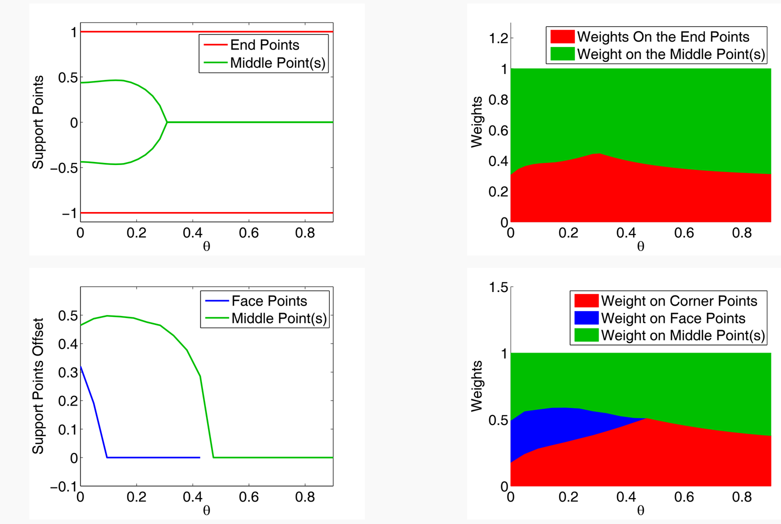
$$\begin{aligned} &\text{Minimize}_{w_j} \ -\log\det \left[\sum_{j=1}^K w_j \boldsymbol{G}(\boldsymbol{v}_j) \tilde{\boldsymbol{\Lambda}}^{-1} \boldsymbol{G}^T(\boldsymbol{v}_j) \right] \\ &\text{Subject to: } \boldsymbol{e}^T \boldsymbol{w} = 1, \\ &\qquad \boldsymbol{w} \ge 0. \end{aligned}$$

Numerical Results for Semi-definite Programming

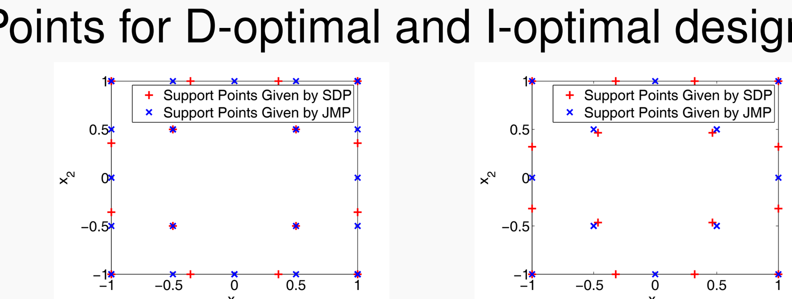
- D optimal design for quadratic model with variable in 1-d and 2-d:



- I optimal design for cubic model with variable in 1-d and 2-d:



- JMP vs SDP (cubic 2-d model)
  - Support Points for D-optimal and I-optimal design



  - Efficiency ratio:

		I Efficiency Ratio				D Efficiency Ratio			
		$1-d$ Cubic		$2-d$ Cubic		$1-d$ Cubic		$2-d$ Cubic	
		Cont	Exact	Cont	Exact	Cont	Exact	Cont	Exact
Number of Sample Points	12	1.0029	1.0004	1.1544	NA	1.0000	1.0000	1.0482	NA
	16	1.0185	0.9994	1.0359	1.0359*	1.0000	1.0000	1.0496	1.0168
	20	1.0007	1.0003	1.0574	1.0471	1.0000	1.0000	1.0403	1.0359
	24	1.0030	1.0005	1.0349	1.0302	1.0000	1.0000	1.0171	1.0171*
	28	1.0031	1.0003	1.0280	1.0280*	1.0000	1.0000	1.0256	1.0256*
	32	1.0004	1.0004	1.0295	1.0295*	1.0000	1.0000	1.0249	1.0211
	36	1.0031	1.0006	1.0327	1.0327*	1.0000	1.0000	1.0185	1.0137
	40	1.0009	1.0004	1.0277	1.0277*	1.0000	1.0000	1.0201	1.0159

Acknowledgements

I would like to thank my advisor Fred J. Hickernell, and our collaborator Dr. Mihai Anitescu, Dr. Jieqiu Chen from Argonne National Lab for the assistance in my research.