Low discrepancy sampling limit aliasing and D-optimal design when gradient information is available

Fred J. Hickernell, Yiou Li

Applied Mathematics Office, Engineering 1 *Building* 10 *West* 32*nd Street, Chicago, IL* 60616

Abstract

Keywords:

1. Introduction

In practice, there are situations that taking sample points is computationally expensive, so one can only afford to take limited number of sample points. Under these circumstances, PRD (polynomial regression with derivative information) could be used to approximate the regression coefficients in the statistical model. In PRD, the statistical model of the output is computed by regressing both output information and its derivatives with respect to the variables computed at a small number of sample points. In addition, in some practice, like nuclear reactor system simulations, it has been found that approximation of the uncertainty effect by PRD (polynomial regression with derivative information) is more precise than linear approximation by an order of magnitude or more. (reference) In our paper, we will discuss the PRD that includes gradient information, since in applications, there are methods like adjoint techniques that can be used for the efficient computation of gradient information.

Efficiency and robustness are both important concepts in experimental design. When the form of the statistical model is specified, optimal designs will lead to efficient estimates of the regression coefficients. However, sometimes one must infer the form of the model from the data using regression diagnostics, like model selection methods. In such cases, aliasing, which acts as high correlation between the regression coefficients, will adversely affect model selection and estimation if the design is not well chosen. As a result,

the robustness of a design is also fairly important. Theorem 1 shows that low discrepancy sampling will limit the adverse effects of aliasing on efficiency and robustness under general assumptions.

Section 2 introduces the statistical model. Section 3

2. Problem definition (statistical model description)

Suppose that an experiment has d variables, and let Ω_j be a measurable set of all possible levels of the jth variable. Common examples of Ω_j are $\{0, \cdots, q_j\}$ and $[0, 1]$. The experimental region, Ω, is some measurable subset of $\Omega_1 \times \cdots \times \Omega_d$. An experimental design with m points, $P = \{\boldsymbol{x}_i = (x_{i1}, \cdots, x_{id}) : i = 1, \cdots, m\}$, is a subset of Ω with multiple copies of the same point allowed.

Define an operator $\mathbf{L}_{\boldsymbol{x}}$, which when applied to a d-variate scalar function $f : \Omega \to \mathcal{R}$, returns its value and gradient information:

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

For a vector function $\boldsymbol{f} = (f_1, \cdots, f_l)^T : \Omega \to \mathcal{R}^l$, we extend the definition of the operator as follows:

$$\mathbf{L}_{\boldsymbol{x}} \mathbf{f}^T = \begin{pmatrix} f_1(\boldsymbol{x}) & f_2(\boldsymbol{x}) & \cdots & f_l(\boldsymbol{x}) \\ \frac{\partial f_1}{\partial x_1}(\boldsymbol{x}) & \frac{\partial f_2}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial f_l}{\partial x_1}(\boldsymbol{x}) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_1}{\partial x_d}(\boldsymbol{x}) & \frac{\partial f_2}{\partial x_d}(\boldsymbol{x}) & \cdots & \frac{\partial f_l}{\partial x_d}(\boldsymbol{x}) \end{pmatrix}$$

Let $\boldsymbol{y}_i$ denote the observed response and its first partial derivatives, when the variables take on the value $\boldsymbol{x}_i$. Then, a linear regression model including gradient information may be written as:

$$\boldsymbol{y}_i = (\mathbf{L}_{\boldsymbol{x}_i} \boldsymbol{g}^T)\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i,$$

where the specified basis g_j are linearly independent, the $\boldsymbol{\varepsilon}_i$ are independent and identically distributed random errors with mean 0 and covariance $\sigma^2 \widetilde{\boldsymbol{\Lambda}}$, with $\widetilde{\boldsymbol{\Lambda}} = \text{Diag}\left(1, \frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_d}\right)$.

This model allows us to estimate the regression coefficient $\boldsymbol{\beta}$ using both objective function values and its first partial derivatives. By assuming the

covariance of the errors to be $\sigma^2\widetilde{\mathbf{\Lambda}}$, we allow that the variances of response and its first partial derivatives to be different, and λ_i is the ratio between the variance of the response and the variance of the first partial derivative with respect to the ith variable.

Define $\mathbf{\Sigma} = \sigma^2 \text{Diag}\left(\widetilde{\mathbf{\Lambda}}, \widetilde{\mathbf{\Lambda}}, \cdots, \widetilde{\mathbf{\Lambda}}\right)$. Then from weighted least squares, the estimation of the regression coefficient $\boldsymbol{\beta}$ is given as, $\hat{\boldsymbol{\beta}} = (\mathbf{G}^T\mathbf{\Sigma}^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{\Sigma}^{-1}\boldsymbol{z}$, where $\boldsymbol{z} = (\boldsymbol{y}_1^T, \cdots, \boldsymbol{y}_m^T)^T$, $\mathbf{G}$ is the collocation matrix defined as:

$$\boldsymbol{G} = \begin{pmatrix} \mathbf{L}_{\boldsymbol{x}_1}\boldsymbol{g}^T \\ \mathbf{L}_{\boldsymbol{x}_2}\boldsymbol{g}^T \\ \vdots \\ \mathbf{L}_{\boldsymbol{x}_m}\boldsymbol{g}^T \end{pmatrix},$$

and $\mathbf{G}^T\mathbf{\Sigma}^{-1}\mathbf{G}$ is assumed to be nonsingular.

3. Scaled Integrated Mean Squared Error

By involving gradient information, one may use fewer sample points to estimate the regression model. For problem of optimal design, although we use gradient information to estimate the model, we may be interested in how the model approximates only the function value. As a result, classical optimality criteria do not apply here, and we need to explore some new optimality criterion.

Define a new function space,$\mathcal{L}^s_{2,F}(\Omega)$, which is the s-fold tensor product of $\mathcal{L}_{2,F}(\Omega)$. $\mathcal{L}^s_{2,F}(\Omega)$ is equipped with inner product defined as:

$$\langle \boldsymbol{f}, \boldsymbol{u} \rangle_{\mathcal{L}^s_{2,F}(\Omega)} = \int_\Omega \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{x}) dF_\Omega(\boldsymbol{x}), \text{ for } \forall \boldsymbol{f}, \boldsymbol{u} \in \mathcal{L}^s_{2,F}(\Omega),$$

where F_Ω is the true or target distribution function of variable $\boldsymbol{x}$ over the domain Ω. Define the norm induced by this inner product as $\|\cdot\|_{\mathcal{L}^s_{2,F}(\Omega)}$.

Define a linear operator $T : \mathcal{H} \to \mathcal{L}^s_{2,F}(\Omega)$, which acts on a scalar function f, and returns a function vector of dimension s. A simple example is: $Tf = f$, which is the embedding operator, and $s = 1$. In this example, we use the function value and gradient information of f to estimate the regression coefficient $\hat{\boldsymbol{\beta}}$, while we are interested in the prediction of the function value only. With natural extension of the definition of T, when it is applied to a

vector function $\boldsymbol{f} = (f_1, \cdots, f_l)^T$, it returns a matrix with each row will be Tf_i.

With the operator T, we could define the scaled integrated mean squared difference between the fitted response, $T\left(\boldsymbol{g}^T\hat{\boldsymbol{\beta}}\right)$, and the true one, $T\left(\boldsymbol{g}^T\boldsymbol{\beta}\right)$ as:

$$\mathrm{IMSE}(P, \boldsymbol{g}) = \frac{m}{\sigma^2} E \left\| T\left(\boldsymbol{g}^T\hat{\boldsymbol{\beta}}\right) - T\left(\boldsymbol{g}^T\boldsymbol{\beta}\right)\right\|^2_{\mathcal{L}^s_{2,F}(\Omega)}.$$

The scaling constant $\frac{m}{\sigma^2}$ is used to eliminate the effects brought by different number of sample points. For the designs with different number of sample points, it is unfair to compare them with no scaling constant. For an extreme example, consider a design P, and another design $\widetilde{P}$ which contains k copies of all the points in P. In this case, $\mathrm{IMSE}(\widetilde{P}, \boldsymbol{g}) = \mathrm{IMSE}(P, \boldsymbol{g})$ for the scaled integrated mean squared error defined above, and it should be the case.

Proposition 1. *The integrated mean squared error* $\mathrm{IMSE}(P, \boldsymbol{g})$ *defined above is identical to the trace of* $\boldsymbol{M}^{-1}\boldsymbol{A}$*, where* $\boldsymbol{M} = \int_\Omega (\mathbf{L}_{\boldsymbol{x}}\boldsymbol{g}^T)^T \widetilde{\boldsymbol{\Lambda}}^{-1} (\mathbf{L}_{\boldsymbol{x}}\boldsymbol{g}^T) dF_P(\boldsymbol{x})$*,* $F_P(\boldsymbol{x})$ *is the empirical distribution of design* P*, and* $\boldsymbol{A} = \left(\langle Tg_i, Tg_j\rangle_{\mathcal{L}^s_{2,F}(\Omega)}\right)_{i,j=1}^{m(d+1)}$.

Proof. By the linearity of T,

$$\left\| T\left(\boldsymbol{g}^T\hat{\boldsymbol{\beta}}\right) - T\left(\boldsymbol{g}^T\boldsymbol{\beta}\right)\right\|^2_{\mathcal{L}^s_{2,F}(\Omega)} = \left\| \left\| T\left(\boldsymbol{g}^T(\cdot)\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right\|^2_{l_2}\right\|^2_{\mathcal{L}^s_{2,F}(\Omega)}.$$

By the definition of $\|\cdot\|^2_{l_2}$, $\left\| T\left(\boldsymbol{g}^T(\cdot)\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right\|^2_{l_2} = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T \left(T\boldsymbol{g}(\cdot)\right)\left(T\boldsymbol{g}(\cdot)\right)^T \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$.
As $\boldsymbol{G}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{z}$, where $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^T, \cdots, \boldsymbol{\varepsilon}_m^T)^T$, we can get $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\boldsymbol{G}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{G}\right)^{-1}\boldsymbol{G}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}$.
Define $\boldsymbol{B} = \left(\boldsymbol{G}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{G}\right)^{-1}\boldsymbol{G}^T\boldsymbol{\Sigma}^{-1}$, thus,

$$\left\| T\left(\boldsymbol{g}^T(\cdot)\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right\|^2_{l_2} = \boldsymbol{\varepsilon}^T\boldsymbol{B}^T(T\boldsymbol{g}(\cdot))(T\boldsymbol{g}(\cdot))^T\boldsymbol{B}\boldsymbol{\varepsilon}.$$

As a result,

$$\left\| T\left(\boldsymbol{g}^T\hat{\boldsymbol{\beta}}\right) - T\left(\boldsymbol{g}^T\boldsymbol{\beta}\right)\right\|^2_{\mathcal{L}^s_{2,F}(\Omega)} = \boldsymbol{\varepsilon}^T\boldsymbol{B}^T\boldsymbol{A}\boldsymbol{B}\boldsymbol{\varepsilon},$$

where $\boldsymbol{A} = \left(\langle Tg_i, Tg_j\rangle_{\mathcal{L}^s_{2,F}(\Omega)}\right)_{i,j=1}^{m(d+1)}$.

Define $\boldsymbol{H} = \boldsymbol{B}^T \boldsymbol{A} \boldsymbol{B}$, then,

$$\begin{aligned}
\text{IMSE}(P, \boldsymbol{g}) &= \frac{m}{\sigma^2} E\left(\left\| T\left(\boldsymbol{g}^T \hat{\boldsymbol{\beta}}\right) - T\left(\boldsymbol{g}^T \boldsymbol{\beta}\right) \right\|^2_{\mathcal{L}^s_{2,F}(\Omega)} \right) \\
&= \frac{m}{\sigma^2} E\left(\boldsymbol{\varepsilon}^T \boldsymbol{H} \boldsymbol{\varepsilon}\right) \\
&= \frac{m}{\sigma^2} \sum_{j,k=1}^{m(d+1)} E\left(\varepsilon_j h_{jk} \varepsilon_k\right) \\
&= \frac{m}{\sigma^2} \operatorname{tr}(\boldsymbol{\Sigma} \boldsymbol{H}) \\
&= \frac{m}{\sigma^2} \operatorname{tr}(\boldsymbol{\Sigma} \boldsymbol{B}^T \boldsymbol{A} \boldsymbol{B}) \\
&= \frac{m}{\sigma^2} \operatorname{tr}(\boldsymbol{B} \boldsymbol{\Sigma} \boldsymbol{B}^T \boldsymbol{A}) \\
&= \operatorname{tr}(\boldsymbol{M}^{-1} \boldsymbol{A})
\end{aligned}$$

where h_{jk} is the element of $\boldsymbol{H}$, ε_j, ε_k is the element of $\boldsymbol{\varepsilon}$, and $\boldsymbol{M} = \left(\frac{m}{\sigma^2} \boldsymbol{B} \boldsymbol{\Sigma} \boldsymbol{B}^T\right)^{-1}$.

Since $\boldsymbol{B} = \left(\boldsymbol{G}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{G}\right)^{-1} \boldsymbol{G}^T \boldsymbol{\Sigma}^{-1}$, then $\boldsymbol{B} \boldsymbol{\Sigma} \boldsymbol{B}^T = \left(\boldsymbol{G}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{G}\right)^{-1}$. Thus,

$$\boldsymbol{M} = \frac{\sigma^2}{m} \boldsymbol{G}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{G} = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{L}_{\boldsymbol{x}_i} \boldsymbol{g}^T)^T \widetilde{\boldsymbol{\Lambda}}^{-1} (\mathbf{L}_{\boldsymbol{x}_i} \boldsymbol{g}^T) = \int_{\Omega} (\mathbf{L}_{\boldsymbol{x}} \boldsymbol{g}^T)^T \widetilde{\boldsymbol{\Lambda}}^{-1} (\mathbf{L}_{\boldsymbol{x}} \boldsymbol{g}^T) dF_P(\boldsymbol{x})$$

□

As discussed above, with the derivative information, I-optimality, which measures scaled integrated mean squared error, will be defined as $\operatorname{tr}\left(\boldsymbol{M}^{-1} \boldsymbol{A}\right)$. To achieve the optimal design under I-optimality, one should try to minimize the trace of $\boldsymbol{M}^{-1} \boldsymbol{A}$.

4. Reproducing Kernel Hilbert Space

Reference!!

To make precise statements about the effects of the design on aliasing, it is necessary to define the space, $\mathcal{H}$, containing the response, $\left(\mathbf{L}_{\boldsymbol{x}} \boldsymbol{g}^T\right) \boldsymbol{\beta}$, and its square. $\mathcal{H}$ is a separable Hilbert space with a reproducing kernel $K(\boldsymbol{x}, \boldsymbol{y})$, and it is called reproducing Hilbert space. This reproducing Hilbert space is uniquely determined by its reproducing kernel $K(\boldsymbol{x}, \boldsymbol{t})$, with

$$K(\cdot, \boldsymbol{w}) \in \mathcal{H}, \ f(\boldsymbol{w}) = \langle f, K(\cdot, \boldsymbol{w}) \rangle_{\mathcal{H}} \ \text{(for all } \boldsymbol{w} \in \Omega, f \in \mathcal{H}).$$

That is, $K(\cdot, \boldsymbol{x})$ is the representor for the functional that that evaluates a function at a point $\boldsymbol{x}$. A reproducing kernel is symmetric in its arguments and semi-positive definite:

$$K(\boldsymbol{x}, \boldsymbol{t}) = K(\boldsymbol{t}, \boldsymbol{x}) \ \ \forall \boldsymbol{x}, \boldsymbol{y} \in \Omega,$$
$$\sum_{i,k} a^{(i)} a^{(k)} K\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(k)}\right) \geq 0 \ \ \forall a^{(i)} \in \mathbb{R}, \boldsymbol{x}^{(i)} \in \Omega.$$

What's more, any function satisfying above 2 conditions is the reproducing kernel for some unique Hilbert space.

Given an integral with the form of $\int_\Omega f(\boldsymbol{x}) dF_\Omega(\boldsymbol{x})$ with integrand $f(\boldsymbol{x})$, we could use the quadrature $\frac{1}{m} \sum\limits_{\boldsymbol{x}_i \in P} f(\boldsymbol{x}_i)$ to approximate the integral, where P is the design and m is the sample size. Then, the quadrature error is defined as:

$$\text{err}(f; P) = I(f) - Q(f; P) = \int_\Omega f(\boldsymbol{x}) dF_\Omega(\boldsymbol{x}) - \frac{1}{m} \sum_{\boldsymbol{x}_i \in P} f(\boldsymbol{x}_i) = \int_\Omega f(\boldsymbol{x}) d[F_\Omega(\boldsymbol{x}) - F_P(\boldsymbol{x})],$$

where $F_\Omega(\boldsymbol{x})$ denotes the true distribution of the parameters over the experimental domain Ω,and $F_P(\boldsymbol{x}) = \frac{1}{m} \sum\limits_{i=1}^{m} 1_{[\boldsymbol{x}_i, +\infty)}(\boldsymbol{x})$ is the empirical distribution function corresponding to a design P, with $1_{\{\cdot\}}(\boldsymbol{x})$ is the indicator function.

Define the norm induced by the inner product on the reproducing kernel space $\mathcal{H}$ as $\|f\|_\mathcal{H} = \langle f, f \rangle_\mathcal{H}^{1/2}$. The variation of a function $f \in \mathcal{H}$ measures how much f fluctuates, and is defined as the norm of the non-constant part of f:

$$V(f) = V(f; K) \ \ = \ \ \begin{cases} \|f\|_\mathcal{H} & \text{if } 1 \notin \mathcal{H}, \\ \left(\|f\|_\mathcal{H}^2 - \langle f, 1 \rangle_\mathcal{H}^2 / \|1\|_\mathcal{H}^2\right)^{1/2} & \text{if } 1 \in \mathcal{H}. \end{cases}$$

The discrepancy $D(P)$, is a measure of the deviation of a design with empirical distribution F_P from the true distribution F_Ω. $D(P)$ is defined as:

$$\begin{aligned} D(P; K) \ \ &= \ \ \left[\int_{\Omega^2} K(\boldsymbol{x}, \boldsymbol{t}) d\{F_\Omega(\boldsymbol{x}) - F_P(\boldsymbol{x})\} d\{F_\Omega(\boldsymbol{t}) - F_P(\boldsymbol{t})\}\right]^{\frac{1}{2}} \\ &= \ \ \left[\int_{\Omega^2} K(\boldsymbol{x}, \boldsymbol{t}) dF_\Omega(\boldsymbol{x}) dF_\Omega(\boldsymbol{t}) - \frac{2}{m} \sum_{i=1}^{m} \int_{\Omega^2} \{K(\boldsymbol{x}_i, \boldsymbol{t})\} dF_\Omega(\boldsymbol{t}) \right. \\ &\quad \left. + \frac{1}{m^2} \sum_{i,k=1}^{m} K(\boldsymbol{x}_i, \boldsymbol{x}_k)\right]^{\frac{1}{2}} \end{aligned}$$

It has been proved that $|\operatorname{err}(f;P)|$ is upper bounded by the product of discrepancy $D(P;K)$ and variation $V(f;K)$, i.e $|\operatorname{err}(f;P)| \leq D(P;K)V(f;K)$. (Reference!!)

5. result

Theorem 1. *Suppose that $\mathcal{W}$ is a reproducing kernel Hilbert space with reproducing kernel K, and that the set of design points is P. Moreover, for all real-valued $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_l)^T$, suppose that $|\boldsymbol{g}^T\boldsymbol{\alpha}|^2 + \sum_{i=1}^{d} \lambda_i \left| \left(\frac{\partial \boldsymbol{g}}{\partial x_i}\right)^T \boldsymbol{\alpha} \right|^2$ is a function in $\mathcal{W}$. Define*

$$V_{\boldsymbol{g}} = \sup_{\|\boldsymbol{\alpha}\|_2 \leq 1} V\left(|\boldsymbol{g}^T\boldsymbol{\alpha}|^2 + \sum_{i=1}^{d} \lambda_i \left| \left(\frac{\partial \boldsymbol{g}}{\partial x_i}\right)^T \boldsymbol{\alpha} \right|^2 ; K \right),$$

which only depend on the form of the model and not on the design.

Then, $\mathrm{IMSE}(P, \boldsymbol{g}) \leq \frac{\operatorname{tr}(\boldsymbol{A})}{1 - D(P;K)V_{\boldsymbol{g}}}$, *provided that $D(P;K)V_{\boldsymbol{g}} < 1$. This upper bound is monotonically decreasing as $D(P;K)$ tends to zeros.*

Proof. Let $\widetilde{\boldsymbol{M}} = I - \boldsymbol{M}$. For any $\boldsymbol{\alpha} \in \mathbb{R}^l$

$$\begin{aligned}
\rho(\widetilde{\boldsymbol{M}}) &= \sup_{\|\boldsymbol{\alpha}\|_2 \leq 1} \left| \boldsymbol{\alpha}^T \widetilde{\boldsymbol{M}} \boldsymbol{\alpha} \right| \\
&= \sup_{\|\boldsymbol{\alpha}\|_2 \leq 1} \left| \boldsymbol{\alpha}^T \left\{ \int_{\boldsymbol{\Omega}} (\mathbf{L}_{\boldsymbol{x}} \boldsymbol{g}^T)^T \widetilde{\boldsymbol{\Lambda}}^{-1} (\mathbf{L}_{\boldsymbol{x}} \boldsymbol{g}^T) dF_{\boldsymbol{\Omega}}(\boldsymbol{x}) - \frac{1}{m} \sum_{i=1}^{m} (\mathbf{L}_{\boldsymbol{x}_i} \boldsymbol{g}^T)^T \widetilde{\boldsymbol{\Lambda}}^{-1} (\mathbf{L}_{\boldsymbol{x}_i} \boldsymbol{g}^T) \right\} \boldsymbol{\alpha} \right| \\
&= \sup_{||\boldsymbol{\alpha}||_2 \leq 1} \left| \boldsymbol{\alpha}^T \left\{ \int_{\boldsymbol{\Omega}} (\mathbf{L}_{\boldsymbol{x}} \boldsymbol{g}^T)^T \widetilde{\boldsymbol{\Lambda}}^{-1} (\mathbf{L}_{\boldsymbol{x}} \boldsymbol{g}^T) dF_{\boldsymbol{\Omega}}(\boldsymbol{x}) - \int_{\boldsymbol{\Omega}} (\mathbf{L}_{\boldsymbol{x}} \boldsymbol{g}^T)^T \widetilde{\boldsymbol{\Lambda}}^{-1} (\mathbf{L}_{\boldsymbol{x}} \boldsymbol{g}^T) dF_P(\boldsymbol{x}) \right\} \boldsymbol{\alpha} \right| \\
&= \sup_{||\boldsymbol{\alpha}||_2 \leq 1} \left| \int_{\boldsymbol{\Omega}} \boldsymbol{\alpha}^T (\mathbf{L}_{\boldsymbol{x}} \boldsymbol{g}^T)^T \widetilde{\boldsymbol{\Lambda}}^{-1} (\mathbf{L}_{\boldsymbol{x}} \boldsymbol{g}^T) \boldsymbol{\alpha} d(F_{\boldsymbol{\Omega}} - F_P)(\boldsymbol{x}) \right| \\
&= \sup_{||\boldsymbol{\alpha}||_2 \leq 1} \left| \operatorname{err}\left(|\boldsymbol{g}^T\boldsymbol{\alpha}|^2 + \sum_{i=1}^{d} \lambda_i \left| \left(\frac{\partial \boldsymbol{g}}{\partial x_i}\right)^T \boldsymbol{\alpha} \right|^2 ; P \right) \right| \\
&\leq D(P;K) \sup_{||\boldsymbol{\alpha}||_2 \leq 1} V\left(|\boldsymbol{g}^T\boldsymbol{\alpha}|^2 + \sum_{i=1}^{d} \lambda_i \left| \left(\frac{\partial \boldsymbol{g}}{\partial x_i}\right)^T \boldsymbol{\alpha} \right|^2 ; K \right) \\
&= D(P;K) V_{\boldsymbol{g}}
\end{aligned}$$

Here ρ denotes the spectral radius of a matrix. If $D(P;K)V_{\boldsymbol{g}} < 1$, then the smallest eigenvalue of $\boldsymbol{M} = I - \widetilde{\boldsymbol{M}}$ is greater than or equal to $1 - D(P;K)V_{\boldsymbol{g}} > 0$. Since $\boldsymbol{A}$ is positive semi-definite,

$$\begin{aligned} \operatorname{tr}(\boldsymbol{M}^{-1}\boldsymbol{A}) &\leq \rho(\boldsymbol{M}^{-1})\operatorname{tr}(\boldsymbol{A}) \\ &\leq \frac{1}{1-\rho(\widetilde{\boldsymbol{M}})}\operatorname{tr}(\boldsymbol{A}) \\ &\leq \frac{\operatorname{tr}(\boldsymbol{A})}{1-D(P;K)V_{\boldsymbol{g}}} \end{aligned}$$

□

6. Numerical Experiments

Fist, we tried a simple univariate case. We choose polynomials up to degree 8, i.e 1, x, x^2, x^3, x^4, x^5, x^6, x^7, x^8, as the basis. Since the basis has with pretty high order polynomials, we choose a relatively fluctuant function, $f(x) = 4 + e^{-2x}\sin(5x)$ as the testing function. We use 16 sample points to estimated the regression coefficients $\boldsymbol{\beta}$ in the model, and 1000 sample points to test the model. We repeat the experiment 100 times, and compute both maximum of the relative error of the function value and the relative root mean squared error of the function value. Figure 1 compares relative prediction errors in function value between models using Sobol points, which is low discrepancy, and models using simple random points. In the left plot, we compares relative root mean squared errors in function value, while in the right plot, we compares relative maximum errors in function value. We conclude from 1 that Sobol points work substantially better than simple random points, and result in better estimation of function values.

Then, we tried a more complicated case with $d = 2$. First, we choose the univariate polynomials that is orthogonal under the inner product, $\langle \boldsymbol{f}, \boldsymbol{h} \rangle = \int_\Omega \mathbf{L}_{\boldsymbol{x}}\boldsymbol{f}\,(\mathbf{L}_{\boldsymbol{x}}\boldsymbol{h})^T\, dF_\Omega(\boldsymbol{x})$(REFERENCE!!). Then, we use tensor product of these univariate polynomials to construct multivariate polynomials up to degree 3. We choose these multivariate polynomials up to degree 3 as our basis, and there will be totally 10 polynomials in the basis. Since compared to univariate case, in this case, we only have polynomials up to degree 3, as a result, we choose a less fluctuant function, $f(\boldsymbol{x}) = 4 + e^{-0.7x_1 - 0.5x_2}\sin(0.3x_1 + 0.6x_2)$ as the testing function. The same as in the univariate experiment, we use 16 sample points to estimate the regression coefficient $\boldsymbol{\beta}$ in the model,

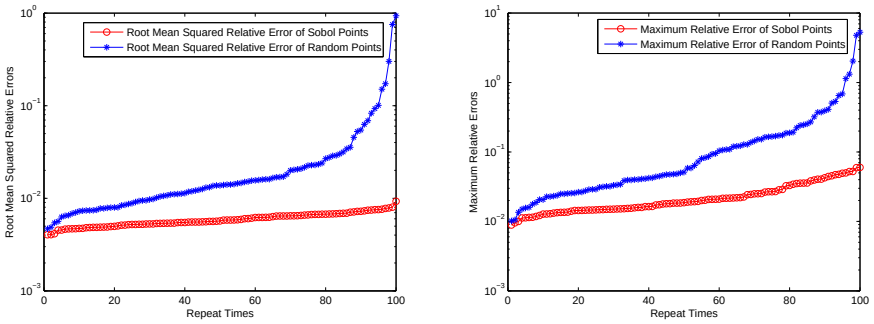

Figure 1: Relative Error Comparison between Sobol Points and Random Points

and use 1000 sample points to test the model. We repeat the experiment 100 times, and compute both maximum of the relative error of the function value and the relative root mean squared error of the function value. Figure 2 compares relative prediction errors in function value between models using Sobol points, which is low discrepancy, and models using simple random points. In the left plot, we compares relative root mean squared errors in function value, while in the right plot, we compares relative maximum errors in function value. We conclude from 2 that Sobol points work substantially better than simple random points, and result in better estimation of function values.

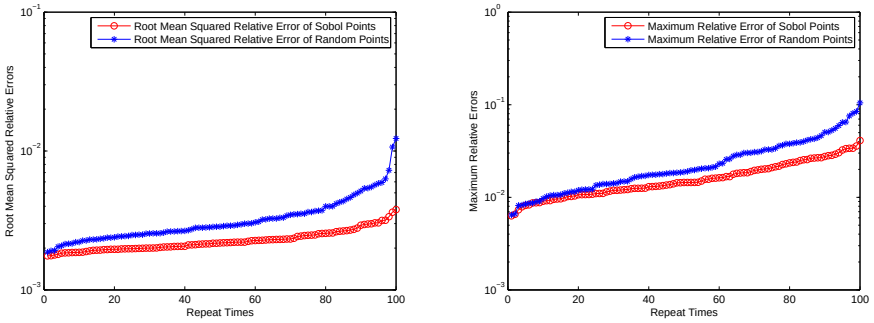

Figure 2: Relative Error Comparison between Sobol Points and Random Points

7. Optimal Design When Gradient Information Is Available

7.1. D-optimality

Continuous designs are represented by the measure ξ over Ω, where Ω is the variable domain as defined previously. If the design has n trials at n distinct points in Ω, we write:

$$\xi = \left\{ \begin{array}{cccc} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_n \\ w_1 & w_2 & \cdots & w_n \end{array} \right\},$$

where w_i is the weight on the sample point $\boldsymbol{x}_i$ with $0 \leq w_i \leq 1$, for $i = 1, \cdots, n$, and $\sum_{i=1}^{n} w_i = 1$.

Optimal designs is a class of experimental designs that are optimal with respect to some statistical criterion. In general, these statistical criterion could be considered as minimizing some convex function $\Psi\{\boldsymbol{M}(\xi)\}$, where $\boldsymbol{M}(\xi)$ is the information matrix, as defined in section 3, and this information matrix will depend on the design ξ. Many of the design criteria are called by a letter of the alphabet, so this large class of design criteria is sometimes called 'alphabetic optimality'. One of the most well-known 'alphabetic optimality' is D-optimality. By definition, a D-optimal design minimizes the content of the confidence region, which is also the volume of the ellipsoid, and this ellipsoid is defined as $(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T \boldsymbol{G}^T \boldsymbol{G}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$. It has been proved that D-optimality is equivalent to minimizing the determinant of $\boldsymbol{M}^{-1}(\xi)$(REFERENCE), in which case, we will have $\Psi\{\boldsymbol{M}(\xi)\} = -\log|\boldsymbol{M}|$.

Let the measure $\bar{\xi}$ put unit mass at the point $\boldsymbol{x}$, then the derivative of Ψ in the direction $\bar{\xi}$ is defined as:

$$\phi(\boldsymbol{x}, \xi) = \lim_{\alpha \to 0^+} \frac{1}{\alpha} \left[\Psi\{(1-\alpha)\boldsymbol{M}(\xi) + \alpha \boldsymbol{M}(\bar{\xi})\} - \Psi\{\boldsymbol{M}(\xi)\} \right].$$

Then, a well-known theorem, **The General Equivalence Theorem** states the equivalence of the following three conditions on ξ^*:

1. The design ξ^* minimizes $\Psi\{\boldsymbol{M}(\xi)\}$.
2. The design ξ^* maximizes the minimum over Ω of $\Phi(\boldsymbol{x}, \xi)$.
3. The minimum over Ω of $\Phi(\boldsymbol{x}, \xi^*) = 0$, the minimum occurring at the points of support of the design.
   As a consequence of 3, we obtain the further condition:
4. For any non-optimum design ξ, the minimum over Ω of $\Phi(\boldsymbol{x}, \xi) < 0$.

By The General Equivalence Theorem, we could check whether a design is optimal or not, and we could even find an optimal design over Ω.

7.2. D-optimal Design When Gradient Information Is Available

A nature question to ask will be whether gradient information will influence the optimal design? Our answer is yes, and we discuss this upon the quadratic regression with scalar variable, on interval $[-1, 1]$, i.e $\Omega = [-1, 1]$.

The classical model will be:

$$y_i = \boldsymbol{g}^T(x)\boldsymbol{\beta} + \varepsilon_i,$$

while, the model with gradient information will be:

$$\boldsymbol{y}_i = (\mathbf{L}_{x_i}\boldsymbol{g}^T)\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \text{(Need more specified??write as a matrix??)}$$

where $\boldsymbol{g}(x) = (1, x, x^2)^T$, ε_i are independent and identically distributed random errors for classical model with mean 0 and variance σ^2, and $\boldsymbol{\varepsilon}_i$ are independent and identically distributed random errors for the model with gradient information, with mean 0 and covariance $\sigma^2\widetilde{\boldsymbol{\Lambda}}$ as defined in section 2.

From the theory of optimal design, it is known that for classical quadratic regression model, the derivative of Ψ will be:

$$\Phi(x, \xi) = 3 - \boldsymbol{g}^T(x)\boldsymbol{M}(\xi)^{-1}\boldsymbol{g}(x),$$

where 3 comes from the number of the terms in the basis, and as a result, a D-optimal design will be(REFERENCE):

$$\xi^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{array} \right\}.$$

For the model with gradient information in univariate, given the design ξ, it is obvious that:

$$\Phi(x, \xi) = 3 - \boldsymbol{g}^T(x)\boldsymbol{M}(\xi)^{-1}\boldsymbol{g}(x) - \lambda\boldsymbol{g}'(x)^T\boldsymbol{M}(\xi)^{-1}\boldsymbol{g}'(x),$$

with $\boldsymbol{M}(\xi) = \frac{\sigma^2\boldsymbol{G}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{G}}{m} = \sum_{i=1}^{n} w_i \left(\boldsymbol{g}(x_i)\boldsymbol{g}^T(x_i) + \lambda\boldsymbol{g}'(x_i)\left(\boldsymbol{g}'(x_i)\right)^T\right)$.

Then, we will have the following theorem for the D-optimal design with the univariate quadratic model involving gradient information.

Theorem 2. *For the univariate quadratic regression model with gradient information defined above, a D-optimal design ξ^* will be:*

1. *When* $0 \leq \lambda < \frac{\sqrt{65}-7}{8}$,

$$\xi^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ w & 1-2w & w \end{array} \right\},$$

with $w = \frac{1}{6} + \frac{1}{2}\lambda + \frac{1}{6}\sqrt{1+9\lambda+21\lambda^2}$.

2. *When* $\lambda > \frac{\sqrt{65}-7}{8}$,

$$\xi^* = \left\{ \begin{array}{cc} -1 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{array} \right\}.$$

Proof. Suppose a D-optimal design is assumed to be:

$$\xi^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ w & 1-2w & w \end{array} \right\}.$$

It is obvious that $w \neq 0$, since if $w = 0$, 0 will be the only sample point, and the information matrix $\boldsymbol{M}(\xi)$ will be singular.

Then, since $\boldsymbol{M}(\xi) = \sum_{i=1}^{n} w_i \left(\boldsymbol{g}(x_i)\boldsymbol{g}^T(x_i) + \lambda \boldsymbol{g}'(x_i) \left(\boldsymbol{g}'(x_i)\right)^T \right)$, we will get,

$$\boldsymbol{M}(\xi^*) = \left(\begin{array}{ccc} 1 & 0 & 2w \\ 0 & 2w+\lambda & 0 \\ 2w & 0 & w(2+8\lambda) \end{array} \right).$$

Then,

$$\boldsymbol{M}^{-1}(\xi^*) = \left(\begin{array}{ccc} -\frac{1+4\lambda}{2w-1-4\lambda} & 0 & \frac{1}{2w-1-4\lambda} \\ 0 & \frac{1}{2w+\lambda} & 0 \\ \frac{1}{2w-1-4\lambda} & 0 & -\frac{1}{2(2w-1-4\lambda)w} \end{array} \right).$$

By $\Phi(x,\xi) = 3 - \boldsymbol{g}^T(x)\boldsymbol{M}(\xi)^{-1}\boldsymbol{g}(x) - \lambda \boldsymbol{g}'(x)^T \boldsymbol{M}(\xi)^{-1}\boldsymbol{g}'(x)$, we will have

$$\begin{aligned} \Phi(x,\xi^*) &= \frac{1}{2(2w-1-4\lambda)w}x^4 + \left(\frac{2\lambda}{(2w-1-4\lambda)w} - \frac{2}{2w-1-4\lambda} - \frac{1}{2w+\lambda} \right) x^2 \\ &\quad +3 - \frac{\lambda}{2w+\lambda} + \frac{1+4\lambda}{2w-1-4\lambda}. \end{aligned}$$

As $2w - 1 \leq 0$, $w > 0$, and $\lambda \geq 0$, $\Phi(x,\xi^*)$ is a parabola of x^2 with the coefficient of x^4, $\frac{1}{2(2w-1-4\lambda)w} < 0$. As a result, the minimum of $\Phi(x,\xi^*)$ can only occurs when $x^2 = 0$ or $x^2 = 1$.

When $x^2 = 0$,

$$\Phi(0, \xi^*) = \frac{12w^2 - 4w - 12w\lambda - \lambda - 4\lambda^2}{(2w + \lambda)(2w - 1 - 4\lambda)} = \frac{(w - w_{0+})(w - w_{0-})}{(2w + \lambda)(2w - 1 - 4\lambda)},$$

where $w_{0+} = \frac{1}{6} + \frac{1}{2}\lambda + \frac{1}{6}\sqrt{1 + 9\lambda + 21\lambda^2}$, and $w_{0-} = \frac{1}{6} + \frac{1}{2}\lambda - \frac{1}{6}\sqrt{1 + 9\lambda + 21\lambda^2} \leq 0$, when $\lambda \geq 0$. It is obvious that $(2w + \lambda)(2w - 1 - 4\lambda) < 0$.

When $x^2 = 1$,

$$\begin{aligned} \Phi(\pm 1, \xi^*) &= \frac{24w^3 - 8w^2 - 32w^2\lambda + 10w\lambda + 4\lambda^2 - 12w^2 + 4w + \lambda}{2(2w + \lambda)(2w - 1 - 4\lambda)w} \\ &= \frac{(w - \frac{1}{2})(w - w_{0+})(w - w_{0-})}{(2w + \lambda)(2w - 1 - 4\lambda)w} \end{aligned}$$

It is obvious that $(2w + \lambda)(2w - 1 - 4\lambda)w < 0$.

1. When $0 \leq \lambda \leq \frac{\sqrt{65} - 7}{8}$, from simple calculation, we will get

   $$0 < w_{0+} = \frac{1}{6} + \frac{1}{2}\lambda + \frac{1}{6}\sqrt{1 + 9\lambda + 21\lambda^2} \leq \frac{1}{2}.$$

   The design,

   $$\xi^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ w & 1 - 2w & w \end{array} \right\},$$

   with $w = \frac{1}{6} + \frac{1}{2}\lambda + \frac{1}{6}\sqrt{1 + 9\lambda + 21\lambda^2}$, will make $\Phi(0, \xi^*) = 0$, and $\Phi(\pm 1, \xi^*) = 0$, which means when $x \in [-1, 1]$, $\Phi(x, \xi^*) \geq 0$, thus, by The General Equivalence Theorem, ξ^* is a D-optimal design.
   What's more, if $w = \frac{1}{2}$, we will get $\Phi(0, \xi^*) < 0$, which means

   $$\xi = \left\{ \begin{array}{cc} -1 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{array} \right\}$$

   could not be a D-optimal design.
2. When $\lambda > \frac{\sqrt{65} - 7}{8}$, the design

   $$\xi^* = \left\{ \begin{array}{cc} -1 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{array} \right\}$$

   will make $\Phi(0, \xi^*) > 0$, and $\Phi(\pm 1, \xi^*) = 0$, which means when $x \in [-1, 1]$, $\Phi(x, \xi^*) \geq 0$, thus, by The General Equivalence Theorem, ξ^* is a D-optimal design.

□

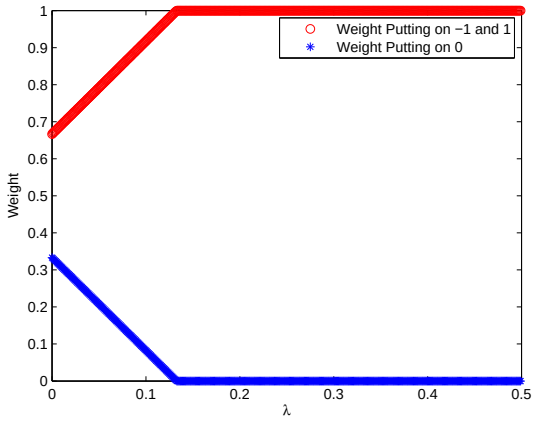

Figure 3: D Optimal Design Design Corresponding to the Change of λ

7.3. I-optimal Design When Gradient Information Is Available

Now, we will define an optimality criteria that corresponding to the scaled integrated mean squared error defined in section 3, for which the optimal design minimizes the scaled integrated mean squared error. We will call this optimality criteria I optimality, since it is analogous to classical I optimality without gradient information(REFERENCE). In this case, as proved in proposition 1, we will have $\Psi\{\boldsymbol{M}(\xi)\} = \mathrm{tr}(\boldsymbol{M}^{-1}(\xi)\boldsymbol{A})$, where as defined in section 3, $\boldsymbol{A} = \left(\langle Tg_i, Tg_j\rangle_{\mathcal{L}^s_{2,F}(\Omega)}\right)_{i,j=1}^{m(d+1)}$.

Proposition 2. *For I-optimality, the derivative of* Ψ *will be:*

1. *for the classical model without gradient information:*

$$y_i = \boldsymbol{g}^T(x)\boldsymbol{\beta} + \varepsilon_i,$$

where ε_i *are independent and identically distributed random errors for classical model with mean* 0 *and variance* σ^2*.*

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

2. *for the model with gradient information:*

$$\boldsymbol{y}_i = (\mathbf{L}_{x_i}\boldsymbol{g}^T)\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i,$$

where $\boldsymbol{\varepsilon}_i$ are independent and identically distributed random errors for the model with gradient information, with mean 0 and covariance $\sigma^2\widetilde{\boldsymbol{\Lambda}}$ as defined in section 2.

$$\begin{aligned}\Phi(x,\xi) &= \operatorname{tr}(\boldsymbol{M}^{-1}(\xi)\boldsymbol{A}) - \boldsymbol{g}^T(\boldsymbol{x})\boldsymbol{M}^{-1}(\xi)\boldsymbol{A}\boldsymbol{M}^{-1}(\xi)\boldsymbol{g}(\boldsymbol{x}) \\ &\quad - \sum_{i=1}^{d}\lambda_i\left(\frac{\partial \boldsymbol{g}}{\partial x_i}\right)^T(\boldsymbol{x})\boldsymbol{M}^{-1}(\xi)\boldsymbol{A}\boldsymbol{M}^{-1}(\xi)\frac{\partial \boldsymbol{g}}{\partial x_i}(\boldsymbol{x})\end{aligned}$$

Proof. First, we compute the Gâteaux derivative of Ψ at $\boldsymbol{M}_1$ in the direction of $\boldsymbol{M}_2$, which is defined as:

$$G_\Psi(\boldsymbol{M}_1,\boldsymbol{M}_2) = \lim_{\varepsilon\to 0^+}\frac{1}{\varepsilon}\left\{\Psi(\boldsymbol{M}_1+\varepsilon\boldsymbol{M}_2) - \Psi(\boldsymbol{M}_1)\right\}.$$

$$\begin{aligned}G_\Psi(\boldsymbol{M}_1,\boldsymbol{M}_2) &= \lim_{\varepsilon\to 0^+}\frac{1}{\varepsilon}\{\operatorname{tr}[(\boldsymbol{M}_1+\varepsilon\boldsymbol{M}_2)^{-1}\boldsymbol{A}] - \operatorname{tr}(\boldsymbol{M}_1^{-1}\boldsymbol{A})\} \\ &= \lim_{\varepsilon\to 0^+}\frac{1}{\varepsilon}\{\operatorname{tr}[((\boldsymbol{M}_1+\varepsilon\boldsymbol{M}_2)^{-1} + \boldsymbol{M}_1^{-1})\boldsymbol{A}]\} \\ &= \lim_{\varepsilon\to 0^+}\frac{1}{\varepsilon}\{\operatorname{tr}[((I+\varepsilon\boldsymbol{M}_1^{-1}\boldsymbol{M}_2)^{-1}\boldsymbol{M}_1^{-1} - \boldsymbol{M}_1^{-1})\boldsymbol{A}]\} \\ &= \lim_{\varepsilon\to 0^+}\frac{1}{\varepsilon}\{\operatorname{tr}[((I-\varepsilon\boldsymbol{M}_1^{-1}\boldsymbol{M}_2+\mathcal{O}(\varepsilon^2))\boldsymbol{M}_1^{-1} - \boldsymbol{M}_1^{-1})\boldsymbol{A}]\} \\ &= \lim_{\varepsilon\to 0^+}\frac{1}{\varepsilon}\{\operatorname{tr}(-\varepsilon\boldsymbol{M}_1^{-1}\boldsymbol{M}_2\boldsymbol{M}_1^{-1}\boldsymbol{A}) + \mathcal{O}(\varepsilon^2)\} \\ &= -\operatorname{tr}(\boldsymbol{M}_1^{-1}\boldsymbol{M}_2\boldsymbol{M}_1^{-1}\boldsymbol{A})\end{aligned}$$

Then, we will compute the Fréchet derivative of Ψ at $\boldsymbol{M}_1$ in the direction of $\boldsymbol{M}_2$, which is defined as:

$$F_\Psi(\boldsymbol{M}_1,\boldsymbol{M}_2) = \lim_{\varepsilon\to 0^+}\frac{1}{\varepsilon}\left\{\Psi((1-\varepsilon)\boldsymbol{M}_1+\varepsilon\boldsymbol{M}_2) - \Psi(\boldsymbol{M}_1)\right\}.$$

Since it is easy to show that $F_\Psi(\boldsymbol{M}_1,\boldsymbol{M}_2) = G_\Psi(\boldsymbol{M}_1,\boldsymbol{M}_2-\boldsymbol{M}_1)$, we will get,

$$\begin{aligned}F_\Psi(\boldsymbol{M}_1,\boldsymbol{M}_2) &= -\operatorname{tr}[(\boldsymbol{M}_1^{-1}\boldsymbol{M}_2\boldsymbol{M}_1^{-1} - \boldsymbol{M}_1^{-1})\boldsymbol{A}] \\ &= -\operatorname{tr}(\boldsymbol{M}_1^{-1}\boldsymbol{M}_2\boldsymbol{M}_1^{-1}\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{M}_1^{-1}\boldsymbol{A}) \\ &= -\operatorname{tr}(\boldsymbol{M}_2\boldsymbol{M}_1^{-1}\boldsymbol{A}\boldsymbol{M}_1^{-1}) + \operatorname{tr}(\boldsymbol{M}_1^{-1}\boldsymbol{A})\end{aligned}$$

Since $\Phi(\boldsymbol{x},\xi) = F_\Psi(\boldsymbol{M}(\xi),\boldsymbol{M}(\bar{\xi}))$, where $\bar{\xi}$ put unit mass at the point $\boldsymbol{x}$, from simple calculation, we will have,

1. for the classical model without gradient information:

$$\Phi(\boldsymbol{x},\xi) = F_\Psi(\boldsymbol{M}(\xi), \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}^T(\boldsymbol{x})) = \mathrm{tr}(\boldsymbol{M}^{-1}(\xi)\boldsymbol{A}) - \boldsymbol{g}^T(\boldsymbol{x})\boldsymbol{M}^{-1}(\xi)\boldsymbol{A}\boldsymbol{M}^{-1}(\xi)\boldsymbol{g}(\boldsymbol{x})$$

2. for the model with gradient information:

$$\begin{aligned}\Phi(x,\xi) &= F_\Psi\left(\boldsymbol{M}(\xi), \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}^T(\boldsymbol{x}) + \sum_{i=1}^{d}\lambda_i(\boldsymbol{x})\frac{\partial \boldsymbol{g}}{\partial x_i}(\boldsymbol{x})\left(\frac{\partial \boldsymbol{g}}{\partial x_i}\right)^T\right)\\ &= \mathrm{tr}(\boldsymbol{M}^{-1}(\xi)\boldsymbol{A}) - \boldsymbol{g}^T(\boldsymbol{x})\boldsymbol{M}^{-1}(\xi)\boldsymbol{A}\boldsymbol{M}^{-1}(\xi)\boldsymbol{g}(\boldsymbol{x})\\ &\quad -\sum_{i=1}^{d}\lambda_i\left(\frac{\partial \boldsymbol{g}}{\partial x_i}\right)^T(\boldsymbol{x})\boldsymbol{M}^{-1}(\xi)\boldsymbol{A}\boldsymbol{M}^{-1}(\xi)\frac{\partial \boldsymbol{g}}{\partial x_i}(\boldsymbol{x})\end{aligned}$$

□

Then, as the same in section 7.2, we will discuss upon the quadratic regression with scalar variable, on interval $[-1,1]$, i.e $\Omega = [-1,1]$. For simplicity, in the following discussion, we will define $Tf = f$, then $\boldsymbol{A} = \int_\Omega \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}^T(\boldsymbol{x})dF_P(\boldsymbol{x})$. In general, we could have T to be any linear and bounded operator, and as long as it is determined, we could always find an I optimal design corresponding to a particular T.

The classical model will be:

$$y_i = \boldsymbol{g}^T(x)\boldsymbol{\beta} + \varepsilon_i,$$

while, the model with gradient information will be:

$$\boldsymbol{y}_i = (\mathbf{L}_{x_i}\boldsymbol{g}^T)\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \text{(Need more specified??write as a matrix??)}$$

where $\boldsymbol{g}(x) = (1, x, x^2)^T$, ε_i are independent and identically distributed random errors for classical model with mean 0 and variance σ^2, and $\boldsymbol{\varepsilon}_i$ are independent and identically distributed random errors for the model with gradient information, with mean 0 and covariance $\sigma^2\widetilde{\boldsymbol{\Lambda}}$ as defined in section 2.

For the classical quadratic regression model, as proved in proposition 2, we have,

$$\Phi(x,\xi) = \mathrm{tr}\left(\boldsymbol{M}^{-1}(\xi)\boldsymbol{A}\right) - \boldsymbol{g}^T(x)\boldsymbol{M}^{-1}(\xi)\boldsymbol{A}\boldsymbol{M}^{-1}(\xi)\boldsymbol{g}(x).$$

It is easy to prove by the 3rd condition of the General Equivalence Theorem that the design:

$$\xi^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 1/4 & 1/2 & 1/4 \end{array} \right\}$$

is an I optimal design.

For the model with gradient information in univariate, given the design ξ, as proved in proposition 2, we have,

$$\Phi(x,\xi) = \mathrm{tr}\left(\boldsymbol{M}^{-1}(\xi)\boldsymbol{A}\right) - \boldsymbol{g}^T(x)\boldsymbol{M}^{-1}(\xi)\boldsymbol{A}\boldsymbol{M}^{-1}(\xi)\boldsymbol{g}(x) - \lambda \boldsymbol{g}'(x)^T\boldsymbol{M}^{-1}(\xi)\boldsymbol{A}\boldsymbol{M}^{-1}(\xi)\boldsymbol{g}'(x).$$

Then, we will have the following theorem for the I optimal design with univariate quadratic model involving gradient information.

Theorem 3. *For the univariate quadratic regression model with gradient information defined above, we will have an I optimal design ξ^*,*

$$\xi^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ w & 1-2w & w \end{array} \right\},$$

with w is the root that lies in $[0, 0.5]$ of the $p(w)$, which is a polynomial of degree 4,

$$960w^4\lambda + \left(128 + 960\lambda^2 + 400\lambda\right) w^3 + \left(240\lambda^3 - 300\lambda^2 - 160\lambda - 32\right) w^2 - \left(36\lambda^2 + 12\lambda\right) w - \left(3\lambda^2 + 12\lambda^3\right).$$

Later in the proof, we will demonstrate numerically that there will be such a solution that lies in $[0, 0.5]$.

Proof. As given,

$$\xi^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ w & 1-2w & w \end{array} \right\},$$

it is obvious that $w \neq 0$, since if $w = 0$, 0 will be the only sample point, and the information matrix $\boldsymbol{M}(\xi)$ will be singular.

Then, since $\boldsymbol{M}(\xi) = \sum_{i=1}^{n} w_i \left(\boldsymbol{g}(x_i)\boldsymbol{g}^T(x_i) + \lambda \boldsymbol{g}'(x_i)\left(\boldsymbol{g}'(x_i)\right)^T\right)$, we will get,

$$\boldsymbol{M}(\xi^*) = \left(\begin{array}{ccc} 1 & 0 & 2w \\ 0 & 2w+\lambda & 0 \\ 2w & 0 & w(2+8\lambda) \end{array} \right).$$

Then,

$$\boldsymbol{M}^{-1}(\xi^*) = \begin{pmatrix} -\frac{1+4\lambda}{2w-1-4\lambda} & 0 & \frac{1}{2w-1-4\lambda} \\ 0 & \frac{1}{2w+\lambda} & 0 \\ \frac{1}{2w-1-4\lambda} & 0 & -\frac{1}{2(2w-1-4\lambda)w} \end{pmatrix}.$$

Since $\boldsymbol{A} = \int_\Omega \boldsymbol{g}(x)\boldsymbol{g}^T(x)dF_P(x)$, and $\boldsymbol{g}(x) = (1, x, x^2)^T$, we will have,

$$\boldsymbol{A} = \frac{1}{2}\int_{-1}^{1} \boldsymbol{g}(x)\boldsymbol{g}^T(x)dx = \frac{1}{2}\int_{-1}^{1} \begin{pmatrix} 1 & x & x^2 \\ x & x^2 & x^3 \\ x^2 & x^3 & x^4 \end{pmatrix} dx = \begin{pmatrix} 1 & 0 & \frac{1}{3} \\ 0 & \frac{1}{3} & 0 \\ \frac{1}{3} & 0 & \frac{1}{5} \end{pmatrix}.$$

As a result,

$$\begin{aligned} \mathrm{tr}\left(\boldsymbol{M}^{-1}\boldsymbol{A}\right) &= \mathrm{tr}\begin{pmatrix} -\frac{2(1+6\lambda)}{3(2w-4\lambda-1)} & 0 & -\frac{2(1+10\lambda)}{15(2w-4\lambda-1)} \\ 0 & \frac{1}{3(2w+\lambda)} & 0 \\ \frac{6w-1}{6w(2w-4\lambda-1)} & 0 & \frac{10w-3}{30w(2w-4\lambda-1)} \end{pmatrix} \\ &= -\frac{240\lambda w^2 + (120\lambda^2 + 50\lambda + 16)w + 3\lambda}{30(2w-4\lambda-1)(2w+\lambda)w} \end{aligned}$$

As a result, we will get,

$$\begin{aligned} & \Phi(x, \xi^*) \\ = \; & \mathrm{tr}\left(\boldsymbol{M}^{-1}(\xi)\boldsymbol{A}\right) - \boldsymbol{g}^T(x)\boldsymbol{M}^{-1}(\xi)\boldsymbol{A}\boldsymbol{M}^{-1}(\xi)\boldsymbol{g}(x) - \lambda \boldsymbol{g}'(x)^T\boldsymbol{M}^{-1}(\xi)\boldsymbol{A}\boldsymbol{M}^{-1}(\xi)\boldsymbol{g}'(x) \\ = \; & -\frac{1}{60(2w-4\lambda-1)^2(2w+\lambda)^2w^2}\left(Coef_2 x^4 + Coef_1 x^2 + Coef_0\right) \end{aligned}$$

where

$$Coef_2 = 240w^4 + (240\lambda - 80)w^3 + \left(60\lambda^2 - 80\lambda + 12\right)w^2 - \left(20\lambda^2 - 12\lambda\right)w + 3\lambda^2,$$

$$\begin{aligned} Coef_1 &= (-960\lambda - 240)w^4 + \left(-960\lambda^2 - 640\lambda - 48\right)w^3 + \left(-240\lambda^3 + 240\lambda^2 + 240\lambda + 20\right)w^2 \\ &\quad + 56\lambda^2 w + 12\lambda^3 \end{aligned}$$

and

$$\begin{aligned} Coef_0 &= 1920\lambda w^5 + \left(1920\lambda^2 + 800\lambda + 256\right)w^4 + \left(480\lambda^3 - 600\lambda^2 - 320\lambda - 64\right)w^3 \\ &\quad - \left(72\lambda^2 + 24\lambda\right)w^2 - \left(24\lambda^3 + 6\lambda^2\right)w. \end{aligned}$$

It is easy to prove that,

$$\begin{aligned} Coef_2 &= 240w^4 + (240\lambda - 80)w^3 + \left(60\lambda^2 - 80\lambda + 12\right) w^2 - \left(20\lambda^2 - 12\lambda\right) w + 3\lambda^2 \\ &= \left(60w^2 - 20w + 3\right)(2w + \lambda)^2 \\ &= \left(60\left(w - \frac{1}{6}\right)^2 + \frac{4}{3}\right)(2w + \lambda)^2 > 0. \end{aligned}$$

As a result, it is obvious that $\Phi(x, \xi^*)$ is a parabola of x^2 with the coefficient of x^4 being negative. Then, the minimum of $\Phi(x, \xi^*)$ can only occur when $x^2 = 0$ or $x^2 = 1$.

When $x^2 = 0$,

$$\Phi(0, \xi^*) = -\frac{p(w)}{30(2w - 4\lambda - 1)^2(2w + \lambda)^2 w},$$

where $p(w) = 960w^4\lambda + (128 + 960\lambda^2 + 400\lambda)\, w^3 + (240\lambda^3 - 300\lambda^2 - 160\lambda - 32)\, w^2 - (36\lambda^2 + 12\lambda)\, w - (3\lambda^2 + 12\lambda^3)$.

When $x^2 = 1$,

$$\Phi(0, \xi^*) = -\frac{\left(w - \frac{1}{2}\right) p(w)}{30(2w - 4\lambda - 1)^2(2w + \lambda)^2 w^2}.$$

It would be easy to check that when $w = \frac{1}{2}$, $\Phi(0, \xi^*) = -\frac{48\lambda^3 + 24\lambda^2 + 64\lambda + 8}{30(2w - 4\lambda - 1)^2(2w + \lambda)^2 w}$, which is always negative when $\lambda \geq 0$ and $w > 0$, and this means that

$$\xi^* = \left\{ \begin{array}{cc} -1 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{array} \right\}$$

can not be an I-optimal design.

As a result, an I-optimal design ξ^* will be, ξ^*,

$$\xi^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ w & 1 - 2w & w \end{array} \right\},$$

with w is the root that lies in $[0, 0.5]$ of the $p(w)$, which is a polynomial of degree 4, $p(w) = 960w^4\lambda + (128 + 960\lambda^2 + 400\lambda)\, w^3 + (240\lambda^3 - 300\lambda^2 - 160\lambda - 32)\, w^2 - (36\lambda^2 + 12\lambda)\, w - (3\lambda^2 + 12\lambda^3)$.

Now, let's look at the numerical result of the I-optimal design done by Matlab , with λ varying from 0 to 99999, which in application 99999 for λ is far large enough.

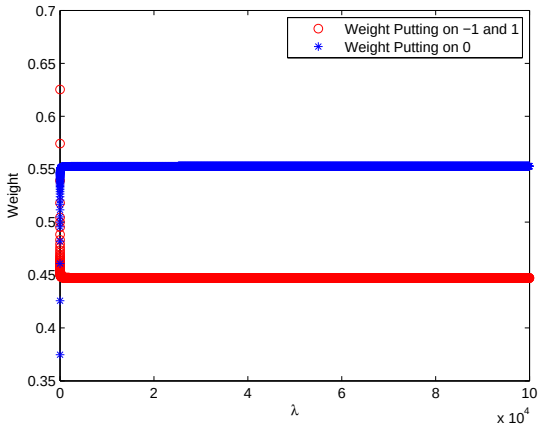

Figure 4: I Optimal Design Design Corresponding to the Change of λ, $\lambda \in [0, 99999]$

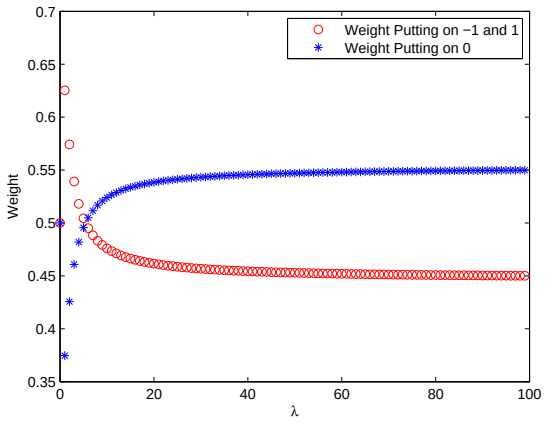

Figure 5: I Optimal Design Design Corresponding to the Change of λ, $\lambda \in [0, 99]$

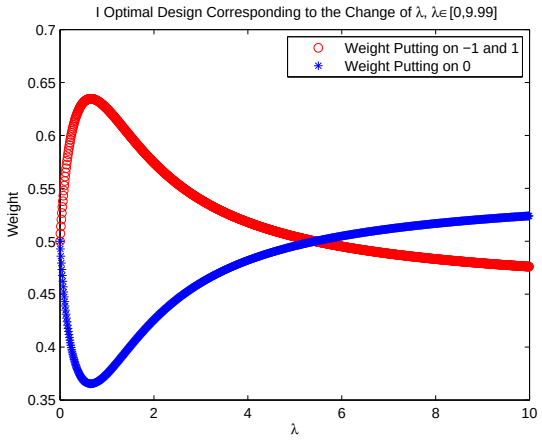

Figure 6: I Optimal Design Design Corresponding to the Change of λ, $\lambda \in [0, 9.99]$

We could see from figure 4 and the numerical result, by the continuity of the root of $p(w)$ with respect to λ, when λ tends to be very large, the I optimal design will tend to be stable, and will be very close to

$$\xi^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 0.2236 & 0.5528 & 0.2236 \end{array} \right\}.$$

From figure 5 and figure 6, we could see that when $\lambda = 0$, the I optimal design will be

$$\xi^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 1/4 & 1/2 & 1/4 \end{array} \right\},$$

which is consistent with the I optimal design for the classical model without gradient information. And, with λ getting large from 0, the weight on -1 and 1 first increases and exceeds the weight on 0, and then decreases and finally will keep below the weight on 0. Also, from the numerical result, when $\lambda = 1$, which means the variance of the function value and the variance of the first derivative value is the same, we will have the I optimal design as:

□