Hellinger Information and Optimal Design for Nonregular Models

by

Yi Lin
B.A, (Loyola University) 2009

Thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Mathmatics
in the Graduate College of the
University of Illinois at Chicago, 2018

Chicago, Illinois

Defense Committee:
Min Yang, Chair and Advisor
Ryan Martin, Co-Advisor, North Carolina State University
Dibyen Majumdar
Jie Yang
George Karabatsos, Educational Psychology

To my family and friends.

ACKNOWLEDGMENTS

First, I would like to express my gratitude and deepest appreciation to my advisers, Professor Ryan Martin and Professor Min Yang, for their guidance and encouragement for the last several years. The inspiration for this project came from classes and conversations with Prof. Martin, who encouraged me to look into nonregular cases in statistics. Prof. Min Yang's deep understanding of topics in optimal design provided invaluable assistance at every stage.

While my advisers provided the most direct help in developing this thesis, I could not have written it without the support of the faculty, graduate students, and administration of the University of Illinois at Chicago. In particular, I thank Dr. Dibyen Majumdar, who first introduced me to the topic of optimal design, and Dr. Jie Yang and Dr. George Karabatsos, who graciously served on my thesis committee, offering valuable comments and advice.

I am indebted to my fellow graduate students who, inside and outside of classrooms, helped me in numerous ways.

I would like to express my love and gratitude to my family, in China and in the USA: My parents, Song Lin and Weilan Fu; my grandparents, whose dedication to scientific inquiry has set a high standard I aspire to meet; and my in-laws, Marie and Lamont Smith. Without their love and support, this thesis, and so much else, would not have been possible. Finally, I would like to thank my husband, Nate Smith, who has helped me along in too many ways to list.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

SUMMARY

Classically, the Fisher information is the relevant object for defining optimal experimental designs. However, for models that lack certain regularity, the Fisher information does not exist and, hence, no notion of design optimality is available in the literature for such situations. This thesis fills this gap by proposing a so-called Hellinger information that generalizes Fisher information in the sense that the two measures agree in regular problems, but the former also exists in certain nonregular problems. A Hellinger information inequality is derived, showing that the Hellinger information defines a lower bound on the local minimax risk of estimators. This provides a connection between features of the underlying model, in particular, the design and the performance of estimators, motivating the use of this new Hellinger information for defining a notion of design optimality in nonregular problems. Hellinger optimal designs are derived for several nonregular regression problems, and numerical results are shown to demonstrate empirically the improved efficiency of these designs compared to alternatives.

CHAPTER 1

INTRODUCTION

1.1 Regular models and Fisher information

In the optimal design literature, the focus is on regular models, where the response variable follows a distribution from an exponential family. In such cases, Fisher information emerges as the relevant object and optimal designs are defined as those with maximal Fisher information. However, for nonregular models, such as a linear regression model with error terms following an exponential distribution, Fisher information does not exist. This project provides a new measure of information that is suitable for constructing optimal designs for nonregular models.

Section 1 of this chapter reviews the theoretical background for optimal design of experiment for regular models, including a discussion of Fisher information. This discussion emphasizes that the primary motivation for the optimization of Fisher information is its connection to the quality of estimator, as this will be a key point in the subsequent chapters. Section 2 covers basic concepts and notations in the mathematics of experimental design and reviews recent developments in the strategy of optimization of the Fisher information matrix. Section 3 introduces the types of nonregular models that can benefit from the approach to optimal design proposed in subsequent chapters. Finally, Section 4 of this chapter outlines the rest of this thesis.

1.1.1 Fisher information and regular models, asymptotic normality

Suppose that the parametric family of probability measures $\{P_\theta, \theta \in \Theta\}$ is defined on a measurable space $(\Omega, \mathcal{B})$, so that P_θ is dominated by some σ-finite measure μ on $\mathcal{B}$, and the choice of dominating measure μ is defined for all points of the parametric family. Following (Lehmann, 1999), the model is regular if it satisfies the following conditions:

- C1) the parameter is identifiable, i.e., for all $\theta \in \Theta, p_\theta \neq p_{\theta'} \Leftrightarrow \theta \neq \theta'$;
- C2) the parameter space $\Theta \subset \mathbb{R}^d, d \geq 1$ is an open set;
- C3) common support: the set $\{y : p_\theta(y) > 0\}$ is independent of θ;
- C4) $p_\theta(y)$ is differentiable with respect to θ for all y.

Under these conditions, Fisher information exists.

Definition 1. (Fisher information) Assume $\{P_\theta, \theta \in \Theta\}$ satisfies $C1) - C4)$ above and define $\dot{l}_\theta = (\frac{\partial \log p_\theta(y)}{\partial \theta_1}, ..., \frac{\partial \log p_\theta(y)}{\partial \theta_d})^\top$. Then the Fisher information matrix has dimension $d \times d$ and is defined as

$$\mathcal{I}(\theta) = E_\theta(\dot{l}_\theta \dot{l}_\theta^\top).$$

Consider the following additional condition:

(C5) The first two derivatives with respect to $p_\theta(y)$ exist for all y, and for all $\theta \in \Theta$, the corresponding differentiation with respect to θ of $\int p_\theta(y)dy$ can be obtained by differentiating under the integral sign. Under the additional condition (C5), there is an alternative formula for Fisher information,

$$\mathcal{I}(\theta) = -E_\theta(\ddot{l}_\theta).$$

There are two properties of Fisher information that will be important in the course of the argument here. First, Fisher information is additive. If $Y_1, ..., Y_n$ are independently but not identically distributed, with densities that share the same parameter $p_{i,\theta}, i = 1, ..., n$, letting $\mathcal{I}_{(i)}(\theta)$ be the Fisher information from the ith observation, then the information obtained in a sample $(y_1, ..., y_n)$, denoted as $\mathcal{I}_n(\theta)$, would be $\sum_{i=1}^{n} \mathcal{I}_{(i)}(\theta)$. If $Y_1, ..., Y_n$ are independently and identically distributed (iid), then information for all y_i would be the same, i.e., $\mathcal{I}_{(i)}(\theta) = \mathcal{I}(\theta)$. So, the information in the sample would be $\mathcal{I}_n(\theta) = n\mathcal{I}(\theta)$. Second, Fisher information has a formulation for reparametrization: If $\eta = g(\theta)$, assuming $g(\theta)$ is one-to-one, differentiable, and $g(.) \neq 0$, then $\tilde{I}(\eta)$ is the Fisher information for η, such that

$$\tilde{\mathcal{I}}(\eta) = \{(g^{-1}(\theta))'\}^T \mathcal{I}(g^{-1}(\eta))\{(g^{-1}(\theta))'\}.$$

A class of regular models is the exponential family of distributions. If $p_\theta(y)$ belongs to the exponential family, then there are some non-negative functions $h(.), T(.), A(.)$, such that $p_\theta(y)$ can be expressed as

$$p_\theta(y) = h(y)e^{\theta^\top T(x) - A(\theta)}, \text{ where } e^{A(\theta)} = \int h(y)e^{\theta^\top T(y)}dy.$$

The normal, exponential, Bernoulli, Poisson distributions, and many other distributions belong to the exponential distribution family. Probability distribution functions of the exponential distribution family satisfy the regularity conditions C(1) through C(5), thus Fisher information exists for them as well. For example, the Fisher information for the location parameter

of normal distribution, $N(\mu, \sigma^2)$, with known variance σ^2, is $I(\mu) = \sigma^{-2}$. The Poisson distribution with probability distribution function, $p_\lambda(y) = \lambda^y e^{-\lambda}(y!)^{-1}$, has Fisher information, $I(\lambda) = \lambda^{-1}$.

Another important concept related to Fisher information is the Cramér-Rao lower bound, which states that Fisher information inversely bounds the variance of any unbiased estimator from below.

Theorem 1 (Cramér-Rao Lower bound)**.** *Consider regular model* $\{P_\theta, \theta \in \Theta\}$*. If* $\mathbf{Y^n} = (Y_1, .., Y_n)^\top, Y_1, .., Y_n \overset{iid}{\sim} p_\theta$ *and if* $T(\mathbf{Y^n})$ *is a real-valued statistics with* $E_\theta(T) = \tau(\theta)$*, then, letting* $\tau'(\theta)$ *be a vector of the derivative of* $\tau(\theta)$*, the Cramér-Rao lower bound is*

$$V_\theta(T) \geq \tau'(\theta)^T (n\mathcal{I}(\theta))^{-1} \tau'(\theta). \tag{1.1}$$

The Cramér-Rao lower bound can be attained only in exponential families. Asymptotically, all regular models belong to the exponential family. (Lehmann and Casella, 1998) Specifically, the bound can be attained by maximum likelihood estimator (MLE), $\hat{\theta}_n$, the solution to the likelihood equation, $\frac{\partial \sum \log p_\theta(y_i)}{\partial \theta} = 0$. MLE is asymptotically normal, i.e.,

$$n^{1/2}(\hat{\theta}_n - \theta) \to N(0, \mathcal{I}(\theta)^{-1}) \text{ as } n \to \infty.$$

Notice that the asymptotic variance of MLE around true parameter value θ becomes smaller as Fisher information becomes larger.

In summary, Fisher information plays an important role in describing the asymptotic property of estimation for regular models, i.e., it determines a lower bound of risk of any arbitrary unbiased estimator. Due to the asymptotic normality of regular model, the inverse of Fisher information is equivalent to the asymptotic variance of the MLE. The regularity conditions C1) to C5) are therefore often described as the condition for the existence of Fisher information and for asymptotic normality. In order to pave the way to the introduction of Hellinger information in Chapter 2, the next subsection reviews a less restrictive condition that nonetheless leads to local asymptotic normality, namely *differentiable in quadratic mean.* An alternative derivation of Fisher information based on the Hellinger distance between distributions is also discussed. The purpose of section 1.1.2 is to provide some background to the definition of Hellinger information described in Chapter 2.

1.1.2 DQM and local asymptotic normality

Asymptotic normality also occurs under weaker conditions than those listed in Section 1.1.1. Specifically, asymptotic normality does not require C(4), and, in fact, $p_\theta(y)$ does not have to be differentiable with respect to θ for all y. For example, take the Laplace distribution's density function with mean parameter μ and, given variance $2b^2$, $p_\mu(y) = \frac{1}{2b} exp(-\frac{|y-\mu|}{b})$. The MLE for μ, $\hat{\mu}$ is the sample median, and $n^{1/2}(\hat{\mu} - \mu)/b$ is asymptotically normal. Distributions that satisfy C1) to C5), as well as distributions like Laplace distribution, satisfy a condition called *differentiable in quadratic mean* (DQM), which implies local asymptotic normality.

Definition 2. (Differentiable in quadratic mean) A model p_θ is differentiable in quadratic mean (DQM) at θ, for some $\theta' \in \Theta$, if there is a quadratic expansion of Hellinger distance between p_θ and $p_{\theta'}$, with some non-negative matrix $J(\theta)$. That is,

$$\int (\sqrt{p_\theta(y)} - \sqrt{p_{\theta'}(y)})^2 dy = (\theta - \theta')^\top J(\theta)(\theta - \theta') + o(|\theta - \theta'|^2) \text{ as } \theta \to \theta'. \tag{1.2}$$

For models satisfying C1) to C5), i.e., when Fisher information, $E_\theta(\dot{l}_\theta \dot{l}_\theta^\top)$, exists, then with regards to the above expression (Equation 1.2), $J(\theta) = \frac{1}{4} E_\theta(\dot{l}_\theta \dot{l}_\theta^\top)$. A brief justification follows. First, consider $\Theta \in R$. Assuming that Fisher information, $E_\theta(\dot{l}_\theta \dot{l}_\theta^\top)$, exists, then C1) to C4) imply that the point-wise derivative of $p_\theta(y)$ exists everywhere (with respect to θ), and so does the point-wise derivative of $\sqrt{p_\theta(y)}$, as follows:

$$\frac{\partial}{\partial \theta} \sqrt{p_\theta(y)} = \frac{1}{2} \frac{1}{\sqrt{p_\theta(y)}} \frac{\partial}{\partial \theta} p_\theta(y) = \frac{1}{2} \frac{\partial \log p_\theta(y)}{\partial \theta} \sqrt{p_\theta(y)}.$$

Therefore, the integral of the squared derivative of the square root of the probability distribution function is proportional to Fisher information:

$$\int \left(\frac{\partial}{\partial \theta} \sqrt{p_\theta(y)} \right)^2 \mathrm{d}y = \int \frac{1}{4} (\frac{\partial \log p_\theta(y)}{\partial \theta})^2 p_\theta(y) \mathrm{d}y = \frac{1}{4} E_\theta(i_\theta^2).$$

Then, by regularity condition C5),

$$\lim_{\varepsilon \to 0} \int (\frac{\sqrt{p_\theta(y)} - \sqrt{p_{\theta+\varepsilon}(y)}}{\varepsilon})^2 \mathrm{d}y = \int \left(\frac{\partial}{\partial \theta} \sqrt{p_\theta(y)} \right)^2 \mathrm{d}y = \frac{1}{4} E_\theta(i_\theta^2).$$

Thus, the squared Hellinger information can be approximated by the following:

$$h(\theta,\vartheta) = \frac{1}{4}\, E_\theta(\dot{l}_\theta^2)\,(\theta-\vartheta)^2 + o(|\theta-\vartheta|^2),\ \text{as } \theta \to \vartheta. \tag{1.3}$$

For the multi-dimensional case, $J(\theta) = \frac{1}{4}E_\theta(\dot{l}_\theta \dot{l}_\theta^\top)$, here is a brief justification for when $\Theta \in R^2$, with $\theta = (\theta_1, \theta_2)$. In this case, one can rewrite the squared Hellinger distance as follows:

$$\begin{aligned}
h(\theta,\vartheta) &= \int \left(\sqrt{p_{\vartheta_1,\vartheta_2}} - \sqrt{p_{\theta_1,\theta_2}}\right)^2 dy \\
&= \int \left(\sqrt{p_{\vartheta_1,\vartheta_2}} - \sqrt{p_{\theta_1,\vartheta_2}} + \sqrt{p_{\theta_1,\vartheta_2}} - \sqrt{p_{\theta_1,\theta_2}}\right)^2 dy \\
&= \int \left(\sqrt{p_{\vartheta_1,\vartheta_2}} - \sqrt{p_{\theta_1,\vartheta_2}}\right)^2 dy + \int \left(\sqrt{p_{\theta_1,\vartheta_2}} - \sqrt{p_{\theta_1,\theta_2}}\right)^2 dy \\
&\quad + 2\int \left(\sqrt{p_{\vartheta_1,\vartheta_2}} - \sqrt{p_{\theta_1,\vartheta_2}}\right)\left(\sqrt{p_{\theta_1,\vartheta_2}} - \sqrt{p_{\theta_1,\theta_2}}\right) dy.
\end{aligned}$$

Define the symmetric 2×2 matrix $J(\theta)$ as $J_{11}(\theta), J_{22}(\theta), J_{12}(\theta)$, and by similar calculation as the one-dimensional case, it can be shown that,

$$\begin{aligned}
J_{11}(\theta) &= \lim_{\varepsilon_1 \to 0} \frac{1}{\varepsilon_1^2} \int \left(\sqrt{p_{\theta_1+\varepsilon_1,\theta_2}} - \sqrt{p_{\theta_1,\theta_2}}\right)^2 dy = \frac{1}{4}E_\theta((\frac{\partial \log p_\theta}{\partial \theta_1})^2) \\
J_{22}(\theta) &= \lim_{\varepsilon_2 \to 0} \frac{1}{\varepsilon_2^2} \int \left(\sqrt{p_{\theta_1,\theta_2+\varepsilon_2}} - \sqrt{p_{\theta_1,\theta_2}}\right)^2 dy = \frac{1}{4}E_\theta((\frac{\partial \log p_\theta}{\partial \theta_2})^2) \\
J_{12}(\theta) &= \lim_{\varepsilon_1,\varepsilon_2 \to 0} \frac{1}{\varepsilon_1\varepsilon_2} \int \left(\sqrt{p_{\theta_1+\varepsilon_1,\theta_2+\varepsilon_2}} - \sqrt{p_{\theta_1,\theta_2+\varepsilon_2}}\right)\left(\sqrt{p_{\theta_1,\theta_2+\varepsilon_2}} - \sqrt{p_{\theta_1,\theta_2}}\right) dy \\
&= \frac{1}{4}E_\theta(\frac{\partial \log p_\theta}{\partial \theta_1}\frac{\partial \log p_\theta}{\partial \theta_2}).
\end{aligned}$$

Thus, the matrix $J(\theta)$ has following appearance:

$$J(\theta) = \begin{bmatrix} J_{11}(\theta) & J_{12}(\theta) \\ \\ J_{12}(\theta) & J_{22}(\theta) \end{bmatrix} = \frac{1}{4} E_\theta(\dot{l}_\theta \dot{l}_\theta^\top).$$

Since none of $J_{11}(\theta), J_{22}(\theta)$, or $J_{12}(\theta)$ depends on the direction of $(\varepsilon_1, \varepsilon_2)^\top$, $(u_1, u_2)^\top = \frac{(\varepsilon_1, \varepsilon_2)^\top}{|(\varepsilon_1, \varepsilon_2)|}$, and squared Hellinger distance between p_θ and p_ϑ has the following approximation, as $\theta \to \vartheta$:

$$\begin{aligned} h(\theta, \vartheta) =& (\theta - \vartheta)^\top \begin{bmatrix} J_{11} & J_{12} \\ \\ J_{12} & J_{22} \end{bmatrix} (\theta - \vartheta) + o(\|\theta - \vartheta\|_2^2) \\ =& \frac{1}{4} (\theta - \vartheta)^\top E_\theta(\dot{l}_\theta \dot{l}_\theta^\top) (\theta - \vartheta) + o(\|\theta - \vartheta\|_2^2). \end{aligned}$$

Notice that obtaining $E_\theta(\dot{l}_\theta \dot{l}_\theta^\top)$ does not require that $\dot{l}_\theta$ exist at every point of the support. If it is not differentiable at a countable number of points, this would still be true. Examples include the triangle distribution, and the Laplace distribution with location parameter.

The above shows that regularity conditions imply DQM; however, there are models that violate certain conditions of regularity but are nonetheless DQM, such as the triangle and Laplace distributions. A characteristic for these distributions is that P_θ is DQM at a specific value θ^*, such that the limit of $p_\theta(y)$ of $y \to \theta^*$ exists, but $p_\theta(y)$ is not necessarily differentiable with respect to θ at the point $y = \theta^*$. This means that for $p_\theta(y)$ to be DQM at some $\theta \in \Theta$, it is not required that C4) or C5) apply.

Example 1. With $\mathbf{1_S}(y)$ as the indicator function for set S, consider a non-symmetric standard triangular distribution with the following density (Shemyakin, 2014):

$$p_\theta(y) = \frac{2y}{\theta} 1_{[0,\theta]}(y) + \frac{2\,(1-y)}{\theta} 1_{[\theta,\leq 1]}(y).$$

$p_\theta(y)$ is differentiable in quadratic mean at θ, because

$$\lim_{\varepsilon \to 0} \int (\frac{\sqrt{p_\theta(y)} - \sqrt{p_{\theta+\varepsilon}(y)}}{\varepsilon})^2 \, \mathrm{d}y = \frac{1}{4\theta(1-\theta)}.$$

One important implication of DQM is that as the sample size becomes large, the local model behaves like normal distribution. A more detailed explanation follows. First, however, a consequence of DQM is that the limit of the likelihood ratio has a quadratic approximation as described in the following theorem.

Theorem 2. *(Van der Vaart, 1998) (Theorem 7.2) Suppose that Θ is an open subset of $\mathbb{R}^k$ and that the model $(P_\theta, \theta \in \Theta)$ is differentiable in quadratic mean at θ_0. Then, $P_{\theta_0} \dot{l}_{\theta_0} = 0$, and the Fisher information matrix $I_{\theta_0} = P_{\theta_0} \dot{l}_{\theta_0} \dot{l}_{\theta_0}^\top$ exists. Define a local parameter $h = \sqrt{n}(\theta - \theta_0)$ with θ_0 as a fixed point, for every converging sequence $h_n \to h$, as $n \to \infty$.*

$$\log \prod_{i=1}^{n} \frac{p_{\theta_0 + \frac{h_n}{\sqrt{n}}}}{p_{\theta_0}}(Y_i) = h^\top \frac{\sum_{i=1}^{n} \dot{l}_{\theta_0}}{\sqrt{n}} - \frac{1}{2} h^T \mathcal{I}_n(\theta_0) h + o_{p_{\theta_0}}(1) \;\; \textit{for every } h_n \to h, \;\; \textit{as } n \to \infty. \quad (1.4)$$

The limit of likelihood ratio above has a similar form to the log likelihood ratio of $N(h, I(\theta_0))$ and $N(0, I(\theta_0))$:

$$\log \frac{dN(h, I(\theta_0))}{dN(0, I(\theta_0))}(Y) = h^T I(\theta_0) Y - \frac{1}{2} h^T I(\theta_0) h. \tag{1.5}$$

To see this, first recall that the expectation of score function, $\dot{l}_{\theta_0}$, is zero, and that its variance is the Fisher information. By the central limit theorem, $\frac{\sum_{i=1}^{n} \dot{l}_{\theta_0}}{\sqrt{n}}$, from the first term of the right hand side of (Equation 1.4), the function follows distribution $N(0, \mathcal{I}_n(\theta_0))$:

$$\frac{\sum_{i=1}^{n} \dot{l}_{\theta_0}}{\sqrt{n}} \xrightarrow{p_{\theta_0}} N(0, \mathcal{I}_n(\theta_0)). \tag{1.6}$$

From the first term of the right hand side of (Equation 1.5), since $Y \sim N(0, I^{-1}(\theta_0))$,

$$\mathcal{I}(\theta_0) Y \sim N(0, \mathcal{I}(\theta_0)).$$

In other words, $h^\top \frac{\sum_{i=1}^{n} \dot{l}_{\theta_0}}{\sqrt{n}} \xrightarrow{p_{\theta_0}} h^\top \mathcal{I}_n(\theta_0) Y$. Hence,

$$\log \prod_{i=1}^{n} \frac{p_{\theta_0 + h_n/\sqrt{n}}}{p_{\theta_0}}(Y_i) \xrightarrow{p_{\theta_0}} \log \frac{dN(h, \mathcal{I}_n(\theta_0))}{dN(0, \mathcal{I}_n(\theta_0))}(Y). \tag{1.7}$$

Given that θ_0 is the true parameter value, the limit of likelihood ratio reflects how much information there is in the model from a sample of n observations about θ, because, if given a sample of size n and if the ratio is large, it follows that the ratio is sensitive to mis-guessing the true value of the parameter. The more sensitive the likelihood ratio is to mis-guessing the true value of the parameter, the more informative to the true parameter value.

In a sense, the limit of likelihood ratio, $\log \prod_{i=1}^{n} \frac{p_{\theta_0+\frac{h_n}{\sqrt{n}}}}{p_{\theta_0}}(Y_i)$ in (Equation 1.4), describes, under DQM, how much information there is for θ_0 when given a sample of size n. Based on (Equation 1.6), the expectation of the first term in the right hand side of (Equation 1.4) is zero. Thus, the "relative peakness" of likelihood ratio at θ_0 is measured by $\frac{1}{2}h^T\mathcal{I}_n(\theta_0)h$. Put differently, under DQM, it is the Fisher information $\mathcal{I}_n(\theta_0)$ that determines the amount of information for a sample size of n on θ_0.

That the limit of likelihood ratio converges to a normal one indicates that, as sample size increases, the local model has similar properties to a normal model. (Equation 1.7) means that when sample size n is large, the local model $(p^n_{\theta_0+h/\sqrt{n}} : h \in R^k)$ and $(N(h, \mathcal{I}_n(\theta_0)^{-1}) : h \in R^k)$ have similar statistical properties. $(N(h, \mathcal{I}_n(\theta_0)^{-1}) : h \in R^k)$ can be considered a limit model of $(p^n_{\theta_0+h/\sqrt{n}} : h \in R^k)$ that is DQM. Discussion of limit local model is helpful because every sequence of statistics in $(p^n_{\theta_0+h/\sqrt{n}} : h \in R^k)$ is matched in the limit by a statistics in its limit model, with the implication that there is no sequence of estimator that can be asymptotically better, in terms of risk, than the best estimator in the limit model. Typically, MLE is matched by MLE in the limit experiment. A rigorous statement of this conclusion is expressed in the Asymptotic Representation Theorem introduced in (Le Cam et al. 1972) and (Van der Vaart, 1991).

For the normal distribution, the best estimator in terms of efficiency for the location parameter is MLE. Accordingly, for models that are DQM, whose limit experiment is Normal, the best estimator shall be MLE $\hat{\theta}$, as well, and it follows normal distribution asymptotically, i.e. $\sqrt{n}(\hat{\theta} - \theta_0) \sim N(0, \mathcal{I}_n^{-1}(\theta_0))$. In other words, for the model that is differentiable in quadratic

mean at θ, the inverse of Fisher information is the asymptotic variance-covariance matrix. For these reasons, for the rest of this thesis, we say that a model is regular as long as it is DQM. Nonregular models refer to those that are not DQM.

1.1.3 Regular regression model and Fisher information

Statistical models help scientists understand how different conditions and features of their observations affect observed outcomes. Linear regression, non-linear regression, and generalized linear models are common statistical modeling tools. For these models, Fisher information depends on covariates. As Fisher information determines the quality of estimation, as discussed previously, in the stage of experimental design, one can choose covariates for observations in an experiment in ways that can improve the quality of estimation.

Linear and nonlinear models with normal distributed error terms can be characterized as follows: with $\mathbf{x}_i$ being the covariate vector for ith observation and θ being the parameter of interest, $\theta \in R^k$, there is a differentiable function $g(\theta, X)$ such that

$$y_i = g(\theta, \mathbf{x}_i) + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2),\ i = 1, ..., n.$$

Denoting $\mathcal{I}_{\mathbf{x}_i}(\theta)$ as the Fisher information based on ith observation, then

$$\mathcal{I}_{\mathbf{x}_i}(\theta) = \sigma^{-2}(\frac{\partial g(\theta, \mathbf{x}_i)}{\theta_1}, ..., \frac{\partial g(\theta, \mathbf{x}_i)}{\theta_d})^\top(\frac{\partial g(\theta, \mathbf{x}_i)}{\partial \theta_1}, ..., \frac{\partial g(\theta, \mathbf{x}_i)}{\partial \theta_d}).$$

By the additivity property, Fisher information of θ based on a sample of n independent observations $(y_i, \mathbf{x}_i), i = 1, ..., n$, $\mathcal{I}_n(\theta)$, would be the summation of $\mathcal{I}_{\mathbf{x}_i}(\theta), i = 1, ..., n$:

$$\mathcal{I}_n(\theta) = \sigma^{-2} \sum_{i=1}^{n} (\frac{\partial g(\theta, \mathbf{x}_i)}{\theta_1}, ..., \frac{\partial g(\theta, \mathbf{x}_i)}{\theta_d})^\top (\frac{\partial g(\theta, \mathbf{x}_i)}{\partial \theta_1}, ..., \frac{\partial g(\theta, \mathbf{x}_i)}{\partial \theta_d}).$$

Based on the above summary of the Fisher information for linear and nonlinear models, one sees that Fisher information is a function of the covariates $\mathbf{x}_i, i = 1, ..., n$ and θ. The familiar example of linear regression illustrates this point.

Example 2. When the regression model is linear, then $g(\theta, \mathbf{x}_i) = \mathbf{x}_i^\top \theta$. Letting Y be the vector of n observation, $Y = (Y_1, ..., Y_n)^\top$, and letting $X = (\mathbf{x}_i, ..., \mathbf{x_n})^\top$ be the design matrix, and $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^\top$ be the vector of error, then the model can be written as

$$Y = X\theta + \varepsilon, \ \varepsilon \sim N_n(0, \sigma^2 I).$$

The Fisher information based on (Y, X) would therefore be $\mathcal{I}_n(\theta) = \sigma^{-2} X^T X$. The best linear unbiased estimator for θ is $\hat{\theta} = (X^T X)^{-1} X^T Y$ and the variance of $\hat{\theta}$ is $Var(\hat{\theta}) = \sigma^2 (X^T X)^{-1} = \mathcal{I}_n(\theta)^{-1}$.

The generalized linear model is commonly used in statistics. The model set-up assumes that observation y is generated from some distribution of the exponential family, with the mean $E(Y)$ determined by independent variable X, which the parameter of interest is determined through a link function $g(.)$, $E(Y) = g^{-1}(X\theta)$. Any generalized linear models belong to the exponential family, where regularity conditions are satisfied, so the Cramér-Rao lower bound applies, and

the MLE is asymptotically normal, with the variance-covariance matrix being equivalent to the inverse of Fisher information. What follows is an example of using a generalized linear model for count data.

Example 3. In a Poisson regression model for independent count data $Y_i, i = 1, \ldots, n$ with covariate vector $\mathbf{x}_i$, $i = 1, \ldots, n$ and link function $\log(\lambda_i) = \mathbf{x}_i^\top \theta$, based on the additivity property and the reparametization rule of Fisher information matrix, the Fisher information matrix for θ can be written as $\mathcal{I}_n(\theta) = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top exp(\mathbf{x}_i^\top \theta)$.

One can see that for generalized linear models, Fisher information also depends on covariates, just as in the case of regression models. Therefore, for linear, nonlinear regression, and generalized linear models, the Fisher information matrix is determined by covariates $\mathbf{x}_i, i = 1, ..., n$.

1.2 Optimal design of experiment

Planning an experiment involves decisions about how many observations of the various experimental conditions should be conducted. Taking into account the purpose of the experiment, the mathematics of optimal design seeks to determine the best design based on a number of goals, such as minimizing the chances of a wrong conclusion or maximizing precision in the estimation of the parameters in the specified statistical model.

As described in the previous section, for both linear and nonlinear regression models, and generalized linear models, Fisher information depends on the the covariates of each observation. Since Fisher information inversely determines the lower bound of mean square error of arbitrary unbiased estimators, maximizing $\mathcal{I}(\theta)$ in some sense among possible experimental designs will optimize design in terms of producing the most efficient estimates.

1.2.1 Approximate design

In the study of experimental design, an exact design of size n includes information about where to take observations from, i.e., location of observation sites, how many different sites there will be, and how many observation are taken from each site. Denote χ as the covariate design space, which is a collection of all sites at which one can make observations. For example, consider the simple linear regression model, $y_i = \theta_0 + \theta_1 x_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma), i = 1, ..., 8$, where the design space is $\chi = [-5, -1]$. An exact design of this experiment contains 4 sites. Site 1 is located at $x_1 = -1$ with 1 observation, site 2 is taken at $x_2 = -2$ with 3 observations, site 3 is located at $x_3 = -4$ with 2 observations, and site 4 is located at $x_5 = 5$ with 2 observations.

(Kiefer, 1974) suggested the use of approximate design to simplify the setup of experimental design. An approximate design has the format of $\xi = \{(w_i, x_i), i = 1, ..., r\}$, $x_i \in \chi$, where $x_i, i = 1, ..., r$ represents r distinct locations of experimental space from which one takes observations, also known as the "design points" of an experiment. In this format, w_i represents the proportion of total observations from site x_i. In an approximate design, all the weights sum up to 1, i.e., $\sum_i^r w_i = 1$, thus ξ over χ is also called a design measure. The approximation of the exact design described previously can then be written as $\xi = \{(\frac{1}{8}, -1), (\frac{3}{8}, -2), (\frac{2}{8}, -4), (\frac{2}{8}, -5)\}$.

Considering optimality criteria described in the following section, optimization with respect to exact design is generally intractable, due to the fact that it is a discrete optimization problem. Optimization with respect to approximate design works around this issue by rendering the optimization problem such that it is continuous. In solving real problems, after obtaining an optimal design with respect to approximate design, one can convert the approximate design to

exact design of n observations by taking $n \times w_i$ observations (or the closest integer to $n \times w_i$ if it is not an integer) at location x_i.

Consider an approximate design $\xi = \{(w_i, x_i), i = 1, ..., r\}$. By additivity of Fisher information, with sample size of n, Fisher information based on design ξ with sample size n can be expressed as

$$I_\xi(\theta) = n \sum_{i=1}^{r} I_{\mathbf{x}_i}(\theta) w_i.$$

Since sample size n is irrelevant in finding approximate optimal design, the matrix

$$M_\xi(\theta) = \sum_{i=1}^{r} I_{\mathbf{x}_i}(\theta) w_i \tag{1.8}$$

is called the information matrix of a design, and finding an approximate optimal design boils down to maximizing some function of $M_\xi(\theta)$, where the choice of function depends on the optimality criterion.

For example, given an approximate design with r many design points, $\xi = \{(w_1, \mathbf{x}_i)...(w_r, \mathbf{x_r})\}$, for Poisson regression from Example 3, the information matrix of ξ would be

$$M_\xi(\theta) = \sum_{i=1}^{r} w_i \mathbf{x}_i \mathbf{x}_i^\top exp(\mathbf{x}_i^\top \theta).$$

For generalized linear models, the information matrices and thus the corresponding designs depend on the unknown model parameters. One way to deal with this is to identify locally optimal designs based on the best guess of the parameters. While the local guessed value

can be provided by an expert, one can also conduct a first stage design with a small number of observations to obtain a reasonable initial estimation, then use it as the guessed value. According to (Ford et al., 1992), a local optimal design can also serve as a benchmark to check the efficiency of other designs.

1.2.2 Goal of optimal design of experiment for regular models

Under regularity conditions, the Cramér-Rao lower bound theorem states that Fisher information is the inverse of the direct or asymptotic variance-covariance matrix of unbiased efficient estimator, such as MLE for linear regression, nonlinear regression, and generalized linear models. When working with these models, the goal of optimal design is to minimize the variance-covariance matrix of efficient estimator by maximizing the Fisher information matrix. To directly compare non-negative definite matrices, one can use Loewner ordering.

Definition 3 (Loewner Order)**.** If a matrix M with dimension $(k \times k)$ is higher in Loewner ordering than matrix $\tilde{M}$, such that $M \geq \tilde{M}$, this means that $M - \tilde{M}$ is a non-negative definite matrix, i.e., the smallest eigenvalue has to be non-negative.

A sufficient condition for $M, \tilde{M}$ such that $M - \tilde{M} \geq 0$ is that $M_{i,j} = \tilde{M}_{i,j:}$, for all $1 \leq i, j \leq k$, except for one i, such that $M_{i',i'} \geq \tilde{M}_{i',i'}$. However, there is no total ordering on the non-negative definite matrices in Loewner Order. Therefore, in practice, a convex scalar function of Fisher information (such as the trace, minimum, and determinate of a matrix) are typically used as criteria for comparison. The following summarizes some commonly used optimality criteria and their statistical meanings (Stufken and Yang, 2012):

- D-optimality. A design is D-optimal for θ if it maximizes the determinant of Fisher information over all possible designs. D-optimal designs minimize the expected volume of the asymptotic $100(1-\alpha)\%$ joint confidence ellipsoid for the elements of θ.
- A-optimality. A design is A-optimal for θ if it minimizes the trace of the inverse of Fisher information over all possible designs. A-optimal designs minimize the sum of the asymptotic variances of the estimators of elements of θ.
- E-optimality. A design is E-optimal for θ if it maximizes the smallest eigenvalue of Fisher information. E-optimal designs minimize the expected length of the longest semi-axis of the asymptotic $100(1-\alpha)\%$ joint confidence ellipsoid for the elements of θ.

(Dette et al., 2011) pointed out that most of the optimality criteria are monotonic with respect to Loewner ordering. Suppose that $\Phi(.)$ is an optimality criterion function. This function relates to non-negative definite matrices $M, \tilde{M}$, like so:

$$M \geq \tilde{M} \Rightarrow \Phi(M) \geq \Phi(\tilde{M}).$$

Furthermore, based on (Yang and Stufken, 2009), if one is interested in estimation of a differentiable function of parameter $g(\theta)$, and if $\hat{\theta}$ is unbiased and an efficient estimator for θ, the asymptotic covariance matrix of $\eta(\hat{\theta})$ under design ξ becomes $Cov_{\xi}(\eta(\hat{\theta})) = (\frac{\partial \eta(\theta)}{\partial \theta^{\top}}) I_{\xi}^{-1}(\theta)(\frac{\partial \eta(\theta)}{\partial \theta^{\top}})^{\top}$. Further, this implies that, for two designs ξ, $\tilde{\xi}$, such that $I_{\xi}(\theta) \geq I_{\tilde{\xi}}(\theta)$, then $Cov_{\xi}(\eta(\hat{\theta})) \leq Cov_{\tilde{\xi}}(\eta(\hat{\theta}))$.

1.2.3 Optimization of information matrix

This section will review some recent developments in the strategy of finding the optimal design of experiment. Taking advantage of the monotone properties of D-, E-, and A-optimality criteria with respect to Loewner ordering, one can first find a small class of designs with simple features such that, for any design outside of this class, there is an equal or better design in the class in terms of Loewner ordering. Thereby, the search for optimal design with respect to a specific criterion can be restricted to this small class of designs, called the "complete class."

Definition 4. (Complete Class) A complete class Ξ is a subclass with a simple format such that for any design $\xi \notin \Xi$, there exists a design, $\xi^* \in \Xi$, that satisfies $M_{\xi^*}(\theta) \geq M_{\xi}(\theta)$.

The process of finding the complete class and then searching for optimal design based on a specific optimality criterion greatly simplifies the optimization problem, as the complete class has a lower number of design points to begin with.

In an early attempt to find the complete class for optimal design, the Caratheodory theorem gives the upper bound for the number of design points that a design can have for any $k \times k$ moment matrix that can be written as a linear combination of at most $\frac{k(k+1)}{2}+1$ many moment matrices of design with only one design point.

(De la Garza, 1954) presented the result that for a k$-$th order polynomial regression model, a complete class for the model consists of designs with precisely $k+1$ points, i.e., for any design with a number of support points larger than $k+1$, there is a design with $k+1$ points that is equal or better. Furthermore, $k+1$ is also the minimum number of support points such that the model is estimable. In other words, the minimum number of support points in optimal

design for such a model would be the same as the dimension of the parameter in the polynomial regression model.

(Yang and Stufken, 2009) has found that the de la Garza phenomenon exists for many other nonlinear models with two parameters, such as logistic and probit models, for which the optimal design would be a two-point design. (Yang, 2010) extends the result from (Yang and Stufken, 2009) to many commonly used nonlinear models, with no limitation on the dimension of the parameter. Furthermore, the procedure proposed in (Yang, 2010) is easy to implement. For example, there would be minimally k many design points in optimal design for the Poisson regression model from Example 2, with $\theta \in R^k$. (Dette et al., 2011) and (Yang et al., 2012) further extend these results to a larger class of nonlinear models. For example, given the double-exponential regrowth model that is used to describe the dynamics of post-irradiated tumors, $y_i = \theta_1 + \log(\theta_2 e^{\theta_3 x_i} + (1-\theta_2)e^{-\theta_4 x_i}) + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$, the designs with at most four points form a complete class. In Chapter 2, we will see that Hellinger information is not in a matrix form, thus some of the techniques of optimization of the non-negative definite matrix developed in the literature on the optimal design of experiments are not applicable; however, the idea of using the complete class to find optimal designs is nonetheless employed. In Chapter 3, optimal design for a simple linear and quadratic nonregular regression model will be presented, and the number of design points in the optimal design for these two models is equal to the number of parameters, coinciding with the de la Garza phenomenon.

1.3 Introduction to nonregular models

Nonregular models include any models that do not satisfy the regularity conditions and, thus, for which Fisher information does not exist. In Chapter 2, we will develop an approach using Hellinger information as the measure of information for a large class of models, including regular models and some nonregular ones. The later chapters of this thesis focus on a class of nonregular models applicable in different areas of scientific studies. This section introduces these nonregular models, including some of their applications, and the available methods for estimations, along with explanations of the challenges that these nonregular models present to optimization of design. In Chapter 3, some optimal design results for nonregular models in this section will be presented, based on the optimization of Hellinger information described in Chapter 2.

1.3.1 Distribution with parameter-dependent support

The type of nonregularity considered in this project can be described as models with parameter-dependent support, a violation of condition C(3). A simple example of this type of model is uniform distribution, where the boundary of the support of the distribution is determined by the parameter itself. Letting $\mathbf{1}_{[\mathbf{0},\theta)]}(.)$ be the indicator function,

$$\text{Uniform distribution: } p_\theta(y) = \frac{1}{\theta}\,\mathbf{1}_{[\mathbf{0},\theta)]}(y).$$

The scope of nonregular models considered in this project can be described in terms of the following conditions. First, $p_\theta(y)$ is strictly positive on support $S(\theta) = [a_1(\theta), a_2(\theta)]$, and

bounded and continuous on any compact set in Θ. Second, $p_\theta(y)$ is absolutely continuous in θ for a fixed y in the interior of support, i.e., where $y \in (a_1(\theta), a_2(\theta))$, where both $a_1(.)$ and $a_2(.)$ are zero functions, and derivative $p'_\theta(y) = \frac{\partial p_\theta(y)}{\partial \theta}$ exists for each θ in the interior of $S(\theta)$. Third, right limit $q_1(\theta) = \lim_{y \searrow a_1(\theta)} p_\theta(y)$ and left limit $q_2(\theta) = \lim_{y \nearrow a_2(\theta)} p_\theta(y)$ are finite.

One such model is a truncated distribution, typically used in situations where the range of random variable is bounded from below or above, such that observations beyond the bound are ignored. A truncated distribution can be formed by forcing a bound on the support of the distribution and normalizing it such that it would still be integrated to one over the new support. It can be viewed as a conditional distribution that results from restricting the domain of some other probability distribution. Specifically, let $p(y)$ be the original probability distribution and let $F(y)$ be the cumulative distribution. To form a truncated distribution with parameter-dependent support, such that $y \in [\theta_1, \theta_2]$, with $\theta = (\theta_1, \theta_2) \in R^2$, one would define the truncated distribution as

$$p_\theta(y) = \frac{p(y)}{F(\theta_2) - F(\theta_1)}.$$

There are many applications for truncated distribution. For instance, in reliability engineering, a product might be expected to exhibit a very low failure rate during the period of warranty, but fail quickly after its designated service life. (Zhang and Xie, 2011) suggested that for the purpose of avoiding over-engineering, it would be useful to employ an upper-truncated distribution to model failure rate of a product. They presented an example utilizing the upper-truncated Weibull model to analyze a set of real test data representing time-to-failure of the

turbocharger of one type of engine. Time to failure $\mathbf{t}$ follows upper-truncated Weibull distribution, with θ as the parameter for the upper truncation point,

$$p_\theta(t) = \frac{\frac{\beta}{\lambda}(\frac{t}{\lambda})^{\beta-1}exp(-(\frac{t}{\lambda})^\beta)}{(1-exp(\frac{\theta}{\beta}))}\mathbf{1}_{(0,\theta]}(t).$$

(Finney and Varley, 1955) use truncated Poisson distribution to model the number of eggs from gall-flies and the number of Gall-cells in flower heads with data that is incomplete. (Zaninetti, 2014) used a truncated gamma distribution to model samples of stars. (DePriest, 1983) used truncated normal distribution for estimations of radiance measurements from satellite-borne infrared sensors. Several further examples of application of truncated distribution are described in (Fu, 2016).

Another commonly used type of nonregular distribution with parameter-dependent support involves shift discontinuities, which can be useful to model minimum values. Shift discontinuity in this context can be conceived of as a change of variable to impose a shift on the support of the distribution. For example, start with variable z with support $[0, \infty)$; given a location of shift, θ, one can form a new variable y such that $y - \theta = z$. The support of the new shifted distribution would depend on θ. The shifted gamma distribution is an example of this kind of irregularity:

$$p_\theta(y) = \frac{1}{\Gamma(\beta)}(y-\theta)^{\beta-1}\exp[-(y-\theta)], y \geq \theta \text{ given } \beta > 0. \tag{1.9}$$

Shifted distribution is used when one wants to include a lower threshold for the observations from models with non-negative observations. Unlike in truncated distribution described

above, where the observations out of the interested range are ignored, the assumption for shifted distribution is that there will not be observations considered below the lower threshold. Many disciplines employ shifted distribution for a variety of purposes, with applications in fields as diverse as physics, material studies, climatology, and psychology, among others (see (Cousineau, 2009)). In studies of the voltage endurance of devices, a common method is to apply the voltage continuously to a test specimen until breakdown occurs, at which point the breakdown voltage is recorded and studied. Usually, there is a threshold voltage below which no breakdown can occur, and it is of interest to determine the value of this threshold in these voltage endurance tests. (Hirose and Lai, 1997) used the shifted-Weibull distribution with density function $p_\tau(y) = \frac{\beta}{\lambda}(\frac{y-\tau}{\lambda})^{\beta-1}exp(-(\frac{y-\tau}{\lambda})^{\beta}), y \geq \tau$, to model breakdown voltage data. The random variable Y represents the breakdown voltage, with threshold parameter τ representing the voltage below which no breakdown occurs. (Allen, 2014) employed a shifted exponential distribution to model computer response time at a computer center, with the shift parameter representing the minimum response time. (Bartolucci et al., 1999) utilized a shifted-Weibull to model time to failure of an adjuvant breast cancer therapy trial. (Bartkut-Norkūnien and Sakalauskas, 2009) deployed shifted Weibull distribution in modeling hourly wind speed for a wide range of locations in Europe. For further examples applying shifted distribution, see (Cousineau, 2009).

1.3.2 Nonregular regression

A nonregular regression model based on shifted distribution allows one to model relationships between features of observations and minimum values. (Smith, 1994) gives the general

set-up for nonregular regression. Let $\mathbf{x}_i = (x_{i1}, ..., x_{id})^T$ be the vector of value of experimental variable of ith observation, and let $g(\theta; \mathbf{x}_i)$ be a differentiable function of $\theta \in \mathbb{R}^d$ in all dimensions, given $\mathbf{x}_i$.

$$y_i = g(\theta; \mathbf{x}_i) + \varepsilon_i, i = 1, ..., n, \theta \in \mathbb{R}^d, \tag{1.10}$$

Here, $\varepsilon_i, i = 1, ..., n$ independently follows a non-negative distribution like gamma distribution, Weibull distribution, Pareto distribution, etc. What follows are some examples in which nonregular regression models apply.

Example 4. (Smith, 1994) has shown that to model the annual minimum temperatures in Gothenburg, Nebraska from 1895−1987, a linear regression model with nonzero residual following Weibull distribution can be used. Let Y_i be the temperature of the ith year, and let $x_i = i - 1941$, with i being the calender year. With that set up, (Smith, 1994) used the model $y_i = \theta_0 + \theta_1 x_i + \varepsilon_i$,with ε_i, an example of a two-parameter Weibull distribution. Another example used by (Smith, 1994) involves a quadratic nonregular regression model to analyze the annual best performance in the mile race from 1931 to 1985.

Example 5. In the study of auctions, (Donald and Paarsch, 2002) and (Chernozhukov and Hong, 2004) show that a nonregular regression model can be used to model winning bids in auctions when information such as number of potential bidders, information available to bidders, and the attitude of bidders toward risk are given and encoded as covariate values. Assuming the bidders' valuations are independent, letting x_i be the covariates for the ith auction, then the winning bid, Y_i can be explained by an efficient cost function $c(X_i)$ that depends on X_i, the

feature of ith auction, and a mark-up function $\beta(n_i) \geq 1$, with n_i as the number of bidders (assuming that $\beta(n_i)$ approaches 1 as number of bidders n_i approaches infinity). In this context, a common auction model can then be written as $Y_i = c(X_i)\beta(n_i) + \varepsilon_i, \varepsilon_i \geq 0$, ε_i according to distributions like Exponential, Pareto, Weibull, etc.

Example 6. (Chernozhukov and Du, 2001) suggest using nonregular regression to model observed capital stock, when it can be assumed that observed capital stock satisfies the equation $Z_i = s(X_i) + v_i$, where X_i are covariates and v_i is a disturbance that is positive most of the time.

Example 7. In a study on the impact of various socioeconomic demographics pertaining to minimum birth-weight, (Chernozhukov and Du, 2001) use a linear nonregular model with covariates including gender of child, age of mother, cigarette consumption of mother, education level of mother, etc.

Example 8. (Hall et al., 2009) presented an example using a nonregular regression model in the study of utility companies in the United States. The dependent observation $y_i = -\log(\frac{c_i}{p_i})$ is the negative log transformation of ratio of the cost c_i and the price of fuel for ith company p_i. The independent variable $X_i = \log(Q_i)$ is the log output of ith company. They find that a linear model is appropriate for these data; that is, $y_i = \theta_0 + \theta_1 x_i + \varepsilon_i$, with ε_i, condition on x_i, following a distribution that satisfies $p(u; x) = b(x)c(x)u^{c(x)-1} + o(u^{c(x)+d-1})$ as $u \to 0$, and $d > 0$, b(.) and c(.) are smooth, strictly positive scalar functions.

In this paper, as a first attempt at the problem of finding optimal designs for nonregular models, we begin with a relatively simple set-up where the error term is iid for all observa-

tions. In this thesis, we focus on the nonregular regression model described in (Smith, 1994), where the error term in nonregular regression, $\varepsilon_i, i = 1, .., n$, identically and independently follows non-negative distributions like gamma, Weibull, some extreme value distribution and Pareto distributions, etc., with a probability distribution function that can be described by the following:

$$p(z) = \alpha\, c z^{\alpha-1} \text{ for some } \alpha > 0, c > 0 \text{ as } z \to 0. \tag{1.11}$$

So far, in practice, most nonregular regression models are used for observation studies. It seems that it is hard to find examples in the literature of nonregular regression being applied in experimental studies. However, for many examples described in Section 1.3.1 and the economic auction models, one might conduct experiments to discover whether different experimental conditions would lead to different minimum values. For instance, in the voltage endurance studies, if the researcher is interested in the relation between certain features of the device and minimum breakdown voltage, then a nonregular regression model could be used and experiments testing this relationship could be conducted.

Experimental auctions compose an area of interest in economics to test certain propositions of auction theory, and, in practice, experiments are conducted to study buyers' willingness to pay for certain products. For details, see, (Umberger and Feuz, 2004) and (Kagel and Levin, 2009). Although nothing, it seems, has been written on the experimental design for nonregular regression auction models described earlier in (Donald and Paarsch, 2002), it may be an area of interest for researchers to conduct experiments testing how different features of the bidder or of the item on auction can change the winning bid.

1.3.3 Parameter estimation in nonregular problems

Under regularity conditions, as discussed in Section 1.1.1, the maximum likelihood estimator is usually considered the best estimator, because it is efficient with respect to the Cramér-Rao bound and it exhibits asymptotic normality, which tends to provide grounds for clear interpretation for estimation and statistical inference.

The models described in Section 1.3.1 and 1.3.2 are nonregular due to their common feature of parameter-dependent support in violation of regularity condition C(3) and C(4). It is true that if the shape parameter of the shifted-gamma or shifted Weibull distribution equals two, then the model is DQM on the shift location parameter, as will be discussed in Chapter 2, and the MLE would exhibit asymptotic normality, despite violation of conditions C(3) and C(4). However, in most cases, the nonregular models described here are not DQM, therefore the MLE can no longer serve as the default choice. In these situations, the behavior of the MLE needs to be examined on a case-by-case basis for different nonregular models.

(Smith, 1994) summarized the maximum likelihood estimation for nonregular distributions described in Section 1.3.1 with the characteristic of

$$p_\theta(y) = \alpha\, c(y-\theta)^{\alpha-1} \text{ for some } \alpha > 0, c > 0 \text{ as } y \to \theta, y \geq \theta. \tag{1.12}$$

When $a > 2$, Fisher Information exists, usually with the asymptotic result about MLE being valid. When $1 < a \leq 2$, local MLE exists, but does not have the usual asymptotic properties, i.e., it is not asymptotically normal, and asymptotic efficiency is an open question. When

$0 < a \leq 1$, no local MLE exists but there are a number of alternative estimators consistent at rate $O(n)$. There are some asymptotically efficient estimators in these cases, however, they depend on the choice of loss function, and the construction of the estimators is complicated, as described in greater detail in (Ibragimov and Hasminskii, 1981).

In addition to the problems of the maximum likelihood estimator described above, (Cousineau, 2009) and (Hirano and Porter, 2003) point out that, for some shifted models, the MLE solution is biased by an unknown amount. (Cousineau et al., 2004) have shown that for shifted Weibull distribution, the bias depends on the shape parameter and sample size.

To address the estimation problem of one-dimensional shift or truncated parameters, there are many proposed estimators other than MLE. (Hirose, 1999) provided a bias-corrected estimator, but it is sensitive to the specific details of the implementation. There is also the maximum product of spacing estimator proposed in (Cheng and Amin, 1979) and (Ranneby, 1984), which avoids inconsistent solutions to the likelihood equation when the shape parameter of shifted Weibull distribution is smaller than one. A generalization of the maximum product of spacing method is the quantile maximum product estimation, discussed in (Heathcote et al., 2002), (Heathcote et al., 2004), (Cousineau et al., 2004), (Speckman and Rouder, 2004), and (Heathcote and Brown, 2004). (Jacquelin, 1993) proposed an alternative to MLE called weighted maximum likelihood estimation, with the purpose of canceling out the bias. (Smith, 1985) suggested the sample minimum for estimation with respect to the shifted models.

For the estimation of nonregular regression models described in Section 1.3.1, (Smith, 1994) proposed an estimator that is based on minimum order statistics. The asymptotic variance-

covariance structure for the estimator provided in (Smith, 1994) is not available. However, the confidence interval for inference can be obtained though simulation. More details about Smith's estimator for nonregular regression will be given in Chapter 3. (Smith, 1994) argues that this method held advantage over MLE due to the fact that, in reality, the value of shape parameter α from the distribution of ε_i (Equation 1.11) is unknown, and MLE behaves differently for different α values, whereas Smith's method works for all α values. Smith's method appears to be the most straightforward way for estimation of nonregular regression; however, it seems that it only works for nonregular linear regression models.

Other estimators for nonregular regression include the Bayesian method and the nonparametric method. (Hirano and Porter, 2003) propose to use a Bayesian estimator for nonregular regression problems, with the error term following exponential distribution, while (Chernozhukov and Hong, 2004) extend the Bayesian approach to a wider class of nonregular regression models. (Hall et al., 2009) use a nonparametric method for nonregular regression models, demonstrating their method by applying it to Example 7.

In summary, maximum likelihood behaves differently depending on the value of other parameters in the model, and these values are quite likely to be unknown in real practical situations. There are other estimators that can be deployed to address the problem presented by nonregularity depending on various attributes, such as different values of the shape parameter. Most of these estimators are characterized by an asymptotic variance-covariance structure that is not analytically available. This implies that it is not feasible to take the same approach for optimal design as in the regular case, i.e., finding the expression of the variance-covariance matrix

then optimizing it with respect to designs. The motivation for this thesis is to find a unified approach, adopting a measure of information that can be optimized in order to determine optimal design, without needing to derive the variance-covariance matrix of specific estimators for specific models.

1.4 Information for nonregular models

Under regularity conditions, Fisher information exists and is the inverse of the asymptotic variance-covariance matrix of MLE. Thus, one can obtain the design with the minimal variance-covariance matrix of MLE among all possible designs by choosing the design that maximizes Fisher information. When there is not asymptotic normality available to the nonregular models, as discussed in Section 1.3, the estimation problem becomes complicated. Most important, the absence of the variance-covariance matrix of estimators and risk bound, such as the Cramér-Rao lower bound, makes it difficult to approach the problem of optimal design of experiment for nonregular models. While it is possible to consider the problem of optimal design for nonregular models on the basis of specific models with specific estimators by closely examining the asymptotic distribution of the specific estimator, a new generalized approach to optimal design for nonregular models is desired.

Any alternative approach to the measure of information in optimal design must meet two criteria. First, the new measure of information has to be a more general form of measuring information, such that Fisher information would be a special case of the new information. Second, it must be related to the quality of estimation such that the optimization of this measure of information with respect to designs can lead to better estimation. In Chapter 2, a

Cramér-Rao-like lower bound is introduced and Hellinger information is defined. This thesis proposes Hellinger information as the information measure to be used in a unified approach to optimal design problems involving nonregular models. In Chapter 3, optimal designs for two-parameter linear and quadratic nonregular regression models based on Hellinger information are derived, and results from simulation studies on the optimal designs are included.

CHAPTER 2

INFORMATION IN NONREGULAR MODELS

This project seeks to define a measure of information that could serve a function similar to Fisher information, but which can be applied to nonregular models, for which Fisher information does not exist. The resulting measure of information, to be applied usefully in the optimal design of experiment, must possess two properties: It has to be general enough to account for regular cases, such that Fisher information could be viewed as a special case of Hellinger information; second, it has to indicate the quality of estimation for arbitrary estimators for both regular and nonregular cases. This chapter offers a definition of Hellinger information that satisfies these two properties and accommodates multidimensional parameters under regular conditions and for a range of nonregular conditions.

As discussed in Chapter 1 Section 1.1.2, Fisher information appears in the quadratic approximation of Hellinger distance as shown in (Equation 1.2). Section 1 of this chapter builds on this fact, showing that for some nonregular models, a non-quadratic approximation of Hellinger distance exists. In section 2, a local minimax risk bound of arbitrary estimation that is based on expansion of Hellinger distance will be presented. In section 3, Hellinger information is defined and some results of its expression for a class of models are presented, with several examples. Throughout the chapter, the relation between Hellinger information and Fisher information will be specified. A summary of results from this chapter, along with a comparison of Hellinger information defined in this project and a previous effort from (Shemyakin, 2014) that explored

alternative measure of information in the context of Bayesian statistics, appear in section 5. This section also argues that the approach of defining Hellinger information presented in this thesis fulfills the two criteria laid out above, and is therefore appropriate in the optimal design of experiment for nonregular cases.

2.1 Hellinger information

2.1.1 Expansion of Hellinger distance

First, consider distributions P_θ, $\theta \in \Theta \subseteq \mathbb{R}^d, d \geq 1$, on Y having μ-densities $p_\theta = dP_\theta/dy$. We can define a model-specific measure of distance on Θ via the Hellinger metric $H(P_\theta, P_\vartheta)$, according to which the function h is defined as the squared Hellinger distance.

Definition 5. (Squared Hellinger distance $h(\theta, \vartheta)$)

$$h(\theta, \vartheta) \equiv H^2(P_\theta, P_\vartheta) := \int (p_\theta^{1/2} - p_\vartheta^{1/2})^2 \, dy = 2 - 2 \int (p_\theta p_\vartheta)^{1/2} \, dy.$$

Recall the definition of DQM from Chapter 1: if $\theta \mapsto p_\theta^{1/2}$ is differentiable in quadratic mean, then, for each θ, the ratio $\|\varepsilon\|_2^{-2} h(\theta, \theta + \varepsilon)$ has a finite and non-zero limit as $\varepsilon \to 0, \varepsilon \in \mathbb{R}^d$, where $\|\cdot\|_2$ denotes the usual ℓ_2-norm. In other words, H^2 has a local quadratic approximation, i.e., with $\dot{l}_\theta = (\frac{\partial \log p_\theta(y)}{\partial \theta_1}, ..., \frac{\partial \log p_\theta(y)}{\partial \theta_d})^\top$, and Fisher information $\mathcal{I}(\theta) = E_\theta(\dot{l}_\theta \dot{l}_\theta^\top)$,

$$h(\theta, \vartheta) = \frac{1}{4} (\theta - \vartheta)^\top \mathcal{I}(\theta) \, (\theta - \vartheta) + o(\|\theta - \vartheta\|_2^2), \text{ as } \theta \to \vartheta. \tag{2.1}$$

For one-dimensional parameters, when p_θ is DQM at θ, then

$$\lim_{\varepsilon\to 0}\frac{h(\theta,\pm\varepsilon)}{|\varepsilon|^2}=\lim_{\varepsilon\to 0}\frac{\int(\sqrt{p_\theta}-\sqrt{p_{\theta+\varepsilon}})^2dy}{|\varepsilon|^2}=\frac{1}{4}\mathcal{I}(\theta).$$

For the nonregular models presented in Chapter 1 Section 1.3, the quadratic approximation of squared Hellinger distance in (Equation 2.1) is no longer available. For some of them, a non-quadratic expansion exists. Here is an one-dimensional example.

Example 9. Uniform distribution has the following expression:

$$p_\theta(y)=\frac{1}{\theta}\,\mathbf{1}_{[\mathbf{0},\theta]}(y).$$

To calculate $\lim_{\varepsilon\searrow 0}|\varepsilon|^{-\alpha}h(\theta,\theta\pm\varepsilon)$, one can calculate $\lim_{\varepsilon\searrow 0}|\varepsilon|^{-\alpha}\int(\sqrt{p_\theta}-\sqrt{p_{\theta+\varepsilon}})^2dy$ and $\lim_{\varepsilon\searrow 0}|\varepsilon|^{-\alpha}\int(\sqrt{p_\theta}-\sqrt{p_{\theta-\varepsilon}})^2dy$ separately, with an appropriate value for α.

$$\begin{aligned}h(\theta,\theta+\varepsilon)=\int_0^\theta(\sqrt{p_\theta}-\sqrt{p_{\theta+\varepsilon}})^2dy&=2-2\int_0^\theta\sqrt{\frac{1}{\theta(\theta+\varepsilon)}}dy\\ =2-2\sqrt{\frac{\theta}{\theta+\varepsilon}}&=2-2\sqrt{1-\frac{\varepsilon}{\theta+\varepsilon}}.\end{aligned}$$

Notice that $\sqrt{1-\frac{\varepsilon}{\theta+\varepsilon}}=1+\frac{1}{2}(-\frac{\varepsilon}{\theta+\varepsilon})+o(\varepsilon)$ as $\varepsilon\to 0$. Letting $\alpha=1$, then applying $\lim_{\varepsilon\searrow 0}\frac{1}{|\varepsilon|}$ to the last expression from above, one obtains

$$\lim_{\varepsilon\searrow 0}\frac{h(\theta,\theta+\varepsilon)}{|\varepsilon|}=\lim_{\varepsilon\searrow 0}\frac{\frac{\varepsilon}{\theta+\varepsilon}+o(\varepsilon)}{|\varepsilon|}=\frac{1}{\theta}.$$

Let $\eta = \theta - \varepsilon$,

$$h(\theta, \theta-\varepsilon) = \int_0^{\theta-\varepsilon} (\sqrt{p_\theta} - \sqrt{p_{\theta-\varepsilon}})^2 dy = \int_0^{\eta} (\sqrt{p_{\eta+\varepsilon}} - \sqrt{p_\eta})^2 dy = h(\theta, \theta+\varepsilon). \tag{2.2}$$

Thus,

$$\lim_{\varepsilon \searrow 0} \frac{h(\theta, \theta-\varepsilon)}{|\varepsilon|} = \lim_{\varepsilon \searrow 0} \frac{\frac{\varepsilon}{\theta} + o(\varepsilon)}{|\varepsilon|} = \frac{1}{\theta}.$$

From the above calculations, it can be shown that when the parameter is one-dimensional, the direction of change (positive or negative) does not matter in finding the expression of the expansion of Hellinger distance.

Therefore, as $\varepsilon \to 0$, for $\mathsf{Unif}(0, \theta)$,

$$h(\theta, \theta \pm \varepsilon) = \frac{1}{\theta}|\varepsilon| + o(|\varepsilon|).$$

This resembles the local Hölder condition on h considered in Chapter 1 Section 6 of (Ibragimov and Hasminskii, 1981). The expansion of $h(\theta, \theta \pm \varepsilon)$ for $\mathsf{Unif}(0, \theta)$ is not quadratic. However, in comparison to the expansion of squared Hellinger distance under DQM case (Equation 2.1), it is apparent that $\frac{1}{\theta}$ takes the place of $\frac{1}{4}\mathcal{I}(\theta)$.

In fact, for many other models with scalar parameter, there exists a positive constant $\alpha < 2$ such that, for each θ, the ratio $|\varepsilon|^{-\alpha} h(\theta, \theta+\varepsilon)$ has a finite and non-zero limit, say $J(\theta)$, such that there is a local approximation,

$$h(\theta, \vartheta) = J(\theta)|\theta - \vartheta|^\alpha + o(|\theta - \vartheta|^\alpha), \text{ as } \vartheta \to \theta. \tag{2.3}$$

Based on Example 9, for $\mathsf{Unif}(0,\theta)$, $J(\theta) = \frac{1}{\theta}$. In order for the definition of "information measure" to be applicable to optimal design of experiment, it must be able to accommodate multi-dimensional parameters. The following provides a definition for expansion of Hellinger information for multi-dimensional parameters.

Definition 6. Let $\Theta \subseteq \mathbb{R}^d$, for $d \geq 1$, and let u denote a generic direction, a d-vector with $\|u\|_2 = 1$. Suppose there exists $\alpha \in (0, 2]$ such that, for all $\theta \in \Theta$ and all directions u, the following limit exists and is neither 0 nor ∞:

$$\lim_{\varepsilon \to 0} \frac{h(\theta, \theta + \varepsilon u)}{|\varepsilon|^\alpha} = J(\theta, u). \tag{2.4}$$

Then the following local approximation holds:

$$h(\theta, \theta + \varepsilon u) = J(\theta, u)|\varepsilon|^\alpha + o(|\varepsilon|^\alpha), \quad \varepsilon \to 0, \tag{2.5}$$

In the above approximation, α is defined as the index of regularity, and $J(\theta, u)$ is defined as Hellinger information at θ in the direction of u.

Here are three brief comments about this definition.

- Based on the fact that squared Hellinger distance is symmetric, i.e., $h(\theta, \theta + u\epsilon) = h(\theta, \theta - u\epsilon)$, this means that $J(\theta, -u) = J(\theta, u)$ is true in general. For one-dimensional parameter, $\theta \in \mathbb{R}$, there are only two possible directions of change: $u = 1, u = -1$, then $J(\theta, 1) = J(\theta, -1)$. Thus, $J(\theta, u)$ is written as $J(\theta)$ in the one-dimensional case.

- This definition includes the case when the regularity index $\alpha = 2$, i.e., when the model is differentiable in quadratic mean at θ, and thus $J(\theta, u) = \frac{1}{4}u^\top \mathcal{I}(\theta)u$.
- Note that this definition does not allow α to depend on u, which means that each component of θ, treated individually, must have the same index of regularity. To better understand this point, consider, for example, a shifted exponential distribution with location parameter θ_1 and rate parameter θ_2. If θ_1 were fixed and θ_2 were the only parameter, then the above definition would hold with $\alpha = 2$. Similarly, if θ_2 were fixed and θ_1 were the only parameter, then it holds with $\alpha = 1$ (see Example 16). However, if both θ_1 and θ_2 are parameters, then the model does not satisfy the conditions of the above definition. To see this, consider two unit vectors $u = (1, 0)$ and $u' = (0, 1)$. If $\alpha = 1$, then $J(\theta, u)$ is in $(0, \infty)$ but $J(\theta, u')$ is zero, which is not allowed; likewise, if $\alpha = 2$, then $J(\theta, u')$ is in $(0, \infty)$ but $J(\theta, u)$ is infinite, which is also not allowed. Therefore, the above definition cannot accommodate situations where the components of θ, treated individually, would have different regularity indices. The design applications we have in mind in this thesis fit naturally within this setting where all components have the same regularity, and the more general case will be presented elsewhere.

In the near-regular Example 1 from Chapter 1 Section 1.2, $J(\theta) = \frac{1}{4\theta(1-\theta)}$ for non-symmetric standard triangular distribution. What follows are some other examples for the one-dimensional case for nonregular models.

Example 10. Following steps similar to those in Example 9, it can be shown that

- If $P_\theta = \mathsf{Unif}(\theta^{-1}, \theta)$, $\theta > 1$, then $\alpha = 1$ and $J(\theta) = (\theta^2 + 1)\{\theta(\theta^2 - 1)\}^{-1}$.

- If $P_\theta = \mathsf{Unif}(\theta, \theta^2)$, $\theta > 1$, then $\alpha = 1$ and $J(\theta) = (2\theta + 1)\{\theta(\theta - 1)\}^{-1}$.

Example 11. Consider a random variable t that adheres to the following version of truncated Weibull with probability distribution function $p_\theta(t)$, with known β, φ,

$$p_\theta(t) = \beta\varphi^\beta t^{\beta-1} \exp\left\{-\varphi^\beta\left(t^\beta - \theta^\beta\right)\right\}, t \in [\theta, \infty), \varphi > 0.$$

If one is only interested in θ, then the regularity index for θ is $\alpha = 1$ and $J(\theta) = \varphi^\beta\theta^{\beta-1}$.

While calculation of $J(\theta)$ for the one-dimensional case is straightforward, this is not true if $d > 1$. Due to the non-quadratic nature of the expansion of Hellinger distance for nonregular cases, it can be difficult to think of the derivation of $J(\theta, u)$ when $\alpha < 2$. However, for the purpose of this project, the multi-dimensional parameter case in the experimental design setting (so far) can all be described as cases in which the multi-dimensional nonregular model can be obtained by reparameterization of a one-dimensional parameter model. Nonregular regression models are described in Chapter 1 Section 1.3.2. The following proposition provides a general result for the expression of Hellinger distance expansion for cases like this.

Proposition 1. *(Reparameterization rule for expansion of Hellinger distance) Consider parametric families, $\{P_\theta, \theta \in \Theta\}, \Theta \subset \mathbb{R}^d$, and $\{Q_\eta, \eta \in H\}$, $H \subset R$. Suppose there exist $\tilde{J}(\eta)$ and $\alpha_\eta > 0$ such that*

$$h(\eta, \eta + \varepsilon) = \tilde{J}(\eta)|\varepsilon|^{\alpha_\eta} + o(|\varepsilon|^{\alpha_\eta}) \text{ as } \varepsilon \to 0.$$

If there is a differentiable function $g(.) : \Theta \to H$, such that for every $\theta \in \Theta$, there is an $\eta = g(\theta) \in H$, and $p_\theta(y) = q_\eta(y)$, for all y, then the regularity index of every $\theta_i, i = 1, .., d$

would be the same as α_η*. Therefore, based on a single observation, with any unit vector* $u = (u_1, ..., u_d)^\top$*, the expansion of Hellinger distance can be expressed as*

$$h(\theta, \theta + u\varepsilon) = \Big| \sum_{j=1}^{d} u_j \frac{\partial g(\theta)}{\partial \theta_j} \Big|^{\alpha_\eta} \tilde{J}(\eta) |\varepsilon|^{\alpha_\eta} + o(|\varepsilon|^{\alpha_\eta}) \text{ as } \varepsilon \to 0.$$

That is, the Hellinger information at θ *in the direction of* u *is*

$$J(\theta, u) = \Big| \sum_{j=1}^{d} u_j \frac{\partial g(\theta)}{\partial \theta_j} \Big|^{\alpha_\eta} \tilde{J}(\eta).$$

Proof. By definition of $\tilde{J}(\eta)$, Hellinger distance of $q_\eta(y)$ and $q_{\eta'}(y)$ has the expansion

$$h(\eta, \eta') = \tilde{J}(\eta) |\eta - \eta'|^{\alpha_\eta} + o(|\eta - \eta'|^{\alpha_\eta}).$$

Since $g(.)$ is differentiable, denote $\dot{g} = (\frac{\partial g(\theta)}{\partial \theta_1}, ..., \frac{\partial g(\theta)}{\partial \theta_d})^\top$. For any given θ, there is an η, such that $g(\theta) = \eta$, with $g(.)$ being differentiable. Thus, for some $\varepsilon > 0$, $g(\theta + \varepsilon u) = g(\theta) + \dot{g}^\top u\varepsilon + o(\varepsilon)$. Furthermore, letting $\Delta = \dot{g}^\top u\varepsilon + o(\varepsilon)$, then $g(\theta + \varepsilon) = \eta + \Delta$. Therefore, $p_\theta(y) = q_\eta(y), p_{\theta+\varepsilon u}(y) = q_{\eta+\Delta}(y)$ which implies $h(\theta, \theta + \varepsilon u) = h(\eta, \eta + \Delta)$. To find $J(\theta, u)$ in the expression of expansion of $h(\theta, \theta + \varepsilon u)$ is to find expression of $\lim_{\varepsilon \to 0} \frac{h(\theta, \theta+\varepsilon u)}{|\varepsilon|^{\alpha_\theta}}$, with α_θ being the regularity index for θ. Notice that

$$\lim_{\varepsilon \to 0} \frac{h(\theta, \theta + \varepsilon u)}{|\varepsilon|^{\alpha_\eta}} = \lim_{\varepsilon \to 0} \frac{h(\eta, \eta + \Delta)}{|\varepsilon|^{\alpha_\eta}} = \lim_{\varepsilon \to 0} \frac{h(\eta, \eta + \Delta)}{|\dot{g}^\top u\varepsilon|^{\alpha_\eta}} |\dot{g}^\top u|^{\alpha_\eta} = \lim_{\Delta \to 0} \frac{h(\eta, \eta + \Delta)}{|\Delta|^{\alpha_\eta}} |\dot{g}^\top u|^{\alpha_\eta}.$$

If one recalls that $\lim_{\Delta\to 0}\frac{h(\eta,\eta+\Delta)}{|\Delta|^{\alpha_\eta}}=\tilde{J}(\eta)$, it can be concluded that regularity index of θ would be the same as that of η, i.e. $\alpha_\theta=\alpha_\eta$, and $h(\theta,u)=\tilde{J}(\eta)\big|\sum_{i=1}^{d}u_i\frac{\partial g(\theta)}{\partial\theta_i}\big|^{\alpha_\eta}$. So, Hellinger distance between $p_\theta(y)$ and $p_{\theta+u\varepsilon}(y)$ has the following expansion:

$$\begin{aligned}h(\theta,\theta+u\varepsilon)&=\tilde{J}(\eta)\big|\dot{g}^\top u\big|^{\alpha_\eta}|\varepsilon|^{\alpha_\eta}+o(|\varepsilon|^{\alpha_\eta})\\&=\tilde{J}(\eta)\big|\sum_{i=1}^{d}u_i\frac{\partial g(\theta)}{\partial\theta_i}\big|^{\alpha_\eta}|\varepsilon|^{\alpha_\eta}+o(|\varepsilon|^{\alpha_\eta}),\ \text{as }\varepsilon\to 0\end{aligned}\tag{2.6}$$

□

The following examples can illustrate how one might apply Proposition 1 to a nonregular model.

Example 12. Consider $\mathsf{Unif}(0,e^{\theta^\top\mathbf{x}}),\theta\in\mathbb{R}^d,d\geq 1$,with fixed $x\in\mathbb{R}^d$.

$$p_\theta(y)=\frac{1}{e^{\theta^\top\mathbf{x}}}\mathbf{1}_{[\mathbf{0},\mathbf{e}^{\theta^\top\mathbf{x}}]}(y).$$

Recall that for $\mathsf{Unif}(0,\eta)$, $J(\theta)=\frac{1}{\eta}$, with 1 as the regularity index. Thus, if θ,x are one-dimensional, by Proposition 1, it follows that

$$J(\theta)=|xe^{\theta x}|\frac{1}{e^{\theta x}}=|x|.$$

If $d>1$, and $\mathbf{x}=(x_1,...,x_d)^\top$, then similarly, Hellinger information at θ in the direction of u would be

$$J(\theta,u)=|e^{\theta^\top\mathbf{x}}\sum_{j=1}^{d}u_jx_j|\frac{1}{e^{\theta^\top\mathbf{x}}}=|\sum_{j=1}^{d}u_jx_j|.\tag{2.7}$$

Proposition 1 is consistent with the reparameterization rule for Fisher information. Under regularity conditions, Fisher information for η and θ exists for $p_\eta(y)$ and $p_\theta(y)$ and $\mathcal{I}(\theta) = (\frac{\partial g(\theta)}{\partial \theta_1}, .., \frac{\partial g(\theta)}{\partial \theta_d})^\top \mathcal{I}(\eta)(\frac{\partial g(\theta)}{\partial \theta_1}, .., \frac{\partial g(\theta)}{\partial \theta_d})$. Notice that this coincides with Proposition 1: when $d = 1$,

$$J(\theta) = \tilde{J}(\eta)\big|\frac{\partial g(\theta)}{\partial \theta}\big|^2 = \frac{1}{4}\mathcal{I}(\eta)\big|\frac{\partial g(\theta)}{\partial \theta}\big|^2 = \frac{1}{4}\mathcal{I}(\theta),$$

and when $d > 1$,

$$J(\theta, u) = |\sum_{i=1}^{d} u_i \frac{\partial g(\theta)}{\partial \theta_i}\big|^2 \tilde{J}(\eta) = u^T(\frac{\partial g(\theta)}{\partial \theta_1}, .., \frac{\partial g(\theta)}{\partial \theta_d})^\top \tilde{J}(\eta)(\frac{\partial g(\theta)}{\partial \theta_1}, .., \frac{\partial g(\theta)}{\partial \theta_d})u = \frac{1}{4}u^\top \mathcal{I}(\theta)u.$$

The next result is useful for finding the expression of Hellinger information in direction u for any function of the parameter.

Proposition 2. *Consider parametric families, $\{P_\theta, \theta \in \Theta\}, \Theta \subset \mathbb{R}^d$, and let $\psi : \Theta \to \mathbb{R}^q, q \leq d$ be a differentiable function with non-singular $q \times d$ derivative matrix $D_\psi(\theta)$. Then, for estimation of $\psi(\theta)$, the Hellinger information of ψ at θ in the direction of u is*

$$J^\psi(\theta, u) = \|D_\psi(\theta)u\|_2^{-\alpha} J(\theta, u).$$

Proof. Letting $\psi : \Theta \to \mathbb{R}^q$ be a differentiable function with non-singular $q \times d$ derivative matrix $D_\psi(\theta)$, when $J(\theta, u)$ is given, but the interest of estimation is $\psi(\theta)$, then the following "chain rule" can be applied in order to find the expression of $J^\psi(\theta, u)$. For some $\varepsilon > 0$,

$\Delta_\psi = \psi(\theta + \varepsilon u) - \psi(\theta) = D_\psi(\theta)u\varepsilon + o(\varepsilon)$. Following a similar argument as in the proof of Proposition 1 above,

$$\begin{aligned} h(\theta, \theta + u\varepsilon) &= J(\theta, u)|\varepsilon|^\alpha + o(|\varepsilon|^\alpha) \\ &= J^\psi(\theta, u)|\Delta_\psi|^\alpha + o(|\varepsilon|^\alpha) \\ &= J^\psi(\theta, u)\|D_\psi(\theta)u\|_2^\alpha|\varepsilon|^\alpha + o(|\varepsilon|^\alpha). \end{aligned}$$

Thus,

$$J^\psi(\theta, u)\|D_\psi(\theta)u\|_2^\alpha = J(\theta, u) \rightarrow J^\psi(\theta, u) = \|D_\psi(\theta)u\|_2^{-\alpha}J(\theta, u).$$

□

2.1.2 Geometric interpretation of $J(\theta, u)$

The following is a brief geometric interpretation of $J(\theta, u)$. As mentioned earlier in the previous Section, whenever the regularity index $\alpha < 2$, one can no longer write the local approximation (Equation 2.5) in a quadratic form; thus, there is not a "direction-free" matrix (i.e., of the kind associated with Fisher information) when $\alpha < 2$. The implication is that the "measure of information" based on the Hellinger expansion for nonregular problems cannot be expressed as a matrix.

One can view $J(\theta, u)$ in (Equation 2.5) as a "directional derivative-like" function in squared Hellinger distance at θ, with exponent α, in the direction u. Given $\varepsilon > 0$, and a point $\theta \in \Theta$, one can compare the approximation to region

$$S(\theta) = \{h(\theta, \theta + \varepsilon u) : |u| = 1\}$$

for the regular case, $\alpha = 2$, and for the nonregular case, $\alpha < 2$.

When $\alpha = 2$, given scaler $\varepsilon > 0$, and some unit vector u, Hellinger distance between p_θ and $p_{\theta+u\varepsilon}$ can be expressed as

$$h(\theta, \theta + u\varepsilon) = \tfrac{1}{4} u^\top \mathcal{I}(\theta)\, u|\varepsilon|^2 + o(|\varepsilon|^2).$$

Due to the fact that $\mathcal{I}(\theta)$ is positive definite, the collection of Hellinger distance between θ and $\theta + u\varepsilon$ over all $\{u : |u| = 1\}$, approximately forms the surface of an ellipsoid with d axes, centered at $0 = h(\theta, \theta)$. Let's denote this as $S_2(\theta)$, expressed as follows:

$$S_2(\theta) = \{\frac{1}{4} u^\top \mathcal{I}(\theta)\, u|\varepsilon|^2 : |u| = 1\}.$$

On the other hand, when $\alpha < 2$, the shape formed by Hellinger distance between θ and $\theta + u\varepsilon$ over all $u : |u| = 1$ is no longer an ellipsoid, but an irregularly shaped surface.

We call $S_\alpha(\theta)$, an irregularly shaped surface because the rate of change of Hellinger distance as $|\varepsilon|$ increases from the center value θ depends on the direction of change in such a way that

it cannot be expressed in a quadratic form. That is, under $\alpha < 2$, an approximation to $S(\theta) = \{h(\theta, \theta + \varepsilon u) : |u| = 1\}$, can be most appropriately expressed in these terms:

$$S_\alpha(\theta) = \{J(\theta, u)|\varepsilon|^\alpha : |u| = 1\}.$$

Apart from its intrinsic interest, the geometric interpretation laid out above will prove useful in understanding the proof of Theorem 1 in the next part, regarding the Hellinger information inequality.

2.2 Hellinger information inequality

Theorem 3 below establishes a suitable connection between the quality of any arbitrary estimator and Hellinger information with direction u. This will provide the necessary foundation for defining a measure of information that is applicable to the question of optimal design for nonregular models.

Suppose the goal is to estimate $\psi(\theta)$ based on sample of size n, $\mathbf{Y^n} = (Y_1, ..., Y_n)$, where $\psi : \mathbb{R}^d \to \mathbb{R}^q$, $q \leq d$, is sufficiently smooth. Let $T_n = T(\mathbf{Y^n})$ be an estimator of $\psi(\theta)$, and measure its quality by the risk,

$$R_\psi(T_n, \theta) = E_\theta \|T_n - \psi(\theta)\|_2^2. \tag{2.8}$$

For the q-vector version of mean square error, expectation, E_θ, is taken with respect to the joint distribution of $Y_1, ..., Y_n$. This covers the case where $\psi(\theta) = \theta$ and $q = d$, (i.e., when the

interest is in the full parameter vector θ), the case where $\psi(\theta)$ is a single component of θ and $q = 1$, as well as other intermediate cases.

Before presenting the Hellinger information inequality, here is how the Hellinger information in the direction of u from a sample of n independent observations is defined. Hellinger information in direction u from a joint distribution can be defined analogously to (Equation 2.4); however, this project defines the independent sample version of $J(\theta, u)$ with the sum formula as presented below. This definition does not affect the finding of the Hellinger information inequality presented in Theorem 3, and it agrees with the familiar notion that information from independent sources is additive.

Definition 7. Let $\mathbf{Y^n} = (Y_1, \ldots, Y_n)$ consist of independent observations with $Y_i \sim P_{i,\theta}$, $i = 1, \ldots, n$, and let $h_{(i)}(\theta, \vartheta)$ denote the i-specific Hellinger distance between $P_{i,\theta}$ and $P_{i,\vartheta}$. Take any fixed $\theta \in \Theta \subseteq \mathbb{R}^d$ and assume that, for any d-dimensional direction u, h_i admits a local expansion at θ as in (Equation 2.5), with $J_{(i)}(\theta, u)$ and regularity index $\alpha \in (0, 2]$. The Hellinger information at θ in the direction of u based on sample $\mathbf{Y^n}$ is defined as

$$\mathcal{J}_n(\theta, u) = \sum_{i=1}^{n} J_{(i)}(\theta, u).$$

Theorem 3. *Let $\mathbf{Y^n} = (Y_1, \ldots, Y_n)$ consist of independent observations with $Y_i \sim P_{i,\theta}$, $i = 1, \ldots, n$, and $\mathcal{J}_n(\theta, u)$, the Hellinger information at θ in the direction of u of sample $\mathbf{Y^n}$ exists. Let $\psi : \Theta \to \mathbb{R}^q$ be a differentiable function with non-singular $q \times d$ derivative matrix $D_\psi(\theta)$, and*

let $T_n = T(\mathbf{Y^n})$ *be any estimator of* $\psi(\theta)$ *with risk in* (Equation 2.8). *If* $\varepsilon_{n,u} = \{3\mathcal{J}_n(\theta,u)\}^{-1/\alpha}$, *and*

$$\liminf_{n\to\infty} n^{-1}\mathcal{J}_n(\theta,u) > 0, \tag{2.9}$$

then

$$R_\psi(T_n, \theta+\varepsilon_u u) + R_\psi(T_n,\theta) \gtrsim \|D_\psi(\theta)\,u\|_2^2\,\mathcal{J}_n(\theta,u)^{-2/\alpha}, \quad \textit{for all large } n. \tag{2.10}$$

Consequently, if $B_n(\theta)$ *is the region whose boundary is determined by the union of* $\{\theta+\varepsilon_{n,u}u\}$ *over all directions* u, *then for all large* n

$$\inf_{T_n}\sup_{\vartheta\in B_n(\theta)} R_\psi(T_n,\vartheta) \gtrsim \big\{\inf_u \|D_\psi(\theta)\,u\|_2^{-\alpha}\,\mathcal{J}_n(\theta,u)\big\}^{-2/\alpha}. \tag{2.11}$$

Proof. See Appendix 2.5.2 □

Three brief comments about the result in Theorem 3:

- The universal constant hidden in "$\gtrsim$" is known and given in the proof.
- There is nothing special about "3" in the definition of ε_u; any number strictly greater than 2 would suffice.
- Based on the reparametization rule of Proposition 1, the Hellinger information of $\psi(.)$ at θ in the direction of u, based on a sample of size n, shall be $J_n^\psi(\theta,u) = \|D_\psi(\theta)u\|_2^{-\alpha} J_n(\theta,u)$.

Thus, one can rewrite the right-hand side of (Equation 2.11), and the inequality would then be expressed as

$$\inf_{T_n} \sup_{\vartheta \in B_n(\theta)} R_\psi(T_n, \vartheta) \gtrsim \left\{ \inf_{u:\|u\|_2=1} \mathcal{J}_n^\varphi(\theta, u) \right\}^{-2/\alpha}. \tag{2.12}$$

Some additional comments about the interpretation of Theorem 3 are in order. First, the reason for the sum of two risks, or the supremum over a "neighborhood" of θ, is that a lucky choice of $T_n \equiv \theta$ has excellent performance at θ, but not such good performance at a nearby ϑ. The theorem says that, if one looks at a locally uniform measure of risk, which prevents "cheating" towards a particular θ, then one cannot do better, in terms of risk, than the lower bound in (Equation 2.11). The classical Cramér–Rao lower bound uses unbiasedness of the estimator to prevent this kind of "cheating".

To assess how sharp the bound in (Equation 2.11) is outside of regular cases, consider the case where $q = 1$, so that $\psi(\theta)$ is a scalar function. Regarding the rate, if we consider the identically independently distributed case, so that $\mathcal{J}_n(\theta, u) = nJ_1(\theta, u)$, it follows that the lower bound is of order $n^{-2/\alpha}$, which agrees with the known minimax rate for estimators in nonregular models (Ibragimov and Hasminski 1981, Chapter 1 Sec.5). Therefore, the bound cannot be improved in terms of how it depends on the sample size. To assess the quality of the lower bound in terms of its dependence on θ, if the observations come from $\mathsf{Unif}(0, \theta)$, which

has $\alpha = 1$ and $J(\theta) = \theta^{-1}$, the maximum likelihood estimator is the sample maximum, and its mean square error is given by

$$\frac{\theta^2 n}{(n+1)^2(n+2)} + \Big(\frac{\theta n}{n+1} - \theta\Big)^2.$$

Asymptotically, this expression is equivalent to $\theta^2 n^{-1}$, which agrees with our lower bound. Therefore, up to universal constants, the bound in Theorem 3 is sharp. The question of whether a particular estimator for a specific nonregular model can attain the bound exactly, or asymptotically, apparently needs to be addressed case by case. Although not the focus here, this deserves further investigation.

How does the bound compare to Cramér–Rao in the regular case, with $\alpha = 2$? The following discussion explores this question.

Corollary 1. *When $\alpha = 2$, Fisher information of a sample of size n, $\mathcal{I}_n(\theta)$, exists. When $\psi(.) : \mathbb{R}^d \to R^q$, letting $\lambda_{min}(.)$ be the function of minimum eigenvector, then the expression of the right-hand side of the risk bound from* (Equation 2.11) *would have the expression of*

$$\big\{\inf_u \|D_\psi(\theta)\, u\|_2^{-\alpha}\, \mathcal{J}_n(\theta; u)\big\}^{-2/\alpha} = \frac{1}{4}\lambda_{min}\big\{(D_\psi(\theta)^\top \mathcal{I}_n^{-1}(\theta) D_\psi(\theta))^{-1}\big\}.$$

Proof. When $\alpha = 2$, Fisher information $\mathcal{I}_n(\theta)$ exists and $\mathcal{J}_n(\theta, u) = \frac{1}{4} u^\top \mathcal{I}_n(\theta) u$. When $\psi(.) : R^d \to R^q$, then $D_\psi(\theta)$ is $q \times d$. So, by (Equation 2.13),

$$\min_{u:\|u\|_2=1} \|D_\psi(\theta)u\|_2^{-2} \mathcal{J}_n(\theta, u) = [\max_{u:\|u\|_2=1} 4\frac{\|u^\top D_\psi(\theta)\|_2^2}{u^\top \mathcal{I}_n(\theta) u}]^{-1}.$$

By Cholesky decomposition, let $\mathcal{I}_n(\theta) = M^\top M$. Since $D_\psi(\theta)^\top \mathcal{I}_n^{-1}(\theta) D_\psi(\theta)$ is positive definite,

$$\begin{aligned}
\frac{\|u^\top D_\psi(\theta)\|_2^2}{u^\top \mathcal{I}_n(\theta) u} &= \frac{\|u^\top M^\top (M^\top)^{-1} D_\psi(\theta)\|_2^2}{u^\top \mathcal{I}_n(\theta) u} \\
&\leq \frac{\|u^\top M^\top M u\|_2 \|D_\psi(\theta)^\top M^{-1} (M^{-1})^\top D_\psi(\theta)\|_2}{u^\top \mathcal{I}_n(\theta) u} \\
&= \|D_\psi(\theta)^\top \mathcal{I}_n^{-1}(\theta) D_\psi(\theta)\|_2 \\
&= \lambda_{max}(\frac{1}{4} D_\psi(\theta)^\top \mathcal{I}_n^{-1}(\theta) D_\psi(\theta)) \\
&= \frac{1}{4} \lambda_{min}[(D_\psi(\theta)^\top \mathcal{I}_n^{-1}(\theta) D_\psi(\theta))^{-1}]
\end{aligned}$$

The above inequality takes an equal sign only when u is dependent with respect to columns in $D_\psi(\theta)^\top$.

In conclusion, when $\alpha = 2$,

$$\big\{\inf_u \|D_\psi(\theta)\, u\|_2^{-\alpha}\, \mathcal{J}_n(\theta; u)\big\}^{-2/\alpha} = \frac{1}{4} \lambda_{min}\big\{(D_\psi(\theta)^\top \mathcal{I}_n^{-1}(\theta) D_\psi(\theta))^{-1}\big\}.$$

□

Now, let's compare this to the Cramér–Rao lower bound for estimation of $\psi(\theta)$. When $q = 1$, the local minimax risk bound (Equation 2.11) from Theorem 3 coincides with the expression of the lower bound in (Equation 1.1), the Cramér-Rao lower bound:

$$\inf_{T_n} \sup_{\vartheta \in B_n(\theta)} R_\psi(T_n, \vartheta) \gtrsim D_\psi(\theta)^\top \mathcal{I}_n^{-1}(\theta) D_\psi(\theta), \quad \text{for all large } n.$$

When the Cramér-Rao lower bound exists, Corollary 1 shows that it is proportional to the lower bound from Theorem 3. As the Hellinger information inequality described in Theorem 3 covers both regular and some nonregular models, we can conclude that Theorem 3 is a more general result.

2.2.1 Definition of Hellinger information

In the context of optimal design, the measure of information must provide a measure of quality of estimation. So far, given independent sample $y_i, i = 1, ..., n$, when the interest of estimation is the parameter θ itself, $\mathcal{J}_n(\theta, u) = \sum_i^n J_{(i)}(\theta, u)$ defines the Hellinger information on θ in the direction u of the sample. However, what determines the quality of arbitrary estimators, based on Theorem 3, is $\mathcal{J}_n(\theta, u)$ at the minimum direction. Based on these observations, this project proposes that the measure of information for parameter with regularity index of $\alpha \leq 2$ would be $\min_{u:\|u\|_2=1} \mathcal{J}_n(\theta, u)$.

Definition 8. Let $Y_i \sim P_{i,\theta}$, $i = 1, \ldots, n$ be independently distributed, with $\theta \in \Theta \subseteq \mathbb{R}^d$ being a fixed but unknown parameter. Further, assume that each $P_{i,\theta}$ has the same index of regularity, $\alpha \in (0, 2]$, and the following expansion exists:

$$h_i(\theta, \theta + \varepsilon u) = J_{(i)}(\theta, u)|\varepsilon|^\alpha + o(|\varepsilon|^\alpha), |u| = 1, \quad \varepsilon \to 0.$$

Then the Hellinger information of function $\psi(.)$ at θ based on the sample is defined as

$$\mathcal{J}_n^\psi(\theta) = \min_{u:\|u\|_2=1} \mathcal{J}_n^\varphi(\theta, u). \tag{2.13}$$

With $\mathcal{J}_n^{\varphi}(\theta, u) = \|D_{\psi}(\theta)\, u\|_2^{-\alpha}\, \mathcal{J}_n(\theta, u)$, the Hellinger information at θ based on the sample is defined as follows: When ψ is identity function, then Hellinger information at θ is

$$\mathcal{J}_n(\theta) = \min_{u:|u|=1} \mathcal{J}_n(\theta, u). \tag{2.14}$$

Under DQM, based on Corollary 1, Hellinger information is proportional to the minimum-eigenvalue of Fisher information of $\psi(\theta)$:

$$\mathcal{J}_n^{\psi}(\theta) = \{\frac{1}{4}\lambda_{max}(D_{\psi}(\theta)^{\top}\mathcal{I}_n^{-1}(\theta)D_{\psi}(\theta))\}^{-1} = \lambda_{min}\{\frac{1}{4}(D_{\psi}(\theta)^{\top}\mathcal{I}_n^{-1}(\theta)D_{\psi}(\theta))\}^{-1}.$$

Example 13. Based on n independent observations $(y_i, \mathbf{x}_i), \mathbf{x}_i = (x_{i,1}...x_{i,d})^{\top}, i = 1, ..., n$ generated from $\mathsf{Unif}(0, e^{\theta^{\top}\mathbf{x}_i}), \theta \in \mathbb{R}^d$, Hellinger information has the following expression:

$$\mathcal{J}_n(\theta) = \min_{u:\|u\|_2=1} |\sum_{i=1}^{n}\sum_{j=1}^{d} u_j x_{i,j}|.$$

2.2.2 A general result for a class of nonregular models

This subsection introduces a general result for Hellinger information for a class of nonregular models discussed in chapter 1 section 1.3.3, described as $p_{\theta}(y)$ from (Equation 1.12). Recall that $p_{\theta}(y) = p_0(y - \theta)$, and $p_0(y)$ from (Equation 1.11) can be further specified as

$$p_0(y) = \alpha f(y) y^{\alpha-1} \mathbf{1}_{[\mathbf{0},\infty]}(y), \tag{2.15}$$

where $f(y)$ is bounded away from zero and infinity as $y \to 0$. (That is, the function is "slowly varying" and is sufficiently smooth.) As described in (Equation 1.11), there is some constant $c > 0$ such that $\lim_{y \to 0} f(y) = c$. A good example is a gamma distribution with shape α, in which case $f(y) = \{\alpha\Gamma(\alpha)\}^{-1} e^{-y}$, possibly including a known scalar parameter.

Proposition 3. *Suppose p_0 such that, for any $\Delta > 0$,*

$$\int_{\Delta}^{\infty} \Big(\frac{d}{dy} \log p_0(y) \Big)^2 p_0(y)\, dy < \infty. \tag{2.16}$$

If $\alpha \in [1, 2)$, then $h(\theta, \theta + \varepsilon) = f(0)\varepsilon^{\alpha} + o(\varepsilon^{\alpha})$, as $\varepsilon \to 0$; that is, Hellinger information at θ is $\mathcal{J}(\theta) = f(0)$.[1]

Proof. See Appendix 2.5.1 □

Example 14. The distribution function of shifted Weibull for a given shape parameter K is

$$p_{\eta}(y) = \frac{K}{\lambda} (\frac{y - \eta}{\lambda})^{K-1} \exp[-(\frac{y - \eta}{\lambda})^{K}], y \geq \eta, 1 \leq K < 2, \lambda > 0.$$

Since the Weibull(K) distribution function is $p_0(z) = K(\frac{1}{\lambda})^K (z)^{K-1} exp(-(\frac{z}{\lambda})^K)$, then as $z \to 0$, let's denote it as

$$p(z) = K(\frac{1}{\lambda})^K (z)^{K-1}. \tag{2.17}$$

---

[1] *Theorem VI.1.1 in (Ibragimov and Hasminskii, 1981) shows that $h(\theta, \theta + \varepsilon) = O(\varepsilon^{\alpha})$ as $\varepsilon \to 0$, but they don't give the constant $f(0)$. Condition* (Equation 2.16) *is basically the same as Condition C_5 in (Woodroofe, 1974), which is basically the same as Assumption 9 in (Smith, 1985). Therefore, it can be checked for the gamma case as well as others.*

Letting $F_0(z)$ be the cumulative distribution of $p_0(z)$, then $F_0(\varepsilon) = 1 - exp(-(\frac{\varepsilon}{\lambda})^K)$. Thus Hellinger information for η would be

$$\mathcal{J}(\eta) = \lim_{\varepsilon \to 0} \frac{1 - exp(-(\frac{\varepsilon}{\lambda})^K)}{(\varepsilon)^K} = (\frac{1}{\lambda})^K, \text{ and the regularity index would be } K.$$

Example 15. The distribution function of the shifted gamma distribution for a given shape parameter $\beta \in [1, 2)$ is

$$p_\eta(y) = \frac{1}{\Gamma(\beta)}(y - \eta)^{\beta - 1} \exp[-(y - \eta)]\mathbf{1}_{[\eta, \infty]}(y).$$

Gamma(β) distribution is $p_0(z) = \frac{1}{\Gamma(\beta)}(z)^{\beta-1} \exp(-z), z \geq 0$, and as $z \to 0$, $\exp(-z) \to 1$, thus $p_0(z) \to \frac{1}{\Gamma(\beta)}(z)^{\beta-1}$. Let $F_0(z)$ be the cumulative distribution of $p_0(z)$, $F_z(\varepsilon) = \frac{\gamma(\beta,\varepsilon)}{\Gamma(\beta)}$. Then, as $\varepsilon \to 0$, $\gamma(\beta, \varepsilon) = \frac{\varepsilon^\beta}{\beta}$, following Proposition 3,

$$\mathcal{J}(\eta) = \lim_{|\varepsilon| \to 0} \frac{\gamma(\beta, \varepsilon)}{\Gamma(\beta)} \frac{1}{|\varepsilon|^\alpha} = \frac{\varepsilon^\beta}{\beta\Gamma(\beta)|\varepsilon|^\alpha} = \frac{1}{\beta\Gamma(\beta)},$$

and the regularity index of η would be β.

Example 16. The distribution function of shifted exponential for a given rate parameter θ_2 is

$$p_{\theta_1}(y) = \theta_2 \exp[-\theta_2(y - \theta_1)]\mathbf{1}_{[\theta_1, \infty]}(y).$$

Based on Proposition 3, the regularity index of the location parameter θ_1 is $\alpha = 1$, and $\mathcal{J}(\theta_1) = \theta_2$.

2.2.3 Expression of Hellinger information for the nonregular regression model

Now we are ready to obtain Hellinger information with respect to the nonregular regression model described in Chapter 1. Recall that the nonregular regression model can be described as

$$y_i = g(\theta; \mathbf{x}_i) + \varepsilon_i, i = 1, ..., n, \theta \in \mathbb{R}^d, \tag{2.18}$$

where ε_i follows distribution that satisfies (Equation 2.15). That is, there are some $\alpha > 0, c > 0$, such that $p_0(\varepsilon_i) = \alpha\, c\, \varepsilon_i^{\alpha-1}$, as $\varepsilon_i \to 0$. Then, distribution of y_i given $\mathbf{x}_i$ would be

$$p_{i,\theta}(y_i) = \alpha c (y_i - g(\theta; \mathbf{x}_i))^{\alpha-1}, \text{ as } y_i \to g(\theta; \mathbf{x}_i).$$

By Proposition 3, Hellinger information for η from $p_\eta(y) = \alpha c (y-\eta)^{\alpha-1}$ as $\alpha \in [1, 2)$ would be $J(\eta) = c$. By Proposition 1,

$$\mathcal{J}_n(\theta, u) = \sum_{i=1}^{n} \Big| \sum_{j=1}^{d} u_j \frac{\partial g(\theta, \mathbf{x}_i)}{\partial \theta_j} \Big|^\alpha c. \tag{2.19}$$

Based on Example 14 and Example 15, and Proposition 1, Hellinger information for an independent sample, $(y_i, x_i), i = 1, ..., n$, from the regression model (Equation 3.1) with Gamma(β) distributed error, would have the following expression:

$$\mathcal{J}_n(\theta) = \min_{u:\|u\|_2=1} \sum_{i=1}^{n} \Big| \sum_{j=1}^{d} u_j \frac{\partial g(\theta, x_i)}{\partial \theta_j} \Big|^{\beta} \frac{1}{\beta\Gamma(\beta)}, \text{with regularity index } \beta;$$

under Weibull(K) distributed error, then,

$$\mathcal{J}_n(\theta) = \min_{u:\|u\|_2=1} \sum_{i=1}^{n} \Big| \sum_{j=1}^{d} u_j \frac{\partial g(\theta, x_i)}{\partial \theta_j} \Big|^{K} (\frac{1}{\lambda})^{K}, \text{with regularity index } K.$$

Since it is assumed that the shape parameter (β for gamma case, and K for Weibull case) in error distribution is either known or considered a nuisance parameter in the experiments where the focus is on the estimation of θ, the value of $\frac{1}{\beta\gamma(\beta)}, (\frac{1}{\lambda})^K$ is not important, nor, in general, is the value of c from the error distribution $p_0(\varepsilon_i)$. The Hellinger information for nonregular regression model with regularity index α can thus be expressed as

$$\mathcal{J}_n(\theta) \propto \min_{u:\|u\|_2=1} \sum_{i=1}^{n} \Big| \sum_{j=1}^{d} u_j \frac{\partial g(\theta, x_i)}{\partial \theta_j} \Big|^{\alpha}. \tag{2.20}$$

2.3 Summary

This chapter presented a definition of Hellinger information as an extension of Fisher information from regular to nonregular models. This definition of Hellinger information is meaningful both in terms of its derivation and its relationship to the quality of estimator. The main

result, the Hellinger information inequality, presented in Theorem 3, states that the Hellinger information determines a lower bound of local minimax risk bound for any estimators.

Although it is not clear what approach is efficient to the risk bound from Theorem 3 in terms of estimation, this provides a guideline to compare designs. In the regression setting, as mentioned in Chapter 1, Fisher information depends on the covariates of a sample, which means that it depends on the design of an experiment. The same can be said about the Hellinger information by its expression in (Equation 2.20). The goal of optimal experimental design for nonregular models can be framed in terms of maximizing the Hellinger information with respect to designs.

In fact, optimization of Hellinger information under regularity condition, based on Corollary 1, is equivalent to optimization of minimum eigenvalue of Fisher information. For, when Fisher information exists,

$$\mathcal{J}_n(\theta) = \min_{u:\|u\|_2=1} u^\top \mathcal{I}(\theta) u.$$

Thus, optimal design obtained from optimization of Hellinger information when $\alpha = 2$ is equivalent to E-optimal design.

The next chapter presents optimal design results based on Hellinger information for several nonregular regression models.

2.4 A comparison with Shemyakin (2014)

One inspiration for this project comes from recent developments in Bayesian statistics, where Fisher information can function as non-informative prior for regular parameters. The idea of

using Hellinger information as an extension of Fisher information for non-informative prior was previously suggested by (Shemyakin, 2014).

The definition of one-dimensional Hellinger information in this thesis mostly coincides with that proposed by (Shemyakin, 2014). Specifically, considering Hellinger information $\mathcal{J}(\theta)$ and regularity index α, (Shemyakin, 2014) suggested to use $\mathcal{J}^{2/\alpha}(\theta)$ as the definition instead. Theorem 3 from (Shemyakin, 2014) presented a result for Hellinger information for one-dimensional parameter in the $\alpha = 1$ case, which can be viewed as a special case for Proposition 2 in this thesis. However, the common ground between this project and Shemyakin's does not extend beyond this point.

First, regarding the information inequality, (Shemyakin, 2014) presented a different inequality that can be viewed as an integral version of the Cramér-Rao inequality for the one-dimensional parameter: Supposing $\pi(\theta)$ is a prior distribution for $\theta \in \Theta$, then with some constant $C(\alpha)$,

$$\int (\hat{\theta} - \theta)^2 p_\theta(y)\pi(\theta)dyd\theta \geq C(\alpha)n^{-2/\alpha} \int_\Theta J^{-2/\alpha}(\theta)\pi(\theta)d\theta + o(n^{-2/\alpha}).$$

A reference to an earlier paper (Shemyakin, 1993) on this result is given but, unfortunately, the details seem not to be available in English. While this is an interesting result, the integration over θ is not appropriate for our non-Bayesian formulation here and, moreover, we require an analogous result for vector parameters as well, and we are not aware of a multi-dimensional

version of the above bound. The result in Theorem 3 does not require integration over the parameter space, and holds for scalar and vector parameters simultaneously.

The second major point of departure between this project and (Shemyakin, 2014) is that Shemyakin defines a so-called "Hellinger information matrix" for multi-dimensional nonregular parameters. He proposes that the Hellinger information matrix be defined by its components as follows:

$$\lim_{\varepsilon\to 0}\frac{D_{i,j}^{1/\alpha_i+1/\alpha_j}(\theta,\alpha_i,\alpha_j)}{|\varepsilon|^{-2}},\ \text{and with}$$

$$D_{i,j}(\theta,\alpha_i,\alpha_j)=\int(\sqrt{p_{\theta_1\ldots\theta_m}(y)}-\sqrt{p_{\theta_1\ldots\theta_i+\varepsilon^{2/\alpha_i}\ldots\theta_m}(y)})(\sqrt{p_{\theta_1\ldots\theta_m}(y)}-\sqrt{p_{\theta_1\ldots\theta_j+\varepsilon^{2/\alpha_j}\ldots\theta_m}(y)}).$$

The definition seems to contradict our claim in Chapter 2 that no such matrix is available. This is not a contradiction, however, because Shemyakin's definition of "Hellinger information matrix" is ad hoc, chosen only so that it agrees with the Fisher information in regular cases, but only in a formal manner. That is, the above matrix does not describe the local behavior of the Hellinger distance and, therefore, does not lead to an information inequality as was developed in Theorem 3.

2.5 Appendix

2.5.1 Proof of Proposition 3

It is easy to check that

$$h(\theta,\theta+\varepsilon)=\int_{\theta}^{\theta+\varepsilon}\left(\sqrt{p_{\theta+\varepsilon}(y)}-\sqrt{p_{\theta}(y)}\right)^2dy+\int_{\theta+\varepsilon}^{\infty}\left(\sqrt{p_{\theta+\varepsilon}(y)}-\sqrt{p_{\theta}(y)}\right)^2dy.$$

The first integral is easy to calculate:

$$\int_{\theta}^{\theta+\varepsilon} \left(\sqrt{p_{\theta+\varepsilon}(y)} - \sqrt{p_{\theta}(y)}\right)^2 dy = \int_{\theta}^{\theta+\varepsilon} p_{\theta}(y)\, dy = f(0)\varepsilon^{\alpha}, \quad \varepsilon \to 0.$$

So, it remains to be shown that the second integral is $o(\varepsilon^{\alpha})$. After some change of variables, this boils down to showing that

$$\int_{0}^{\infty} \left(\sqrt{p_0(z+\varepsilon)} - \sqrt{p_0(z)}\right)^2 dz = o(\varepsilon^{\alpha}).$$

Take $\Delta > 0$ as in (Equation 2.16). Then we can split the integral above, like so:

$$\int_{0}^{\Delta} \left(\sqrt{p_0(z+\varepsilon)} - \sqrt{p_0(z)}\right)^2 dz + \int_{\Delta}^{\infty} \left(\sqrt{p_0(z+\varepsilon)} - \sqrt{p_0(z)}\right)^2 dz.$$

It follows from (Equation 2.16) and the dominated convergence theorem that the second term satisfies

$$\int_{\Delta}^{\infty} \left(\sqrt{p_0(z+\varepsilon)} - \sqrt{p_0(z)}\right)^2 dz = O(\varepsilon^2) = o(\varepsilon^{\alpha}), \quad \varepsilon \to 0.$$

Thus, we must prove that $\int_0^{\Delta} (\sqrt{p_0(z+\varepsilon)} - \sqrt{p_0(z)})^2\, dz = o(\varepsilon^{\alpha})$. To start, adding and subtracting $\sqrt{f(z+\varepsilon)z^{\alpha-1}}$ yields the equation

$$\begin{aligned}
&\sqrt{p_0(z+\varepsilon)} - \sqrt{p_0(z)} \\
&\qquad = \sqrt{f(z+\varepsilon)}\left((z+\varepsilon)^{\frac{\alpha-1}{2}} - z^{\frac{\alpha-1}{2}}\right) + \left(\sqrt{f(z+\varepsilon)} - \sqrt{f(z)}\right) z^{\frac{\alpha-1}{2}},
\end{aligned}$$

so that the integrand becomes

$$\big(\sqrt{p_0(z+\varepsilon)}-\sqrt{p_0(z)}\big)^2 = I_1(z;\varepsilon)+I_2(z;\varepsilon)+I_3(z;\varepsilon),$$

where

$$\begin{aligned}
I_1(z;\varepsilon) &= f(z+\varepsilon)\big((z+\varepsilon)^{\frac{\alpha-1}{2}}-z^{\frac{\alpha-1}{2}}\big)^2\\
I_2(z;\varepsilon) &= \big(\sqrt{f(z+\varepsilon)}-\sqrt{f(z)}\big)^2 z^{\alpha-1}\\
I_3(z;\varepsilon) &= 2\sqrt{f(z+\varepsilon)}\big(\sqrt{f(z+\varepsilon)}-\sqrt{f(z)}\big)\big((z+\varepsilon)^{\frac{\alpha-1}{2}}-z^{\frac{\alpha-1}{2}}\big)z^{\frac{\alpha-1}{2}}.
\end{aligned}$$

The second term, I_2, is the easiest to deal with, so we take this one first. Because f is smooth and slowly varying near zero, the mean value theorem says that $\sqrt{f(z+\varepsilon)}-\sqrt{f(z)} \lesssim \varepsilon$, which implies that

$$\int_0^\Delta I_2(z;\varepsilon)\,dz \lesssim \varepsilon^2\int_0^\Delta z^{\alpha-1}\,dz \lesssim \varepsilon^2 = o(\varepsilon^\alpha), \quad \varepsilon\to 0.$$

The third term, I_3, is similar. That is, after applying the mean value theorem to both of the differences in I_3, we discover that

$$\int_0^\Delta I_3(z;\varepsilon)\,dz \lesssim \varepsilon^2\int_0^\Delta z^{-\frac{3-\alpha}{2}}z^{\frac{\alpha-1}{2}}\,dz = \varepsilon^2\int_0^\Delta z^{\alpha-2}\,dz = o(\varepsilon^\alpha) \text{ if } \alpha\geq 1.$$

Since the integral converges, the upper bound is $O(\varepsilon^2) = o(\varepsilon^\alpha)$ as $\varepsilon \to 0$. Since $(z+\varepsilon)^\beta - z^\beta$, for $\beta < 1$, the first term, I_1, is maximized at $z = 0$, and so we have

$$I_1(z;\varepsilon) \lesssim \varepsilon^{\alpha-1} |(z+\varepsilon)^{\frac{\alpha-1}{2}} - z^{\frac{\alpha-1}{2}}|.$$

Apply the mean value theorem to the difference in this upper bound, and we get

$$I_1(z;\varepsilon) \lesssim \varepsilon^\alpha z^{-\frac{3-\alpha}{2}}.$$

As with I_3, the integral of this upper bound is finite and, in particular, is proportional to $\Delta^{\frac{\alpha-1}{2}}$. Putting everything together, we have

$$\int_0^\infty \left(\sqrt{p_0(z+\varepsilon)} - \sqrt{p_0(z)}\right)^2 dz = \Delta^{\frac{\alpha-1}{2}} \varepsilon^\alpha + o(\varepsilon^\alpha), \quad \varepsilon \to 0.$$

Since Δ can be arbitrarily small, the right-hand side is $o(\varepsilon^\alpha)$, which completes the proof.

2.5.2 Proof of Theorem 3

A key to the proof of Theorem 3 is to make a connection between the Hellinger distance and the risk of an estimator. This first step is based, in part, on the analysis in Chapter 1 Section 6 of (Ibragimov and Hasminskii, 1981), although our setup and conclusions are more general in certain ways. We summarize this first step in the following lemma.

Lemma 1. *For data $Y \in \mathbb{Y}$, consider a model P_θ, with μ-density p_θ, indexed by a parameter $\theta \in \Theta \subseteq \mathbb{R}^d$. Let $\psi = \psi(\theta)$ be the interest parameter, where $\psi : \mathbb{R}^d \to \mathbb{R}^q$. For an estimator $T = T(Y)$ of ψ, the risk function $R_\psi(T, \theta)$ for the estimator T satisfies*

$$R_\psi(T,\theta) + R_\psi(T,\vartheta) \geq \min\Big\{\frac{1-h(\theta,\vartheta)}{4h(\theta,\vartheta)}, \frac{1}{16}\Big\}\|\psi(\theta)-\psi(\vartheta)\|^2.$$

Proof. Define the mean function of the estimator T, i.e., $m_\psi(\theta) = E_\theta(T)$. Since integration of a constant function with respect to the (signed) measure with density $p_\theta - p_\vartheta$ is zero, we have the following identity:

$$m_\psi(\theta) - m_\psi(\vartheta) = \int \big[T(y) - \tfrac{1}{2}\{m_\psi(\theta) + m_\psi(\vartheta)\}\big]\big[p_\theta(y) - p_\vartheta(y)\big]\,\mu(dy).$$

Write $v_{\theta,\vartheta}(y) = T(y) - \frac{1}{2}\{m_\psi(\theta) + m_\psi(\vartheta)\}$. Now bound the norm of the quantity in the above display:

$$\begin{aligned}\|m_\psi(\theta) - m_\psi(\vartheta)\| &= \Big\|\int v_{\theta,\vartheta}(p_\theta - p_\vartheta)\,dy\Big\| \\ &\leq \int \|v_{\theta,\vartheta}\|\,|p_\theta^{1/2} + p_\vartheta^{1/2}|\,|p_\theta^{1/2} - p_\vartheta^{1/2}|\,dy.\end{aligned}$$

Next, apply the Cauchy–Schwartz inequality, to get

$$\|m_\psi(\theta) - m_\psi(\vartheta)\|^2 \leq \int \|v_{\theta,\vartheta}\|^2\,|p_\theta^{1/2} + p_\vartheta^{1/2}|^2\,dy \cdot h(\theta,\vartheta).$$

For two non-negative numbers a and b, we have $(a+b)^2 \leq 2(a^2+b^2)$, so the first term in the above upper bound is itself bounded by

$$2\int \|v_{\theta,\vartheta}\|^2 p_\theta \, dy + 2\int \|v_{\theta,\vartheta}\|^2 p_\vartheta \, dy.$$

If we rewrite $v_{\theta,\vartheta}$ as

$$v_{\theta,\vartheta}(y) = \{T(y) - m_\psi(\theta)\} + \tfrac{1}{2}\{m_\psi(\vartheta) - m_\psi(\theta)\},$$

and use the fact that $\int \{T - m_\psi(\theta)\} p_\theta \, dy = 0$, then we get

$$\int \|v_{\theta,\vartheta}\|^2 \, p_\theta \, dy = R_\psi(T,\theta) + \tfrac{1}{4}\|m_\psi(\theta) - m_\psi(\vartheta)\|^2.$$

An analogous bound holds for $\int \|v_{\theta,\vartheta}\|^2 \, p_\vartheta \, dy$, so we get

$$\|m_\psi(\theta) - m_\psi(\vartheta)\|^2 \leq 2h(\theta,\vartheta)\big\{R_\psi(T,\theta) + R_\psi(T,\vartheta) + \tfrac{1}{2}\|m_\psi(\theta) - m_\psi(\vartheta)\|^2\big\}.$$

Rearranging terms gives the bound

$$R_\psi(T,\theta) + R_\psi(T,\vartheta) \geq \frac{1 - h(\theta,\vartheta)}{h(\theta,\vartheta)} \|m_\psi(\theta) - m_\psi(\vartheta)\|^2.$$

Finally, write $b_\psi(\theta) = m_\psi(\theta) - \psi(\theta)$ for the bias function of T, and consider the following two exhaustive cases based on the magnitude of the bias.

- Suppose that $\max\{|b_\psi(\theta)|, |b_\psi(\vartheta)|\} < \frac{1}{4}\|\psi(\theta) - \psi(\vartheta)\|$. Then it follows from the triangle inequality that

$$\|m_\psi(\theta) - m_\psi(\vartheta)\| = \|\psi(\theta) - \psi(\vartheta) + b_\psi(\theta) - b_\psi(\vartheta)\| \geq \tfrac{1}{2}\|\psi(\theta) - \psi(\vartheta)\|.$$

- Next, suppose that, say, $\|b_\psi(\theta)\| \geq \frac{1}{4}\|\psi(\theta) - \psi(\vartheta)\|$. Then we trivially have $R_\psi(T, \theta) \geq \|b_\psi(\theta)\|^2$ and, therefore, $R_\psi(T, \theta) + R_\psi(T, \vartheta) \geq \frac{1}{16}\|\psi(\theta) - \psi(\vartheta)\|^2$.

Putting these two cases together proves the claim. □

Returning to the proof of Theorem 3, recall that $\mathbf{Y^n} = (Y_1, \ldots, Y_n)$ is an n-vector of independent but not iid random variables, i.e., $Y_i \sim P_{i,\theta}$, for $i = 1, \ldots, n$. Then the squared Hellinger distance between joint distributions P_θ^n and P_ϑ^n is given by

$$h^n(\theta, \vartheta) := H^2(P_\theta^n, P_\vartheta^n) = 2\Big[1 - \prod_{i=1}^n \Big\{1 - \frac{h_i(\theta, \vartheta)}{2}\Big\}\Big],$$

where $h_i(\theta, \vartheta) = H^2(P_{i,\theta}, P_{i,\vartheta})$ is the squared Hellinger distance between individual components. If θ and ϑ are sufficiently close, in the sense that $h_i(\theta, \vartheta) \leq 1$ for each $i = 1, \ldots, n$, then, given the following inequalities,

$$1 - x \leq -\log x \quad \text{and} \quad -\log(1 - x) \leq 2x, \quad x \in [0, 1/2],$$

it follows that

$$h^n(\theta,\vartheta) \le -2\sum_{i=1}^{n}\log\Big\{1-\frac{h_i(\theta,\vartheta)}{2}\Big\} \le 2\sum_{i=1}^{n} h_i(\theta,\vartheta). \tag{2.21}$$

According to our assumption about local expansion of the individual h_i's, if $\vartheta = \theta + \varepsilon\, u$ for a unit vector u, then

$$h^n(\theta,\theta+\varepsilon\, u) \le 2\mathcal{J}_n(\theta,u)\,\varepsilon^{\alpha} + o(n\varepsilon^{\alpha}), \quad \varepsilon \to 0.$$

When we take ε equal to $\varepsilon_{n,u} = \{3\mathcal{J}_n(\theta,u)\}^{-1/\alpha}$, then we get

$$h^n(\theta,\theta+\varepsilon_{n,u}\, u) \le \tfrac{2}{3} + o(1), \quad n \to \infty,$$

where the latter "$o(1)$" conclusion is justified by the assumption (Equation 2.9) about the rate of information accumulation. Therefore, for large enough n, with $\vartheta_{n,u} = \theta + \varepsilon_{n,u}\, u$, $h^n(\theta,\vartheta_{n,u}) \le \frac{3}{4}$, it follows from the above lemma that

$$R_\psi(T_n,\theta) + R_\psi(T_n,\vartheta_{n,u}) \ge \tfrac{1}{16}\|\psi(\theta) - \psi(\vartheta_{n,u})\|^2.$$

Since ψ is differentiable, there is a Taylor approximation at θ:

$$\psi(\theta) - \psi(\vartheta_{n,u}) = D_\psi(\theta)(\theta - \vartheta_{n,u}) + o(\|\theta - \vartheta_{n,u}\|),$$

where the latter little-oh means a q-vector whose entries are all of that magnitude. Plugging in the definition of $\vartheta_{n,u}$ gives

$$\psi(\theta) - \psi(\theta + \varepsilon_{n,u}\, u) = -\varepsilon_{n,u} D_\psi(\theta)\, u + o(\varepsilon_{n,u}), \quad n \to \infty,$$

and, hence,

$$\|\psi(\theta) - \psi(\theta + \varepsilon_{n,u}\, u)\|^2 = \varepsilon_{n,u}^2 \|D_\psi(\theta)\, u + o(1)\|^2 \geq \tfrac{1}{2}\varepsilon_{n,u}^2 \|D_\psi(\theta)\|^2.$$

Plugging in the definition of $\varepsilon_{n,u}$ establishes the first claim, (Equation 2.10), and the constant attached to the lower bound is $\frac{1}{32}3^{-2/\alpha}$. The second claim, (Equation 2.11), follows from the first and the general fact that, for a function f defined on a set A, $f(y_1) + f(y_2)$ is smaller than $2\sup_A f(y)$.

CHAPTER 3

OPTIMAL DESIGN FOR NONREGULAR REGRESSION BASED ON HELLINGER INFORMATION

Classically, optimal design of experiment under regularity conditions boils down to the optimization of Fisher information, which is inversely proportional to a lower risk bound of arbitrary estimator. However, this approach is only viable for regular models, as Fisher information only exists for regular models. Thus, there is not a systematic approach available to the problem of optimal design for nonregular models. The absence of Fisher information is perhaps the greatest, but by no means the only, obstacle; Chapter 1 discussed other issues in the optimal design of experiment, for nonregular models, including non-normal behavior of the usual "go-to" estimators, such as MLE, and the absence of variance-covariance structure for available estimators. Moreover, a variance-covariance structure is enough for inference only if the distribution is normal.

Chapter 2 presents a Hellinger information inequality from Theorem 3 that can be viewed as an extension of the Cramér-Rao bound to nonregular models, while the Hellinger information determines a lower bound of risk of arbitrary estimators. Chapter 2 lays out theoretical justification for using Hellinger information in the optimal design of experiment for nonregular models in a manner similar to Fisher information in regular cases. This approach to optimal design based on Hellinger information is applicable to a wide range of nonregular models

without requiring derivation of the asymptotic distribution of estimator on a case-by-case or model-by-model basis.

In this chapter, we put this approach to practice, specifically applying it to nonregular regression models whose Hellinger information was derived in Chapter 2, Section 2.2.3. Section 3.1 consists of a brief review of the nonregular regression model and the estimation method. Definition of optimal design for the nonregular regression model is presented in Section 3.2. Optimal design obtained by optimization of Hellinger information for a linear and a quadratic nonregular regression model are derived and presented in Theorem 3 and Theorem 4 in Section 3.3 and Section 3.4. Simulations comparing optimal designs to other designs are also presented.

3.1 Nonregular regression problem and method of estimation

According to (Smith, 1994), a class of nonregular regression models proceeds as follows: Let $\mathbf{x}_i = (x_{i1}, ..., x_{id})^\top$ be the vector of value of experimental variable of ith observation, and let $g(\theta; \mathbf{x}_i)$ be a differentiable function of $\theta \in \mathbb{R}^d$ in all dimensions, given $\mathbf{x}_i$. The probability distribution of the error term, $p(\varepsilon_i)$, then satisfies the following property: There is some constant $\alpha \in [1, 2), c > 0$ such that

$$y_i = g(\theta; \mathbf{x}_i) + \varepsilon_i, i = 1, ..., n, \theta \in \mathbb{R}^d, \text{ for some } \alpha \in [1, 2). \tag{3.1}$$

The error term ε_i follows distribution p_0 that fits the description in (Equation 2.15) from Chapter 2 Section 2.2. As discussed earlier in Chapter 1 the error distribution that fits the description above includes the three-parameter Weibull and gamma distribution, as well as

the generalized extreme value and generalized Pareto distributions. For the rest of this thesis, "nonregression model" would be the model referring to (Equation 3.1).

For nonregular regression model (Equation 3.1), where $g(\theta, x_i) = \theta_0 + \sum_{j=1}^{p} \theta_j x_{ij}$, (Smith, 1994) suggested an estimation method which requires the experimental variable $\mathbf{x}_i = (x_{i1}, ..., x_{id})^\top$, $i = 1, ..., n$ to be a zero-sum vector, i.e. $\sum_{i=1}^{n} x_{ij} = 0, j = 1, ..., p$. Centering the experimental variable to zero is used in such cases. According to (Koenker and Hallock, 2001), it is standard practice in quantile regression to center the covariates at the origin so that the intercept in the regression model can be interpreted as a conditional quantile function for some representative case, in order to avoid extrapolation of the model, since the case when all the covariates are at zero might not exist in reality. The estimator for θ would then be the solution to the linear programming problem of choosing $\theta_0, ..., \theta_p$ to satisfy the following:

$$\text{maximize } \theta_0 \text{ subject to } Y_i \geq \sum_{j=0}^{p} x_{ij}\theta_j \text{ for all i.} \tag{3.2}$$

Although there is an analytic expression for the distribution of this estimator, there is no closed-form expression for asymptotic variance of Smith's estimator, and the confidence interval has to be numerically determined. (Smith, 1994) argued that this method is superior to those that use the MLE. Besides the various problems with MLE mentioned in Chapter 1, Section 1.1.2, in reality it is often unknown what value α takes, especially when $\alpha < 1$. Smith's estimator can be employed for all $\alpha > 0$, so, it seems to be the main tool for nonregular regression.

Accordingly, it is used as the method of estimation in the simulation studies presented in the later part of this chapter.

3.2 Hellinger Information for nonregular regression

3.2.1 Hellinger Information for nonregular regression based on an approximate design

Chapter 2, Section 2.3 has shown that Hellinger information based on a sample of size n is proportional to (Equation 2.20). Accordingly, what follows is the definition of Hellinger information based on an approximate design.

Definition 9. Consider the nonregular regression model of (Equation 3.1), with any m point approximated design with the format of $\xi = \{(w_i, x_i), i =, ..., m\}$. Hellinger information based on ξ is defined as

$$\mathcal{J}_{\xi}(\theta) = \min_{u:\|u\|_2=1} \sum_{i=1}^{m} w_i \Big| \sum_{j=1}^{d} u_i \frac{\partial g(\theta, x_i)}{\partial \theta_j} \Big|^{\alpha}.$$

Definition 10. (Optimal design under Hellinger information inequality) A design ξ_{opt} is considered optimal for the nonregular regression model (Equation 3.1) if it minimizes the local minimax risk bound of the arbitrary estimator.

$$\xi_{opt} = \operatorname*{argmax}_{\xi} \mathcal{J}_{\xi}(\theta) = \operatorname*{argmax}_{\xi} \min_{u:\|u\|_2=1} \sum_{i=1}^{m} w_i \Big| \sum_{j=1}^{d} u_i \frac{\partial g(\theta, x_i)}{\partial \theta_j} \Big|^{\alpha}$$

When $\alpha \geq 2$ from (Equation 3.1), or for regular regression models where error follows Normal distribution, Hellinger information based on design ξ can be rewritten as

$$\mathcal{J}_\xi(\theta) = \frac{1}{4}\min_{u:\|u\|_2=1} u^\top \Big(\sum_{i=1}^{m} w_i \big(\frac{\partial g(\theta, x_i)}{\partial \theta}\big)\big(\frac{\partial g(\theta, x_i)}{\partial \theta}\big)^\top\Big) u = \frac{1}{4}\lambda_{min}\Big(\sum_{i=1}^{m} w_i \big(\frac{\partial g(\theta, x_i)}{\partial \theta}\big)\big(\frac{\partial g(\theta, x_i)}{\partial \theta}\big)^\top\Big).$$

Therefore, optimal design based on Hellinger information is equivalent to E-optimal design under regularity conditions.

The following two sections present some analytic results for optimal designs and some simulation results using Smith's estimator (Smith, 1994) for simple linear nonregular regression models and for the quadratic regression model. The requirement of centering the experimental variable at zero, as discussed in Section 3.1, implies that suitable designs for these models must belong to the collection of all balanced designs, which is denoted as Ξ,

$$\Xi = \big\{\{(w_i, x_i)...., \}, \sum w_i x_i = 0, x_i \in [-A, A]\big\}.$$

3.3 Nonregular regression example: linear model

3.3.1 Optimal design result on nonregular polynomial regression model

The Hellinger information for nonregular regression model (Equation 3.1) based on design ξ, according to Definition 10, would be

$$\mathcal{J}_\xi(\theta) = \min_{u:\|u\|_2=1} \sum_{i=1}^{m} w_i \Big| \sum_{j=1}^{d} u_i \frac{\partial g(\theta, x_i)}{\partial \theta_j} \Big|^\alpha.$$

Letting $u = (u_1, ..., u_{p+1})$ be a unit vector, the optimal design of a nonregular polynomial regression model, $g(\theta, x) = \theta_0 + \sum_{j=1}^{p} \theta_j x^j$, is

$$\xi_{opt} = \max_{\xi \in \Xi} \min_{u:\|u\|_2=1} \sum_{i=1}^{m} w_i |\sum_{j=0}^{p} u_{j+1} x_i^j|^\alpha \text{ with } \Xi = \big\{\{(w_i, x_i)...., \}, \sum w_i x_i = 0, x_i \in [-A, A]\big\}.$$

The following lemma contains a result about optimal design for nonregular polynomial regression models.

Lemma 2. *Denote* $\Xi^* = \big\{\{(w_1, -x_1), ..., (w_k, -x_m), (w_1, x_1), ..., (w_k, x_m)\}, \sum w_i = 0.5, x_i \in [0, A], m > 1\big\}$ *as a collection of symmetric designs that belongs to* Ξ*. The optimal design for any nonregular polynomial regression model* (Equation 3.1) *with* $g(\theta, \mathbf{x}_i) = \theta_0 + \sum_{j=1}^{p} \theta_j x_i^j$ *must therefore be a symmetric design, i.e.,*

$$\max_{\xi^* \in \Xi^*} \min_{u:\|u\|_2=1} J_\xi(\theta, u) = \max_{\xi \in \Xi} \min_{u:\|u\|_2=1} J_\xi(\theta, u).$$

Proof. see Appendix 3.6.1 □

Lemma 2 suggests that when searching for optimal design for polynomial nonregular regression models, one can restrict the search to symmetric designs.

3.3.2 Optimal design result on nonregular linear regression model

For simple linear nonregular regression models, we have the following result.

Lemma 3. *Denote* $\min_{u:\|u\|_2=1} J_{\{(0.5,\pm A)\}}(\theta, u)$ *as the Hellinger information of* θ *of model* (Equation 3.1) *with* $g(\theta, \mathbf{x}_i) = \theta_0 + \theta_1 x_i$, *based on design* $\{(0.5, -A), (0.5, A)\}$. *Then, for any symmetric design* $\xi^* \in \Xi^*$,

$$\min_{u:\|u\|_2=1} J_{(0.5,\pm A)}(\theta, u) \geq \max_{\xi^* \in \Xi^*} \min_{u:\|u\|_2=1} J_{\xi^*}(\theta, u).$$

Proof. see Appendix 3.6.2 □

By Lemma 2 and Lemma 3, we have the following theorem.

Theorem 4. *The optimal design for nonregular regression model* (Equation 3.1) *with* $g(\theta, x) = \theta_0 + \theta_1 x$, *is*

$$\xi_{opt} = \{(0.5, -A), (0.5, A)\}$$

Proof. see Appendix 3.6.3 □

3.3.3 Simulation

The optimal design based on Hellinger inequality in Theorem 3 is to minimize the lower bound of local minimax risk, which is defined as the sum of mean square error of estimator. The objective of the following simulation is to evaluate the performance of the optimal design compared with naive designs in terms of the sum of mean square error.

3.3.3.1 Simulation setup

The following outlines the simulation steps for each design. (In the next section, this setup will also be used for simulation studies involving quadratic models.)

- For given true value of θ, with shape parameter $\alpha \in [1, 2)$, and given the distribution of error term Gamma (α), generate n observations for approximate design $\xi = \{(w_i, x_i), i, = , ..., k\}$, i.e., generate nw_i for each x_i.
- Derive estimation of θ. For a linear model, estimation comes from R-package 'Rglpk' which is used for solving linear programing problems. For quadratic models, estimation comes from R-package "quantreg" for quantile regression of zero percentile. [1]
- Repeat each experiment ξ for $M = 1000$ times, denoted by $\hat{\theta}_j^m, m = 1, ...M$ for each $j = 0, ..., p + 1$.
- Compare the mean square error for each parameter for each design, ξ. For example, with linear model and design, ξ,

$$MSE(\theta_j) = \frac{\sum_{m=1}^{M}(\hat{\theta}_j^m - \theta_j)^2}{M}, j = 0, ..., 1 + p,$$

the risk is $\sum_i^{1+p} MSE(\theta_j)$.

3.3.3.2 Simulation results

Recall that the optimal design for two-parameter linear nonregular regression model is $\xi_{opt} = \{(-A, 0.5), (A, 0.5)\}$. In the following simulation study, the purpose is to compare optimal design to five-, ten-, and fifteen-point equidistant equal weight designs across the experimental variable space, $[-A, A]$, denoted as ξ_{5pt}, ξ_{10pt} and ξ_{15pt}.

---

[1]Note: "quantreg" and "Rglpk" both obtain the same estimation for linear model, but "Rglpk" does not seem to be able to produce estimation for quadratic models.

For the rest of Chapter 3, all examples have the same sample size of 120 for each run of experiment, and the result of mean square error came from 1000 repeats of each experiment. The number in parentheses is the standard error of mean square error.

Based on the results regarding Hellinger information presented in Chapter 2, the optimal design result for two-parameter linear nonregular regression models shall be the same regardless of the specific distribution of error distribution and the boundary of the experimental space, A. That is, one expects to see the optimal design perform better than other designs in simulations, regardless of the choice of the error distribution, the choice value of scalar parameter (if the distribution is gamma or Weibull), the value of shape parameter (α), or the value of A. The three following examples confirm this optimal design result.

Example 17 explores simulation settings with different error distributions with the same true parameter value, while Example 18 checks if choosing a different true parameter value would change the conclusion. Example 19 presents simulation results from settings with different values of A, α with gamma distributed error.

Example 17. Table 1, Table 2 and Table 3 contain simulation results from gamma distributed error with shape parameter value 1.2 and with scale parameter set at 10, 1, and 0.5, respectively. It appears that MSE across parameters increases as scale parameter increases. With scale parameter larger than 1, the mean square error level is higher, whereas, while the scale parameter decreases, the mean square error level is reduced. However, the ranking of MSE for each parameter and the sum of MSE among designs remain the same. Table 4, Table 5, and Table 6 contain results from a simulation with Weillbull distributed error and shape parameter

value of 1.2, and with scale parameters of 10, 1, and 0.5, respectively. One can draw the same conclusion from Table 1 to Table 3. For all the tables below, among all designs' sum of MSE, the one from optimal design is two standard error away from the next smallest one. Although under different settings, the MSE and sum of MSE differ among the designs, the relative ranking between designs for each setting remains mostly the same.

This example confirms that the optimal design does not depend on the choice of scale parameter or the choice of distribution of error, as long as the error distribution fits the description from Proposition 2. Thus, for the rest of the simulation, without loss of generality, only gamma distributed error will be used (when $\alpha = 1$, the distribution is exponential).

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.15649 (0.00562)	0.04026 (0.00207)	0.19674 (0.00599)	ξ_{opt}
0.12490 (0.00525)	0.10168 (0.00501)	0.22659 (0.00726)	ξ_{5pt}
0.13218 (0.00533)	0.14636 (0.01198)	0.27855 (0.01311)	ξ_{10pt}
0.13439 (0.00659)	0.14481 (0.00988)	0.27920 (0.01188)	ξ_{15pt}

TABLE 1: $A = 1, \alpha = 1, \theta_0 = 2, \theta_1 = 0.5, scale = 10$

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.00161 (0.00006)	0.00041 (0.00003)	0.00202 (0.00007)	ξ_{opt}
0.00131 (0.00006)	0.00202 (0.00016)	0.00333 (0.00017)	ξ_{5pt}
0.00140 (0.00006)	0.00212 (0.00017)	0.00352 (0.00018)	ξ_{10pt}
0.00140 (0.00006)	0.00246 (0.00019)	0.00386 (0.00020)	ξ_{15pt}

TABLE 2: $A = 1, \alpha = 1, \theta_0 = 2, \theta_1 = 0.5, scale = 1$

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.00041 (0.00002)	0.00011 (0.00001)	0.00052 (0.00002)	ξ_{opt}
0.00032 (0.00001)	0.00048 (0.00003)	0.00081 (0.00004)	ξ_{5pt}
0.00037 (0.00001)	0.00046 (0.00003)	0.00083 (0.00004)	ξ_{10pt}
0.00037 (0.00002)	0.00058 (0.00004)	0.00095 (0.00005)	ξ_{15pt}

TABLE 3: $A = 1, \alpha = 1, \theta_0 = 2, \theta_1 = 0.5, scale = 0.5$

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.12340 (0.00471)	0.03116 (0.00171)	0.15456 (0.00501)	ξ_{opt}
0.11565 (0.00458)	0.08444 (0.00346)	0.20010 (0.00574)	ξ_{5pt}
0.11716 (0.00477)	0.12004 (0.00860)	0.23720 (0.00984)	ξ_{10pt}
0.10780 (0.00388)	0.11838 (0.00766)	0.22618 (0.00859)	ξ_{15pt}

TABLE 4: $A = 1, \alpha = 1, \theta_0 = 2, \theta_1 = 0.5, scale = 10$, Weibull Distributed

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.00130 (0.00005)	0.00035 (0.00002)	0.00165(0.00005)	ξ_{opt}
0.00112 (0.00005)	0.00173 (0.00014)	0.00284 (0.00014)	ξ_{5pt}
0.00116 (0.00005)	0.00153 (0.00012)	0.00268 (0.00012)	ξ_{10pt}
0.00114 (0.00004)	0.00166 (0.00013)	0.00280 (0.000014)	ξ_{15pt}

TABLE 5: $A = 1, \alpha = 1, \theta_0 = 2, \theta_1 = 0.5, scale = 1$, Weibull Distributed

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.00036 (0.00002)	0.00009 (0.00001)	0.00046(0.00002)	ξ_{opt}
0.00029 (0.00001)	0.00043 (0.00003)	0.00072 (0.00004)	ξ_{5pt}
0.00030 (0.00001)	0.00039 (0.00002)	0.00069 (0.00003)	ξ_{10pt}
0.00030 (0.00001)	0.00046 (0.00003)	0.00077 (0.00003)	ξ_{15pt}

TABLE 6: $A = 1, \alpha = 1, \theta_0 = 2, \theta_1 = 0.5, scale = 0.5$, Weibull Distributed

Example 18. This example is meant to check if the true parameter value matters to the performance of design in simulation. Table 7 and Table 8 contains results from two runs of simulations with $A = 1$, exponential distribution error ($\alpha = 1$), and with different true parameter values.

Table 7 has true parameter value of $\theta_0 = 100, \theta_1 = 50$, whereas Table 8 has true parameter value of $\theta_0 = 2, \theta_1 = 0.5$. One can see that although the true values are quite different from each other, the level of mean square error of each parameter, the level of standard deviation of mean square error and the sum of mean square error are close between these two tables. Thus, for the rest of the simulations examples, without loss of generality, all the true value of the parameters will be the same.

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.00045 (0.00002)	0.00015 (0.00001)	0.00060 (0.00002)	ξ_{opt}
0.00035 (0.00002)	0.00064 (0.00005)	0.00099 (0.00006)	ξ_{5pt}
0.00039 (0.00002)	0.00068 (0.00005)	0.00107 (0.00006)	ξ_{10pt}
0.00037 (0.00002)	0.00068 (0.00005)	0.00106 (0.00006)	ξ_{15pt}

TABLE 7: $A = 1, \alpha = 1, \theta_0 = 100, \theta_1 = 50$

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.00045 (0.00002)	0.00015 (0.00001)	0.00061 (0.00003)	ξ_{opt}
0.00034 (0.00002)	0.00068 (0.00006)	0.00102 (0.00006)	ξ_{5pt}
0.00040 (0.00002)	0.00065 (0.00006)	0.00105 (0.00006)	ξ_{10pt}
0.00040 (0.00002)	0.00074 (0.00006)	0.00114 (0.00007)	ξ_{15pt}

TABLE 8: $A = 1, \alpha = 1, \theta_0 = 2, \theta_1 = 0.5$

Comparing the results from the previous simulations with different true parameter values and different error distribution settings, it turns out that the ranking of sum of mean squre error between designs is the same, and the optimal design is always significatly better than others. Based on this, the simulations for the next example would have the same true parameter value and error distribution.

Example 19. To see how optimal design performs for different combinations of A, α values, the following two tables provide some comparisons.

Setting true value as $\theta_0 = 6, \theta_1 = 0.5$, with error following exponential distribution, Table 9 records sum of mean square error from three different runs of simulations with A set to 1, 2, and 5 respectively.

Design	$A = 1, \alpha = 1$	$A = 2, \alpha = 1$	$A = 5, \alpha = 1$
ξ_{opt}	0.00061 (0.00002)	0.00040 (0.00002)	0.00039 (0.000018)
ξ_{5pt}	0.00105 (0.00006)	0.00053 (0.00002)	0.00041 (0.000022)
ξ_{10pt}	0.00102 (0.00006)	0.00049 (0.00002)	0.00038 (0.000018)
ξ_{15pt}	0.00110 (0.00006)	0.00060 (0.00002)	0.00042 (0.000020)

TABLE 9: $\theta_0 = 6, \theta_1 = 0.5, \alpha = 1$

Setting true value as $\theta_0 = 6, \theta_1 = 0.5$, with error following gamma distribution with shape parameter of $\alpha = 1.4$, Table 10 records sum of mean square error from three different runs of simulations with A set to 1, 2 and 5 respectively.

Design	$A = 1, \alpha = 1.4$	$A = 2, \alpha = 1.4$	$A = 5, \alpha = 1.4$
ξ_{opt}	0.00532 (0.00015)	0.00426 (0.00013)	0.00448 (0.000155)
ξ_{5pt}	0.00806 (0.00034)	0.00495 (0.00017)	0.00409 (0.000149)
ξ_{10pt}	0.00825 (0.00031)	0.00513 (0.00016)	0.00423 (0.000146)
ξ_{15pt}	0.00936 (0.00039)	0.00500 (0.00016)	0.00457 (0.000154)

TABLE 10: $\theta_0 = 6, \theta_1 = 0.5, \alpha = 1.4$

From other runs of simulation (not presented here), increasing α value while holding other settings the same would increase sum of mean square error level for each simulation, and the level of standard deviation.

When $A = 1$, $A = 2$, under both $\alpha = 1$ and $\alpha = 1.4$ cases, optimal design is superior, since sum of mean square error from other designs are all larger than two standard deviation away from that of ξ_{opt}.

From the result for $A = 5$, it appears that the design with the smallest sum of mean square error is not the optimal design; however, the Welch modified two-sample t-test result[1] shows that the difference is not significant. For $\alpha = 1$ case, notice that $Sum(MSE)(\xi_{opt}) = 0.000396$, which is larger than that from $Sum(MSE)(\xi_{10pt}) = 0.000377$; however, the difference is not significant. The t-score is $\frac{0.000396-0.000377}{\sqrt{2*0.000018^2}} = 0.7463905$, degree of freedom is 1998, and the p-value for two-sided test is 0.4555. For $\alpha = 1.4$ case, $Sum(MSE)(\xi_{opt}) = 0.004484$, which is larger than that from $Sum(MSE)(\xi_{opt})\xi_{5pt} = 0.004087$, though, again, the difference is not significant. The t-score is $\frac{0.004484-.004087}{\sqrt{0.000155^2+0.000149^2}} = 1.846492$, and the p-value for two-sided test is 0.06497.

Table 11 to Table 13 provide greater detail of Table 9. Likewise, Table 14 to Table 16 provide greater detail of Table 10. What follows is a list of several observations from these tables:

---

[1]The result is calculated in R with function tsum.test

- The optimal design for each combination of A, α values is significantly superior to other designs in estimation of θ_1, while not always significantly inferior in estimation of θ_0. To see this, from Table 11 to Table 16, ξ_{opt}'s $MSE(\theta_0)$ value is either the largest or the second largest; however, among these 6 simulations, not every ξ_{opt}'s $MSE(\theta_0)$ value is significantly larger than the smallest $MSE(\theta_0)$ value. For example, in Table 15, $MSE(\theta_0)(\xi_{opt}) = 0.00404$ and $MSE(\theta_0)(\xi_{15pt}) = 0.00376$ (the smallest) are not significantly different according to the t-test. On the estimation of θ_1, for each of Table 11 to Table 16, ξ_{opt} has the smallest $MSE(\theta_1)$ and is significantly smaller than the second smallest $MSE(\theta_1)$ from other designs.

- On observing the effect of different A values, it appears that when A is small, the MSE for θ_0 and θ_1 are at the same scale, but when A becomes large, the scale of MSE for different parameters becomes drastically different.

  From Table 11 and Table 12, among $MSE(\theta_0)$, and $MSE(\theta_1)$ values, almost all of them have the first non-zero digit at the 4th decimal place. (Except for optimal design for $A = 2$ from Table 12: the $MSE(\theta_1)$ first non-zero digit is at the 5th decimal place.)

  For $A = 5$, using Table 13 ($\alpha = 1$) as an example, one can observe that $MSE(\theta_0)$ for the designs considered are still around 0.0003 and 0.00004, which are at the same level from Table 11 ($A = 1$) and Table 12 ($A = 2$), while $MSE(\theta_1)$ reduces drastically in comparison to $MSE(\theta_0)$: the first non-zero digit in Table 13 is at the 5th digit. This means that, for large A, $MSE(\theta_1)$ only contributed about a hundredth of value from $MSE(\theta_0)$ in calculation of $Sum(MSE)$. Thus, the ranking of $Sum(MSE)$ among designs for large A

is identical to the ranking of $MSE(\theta_0)$, since the contribution of $MSE(\theta_1)$ to the value of $Sum(MSE)$ is negligible. These same observations can be seen for $\alpha = 1.4$ cases from Table 14 to Table 16

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.00045 (0.00002)	0.00016 (0.00001)	0.00061 (0.00002)	ξ_{opt}
0.00034 (0.00002)	0.00071 (0.00006)	0.00105 (0.00006)	ξ_{5pt}
0.00036 (0.00002)	0.00065 (0.00006)	0.00102 (0.00006)	ξ_{10pt}
0.00034 (0.00002)	0.00076 (0.00006)	0.00110 (0.00006)	ξ_{15pt}

TABLE 11: $A = 1, \alpha = 1$

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.000373 (0.000016)	0.000031 (0.000020)	0.000404 (0.000016)	ξ_{opt}
0.000364 (0.000020)	0.000165 (0.000015)	0.000529 (0.000025)	ξ_{5pt}
0.000360 (0.000018)	0.000131 (0.000010)	0.000492 (0.000021)	ξ_{10pt}
0.000408 (0.000020)	0.000192 (0.000014)	0.000600 (0.000024)	ξ_{15pt}

TABLE 12: $A = 2, \alpha = 1$

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.000391 (0.000018)	0.000005 (0.000000)	0.000396 (0.000018)	ξ_{opt}
0.000378 (0.000021)	0.000029 (0.000002)	0.000407 (0.000022)	ξ_{5pt}
0.000350 (0.000018)	0.000027 (0.000002)	0.000377 (0.000018)	ξ_{10pt}
0.000386 (0.000020)	0.000032 (0.000003)	0.000418 (0.000020)	ξ_{15pt}

TABLE 13: $A = 5, \alpha = 1$

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.00443 (0.00015)	0.00089 (0.00005)	0.00532 (0.00015)	ξ_{opt}
0.00388 (0.00014)	0.00418 (0.00031)	0.00806 (0.00034)	ξ_{5pt}
0.00377 (0.00013)	0.00448 (0.00028)	0.00825 (0.00031)	ξ_{10pt}
0.00401 (0.00013)	0.00534 (0.00037)	0.00936 (0.00039)	ξ_{15pt}

TABLE 14: $A = 1, \alpha = 1.4$

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.00404 (0.00013)	0.00022 (0.00005)	0.00426 (0.00013)	ξ_{opt}
0.00398 (0.00014)	0.00097 (0.00010)	0.00495 (0.00017)	ξ_{5pt}
0.00408 (0.00014)	0.00105 (0.00007)	0.00513 (0.00016)	ξ_{10pt}
0.00376 (0.00013)	0.00124 (0.00010)	0.00500 (0.00016)	ξ_{15pt}

TABLE 15: $A = 2, \alpha = 1.4$

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.004449 (0.000155)	0.000035 (0.000002)	0.004484 (0.000155)	ξ_{opt}
0.003888 (0.000148)	0.000200 (0.000016)	0.004087 (0.000149)	ξ_{5pt}
0.004063 (0.000146)	0.000166 (0.000016)	0.004229 (0.000146)	ξ_{10pt}
0.004364 (0.000154)	0.000201 (0.000013)	0.004565 (0.000154)	ξ_{15pt}

TABLE 16: $A = 5, \alpha = 1.4$

3.4 Nonregular regression example: quadratic model

3.4.1 Optimal design results

Consider the nonregular quadratic regression model, i.e., $g(\theta, x) = \theta_0 + \theta_1 x_i + \theta_2 x_i^2$ in (Equation 3.1). Based on Definition 10, optimal design for the nonregular quadratic model is

$$\xi_{opt} = \max_{\xi \in \Xi} \min_{u:\|u\|_2=1} \sum_{i=1}^{n} w_i |u_1 + u_2 x_i + u_3 x_i^2|^{\alpha}.$$

The following theorem provides a result for the $\alpha = 1$ case.

Theorem 5. *When $\alpha = 1$, optimal design for nonregular quadratic regression model belongs to the collection of all three-point symmetric designs centered at zero and with two points at boundary, denoted as Ξ_{3pt},*

$$\Xi_{3pt} = \big\{\xi_w = \{(\frac{1-w}{2}, -A), (w, 0), (\frac{1-w}{2}, A)\}, w \in (0, 1)\big\}. \tag{3.3}$$

Proof. see Appendix 3.6.4 □

Theorem 5 means that for the $\alpha = 1$ case, Hellinger information in the direction of u based on design $\xi_w \in \Xi_{3pt}$ has the expression of

$$J_{\xi_w}(\theta, u) = w|u_1| + \frac{1-w}{2}(|u_1 + u_2 A + u_3 A^2| + |u_1 - u_2 A + u_3 A^2|),$$

and optimal design would be

$$J_{\xi_w^*} = \underset{w\in[0,1]}{\operatorname{argmax}} \min_{u:\|u\|_2=1} J_{\xi_w}(\theta, u).$$

Unfortunately, the analytic solution to the optimal weight at point zero from (Equation 3.3),$w^* = \underset{w\in[0,1]}{\operatorname{argmax}} J_w(\theta, u)$, is not available. However, one can use a numerical search method to find the approximated result.

The result in Theorem 5 is only for $\alpha = 1$, although it is quite likely that the result from Theorem 5 can be extended to all nonregular quadratic models with regularity index value between 1 and 2. Simulation results presented later in Section 3.4.3 support this hypothesis.

The following steps outline the numerical search for the optimal weight at point zero, i.e., searching numerically for

$$w^* = \underset{w\in[0,1]}{\operatorname{argmax}} \min_{u:\|u\|_2=1} J_{\xi_w}(\theta, u). \tag{3.4}$$

To accomplish this, for a given w, search for minimum value of

$$J_{\xi_w}(\theta, u) = w|u_1|^\alpha + \frac{1-w}{2}(|u_1 + u_2 A + u_3 A^2|^\alpha + |u_1 - u_2 A + u_3 A^2|^\alpha), u = (u_1, u_2, u_3), |u| = 1,$$

over unit circle; repeat this for $w \in [0, 1]$ over mesh size of 0.05, then choose the w value that obtains the maximum $\min_{u:\|u\|_2=1} J_{\xi_w}(\theta, u)$. The following describes the steps of a grid search over the unit circle.

First, notice that for any unit vector u from $(+,-,+),(-,-,-)$, and $(-,+,-)$, one can find a vector from $(+,+,+)$ such that they would have the same $J_{\xi_w}(\theta,u)$ value. Similarly, for any unit vector from $(+,-,-),(-,-,+)$, and $(-,+,+)$, one can find a unit vector from $(+,+,-)$ such that $J_{\xi_w}(\theta,u)$ would have the same value. Thus, one only needs to search over the first octant $(+,+,+)$ and fourth octant $(+,+,-)$. The method of generating these unit-vectors is based on the spherical coordinates of unit sphere (u_1,u_2,u_3) with θ as the azimuthal coordinate, and φ as the polar coordinate. Accordingly, any unit-vectors would have the expression of

$$u = \big(\sin(\varphi)\cos(\theta), \sin(\varphi)\sin(\theta), \cos(\varphi)\big), 0 \leq \theta \leq 2\pi, 0 \leq \varphi \leq \pi.$$

Points from the first octant $(+,+,+)$ are

$$u_{(i,j)} = \big(\sin(\varphi_i)\cos(\theta_j), \sin(\varphi_i)\sin(\theta_j), \cos(\varphi_i)\big) \quad \varphi_i, \theta_j = 0, 0.01, ..., 1.57 \approx \frac{\pi}{2}.$$

With the mesh size set for φ, θ above, there are $157^2 = 24649$ many $u_{(i,j)}$ from $(+,+,+)$. To obtain points from fourth octant $(+,+,-)$, one can simply change the sign of all u_3 to the negative of these 24649 $u_{(i,j)}$ of the first octant, i.e., the points from $(+,+,-)$ would be

$$u_{(i,j)} = \big(\sin(\varphi_i)\cos(\theta_j), \sin(\varphi_i)\sin(\theta_j), -\cos(\varphi_i)\big) \quad \varphi_i, \theta_j = 0, 0.01, ..., 1.57.$$

This method generates $24649 \times 2 = 49928$ many unit vectors to search over. The numerical solution to (Equation 3.4) for $\alpha = 1$ case is presented in Table 17.

A	α	$\underset{w\in[0,1]}{\text{argmax}}\ \underset{u:\|u\|_2=1}{\text{argmin}}\ J_w(\theta,u)$	$u_{(min)}$
1	1	0.5	(0.70328, 0, -0.71091)
2	1	0.65	(0.00071, 0.8957, 0.44466)
5	1	0.8	(0.99917, 0, -0.04079)
10	1	0.9	(0.99994, 0, -0.0108)

TABLE 17: Numerical search result: w^* from (Equation 3.4) for $\alpha = 1$ case

At this point, we do not have the proof that Theorem 5 can be extended to $\alpha \in (1, 2)$. However, some simulation studies in the later section show that it is likely that the extension is true. Table 18 provides the results for the numerical solution to (Equation 3.4) for different α values.

	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 1.8$	$\alpha = 1.9$
A=1	0.5	0.6	0.6	0.6
A=2	0.65	0.75	0.75	0.8
A=5	0.8	0.9	0.95	0.95
A=10	0.9	0.95	0.95	0.95

TABLE 18: Numerical search result: w^* from (Equation 3.4) for different A, α

3.4.2 Simulation results for quadratic regression model when $\alpha = 1$

Following the same simulation steps described in Section 3.3.3.1, the simulation results in this section consider the optimal designs listed in Table 19 and 5-, 10-, and 15-point uniform designs for comparison. For all the simulations in this section, sample size is set to $N = 80$, and to obtain mean square errors, each experiment is repeated $M = 1000$ times, with true parameters' values set at $\theta = (2, 4, 0.8)$.

A	Optimal Design $\xi_{Optimal}$
1	$\xi_{0.5} = \{(-A, 0.25), (0, 0.5), (A, 0.25)\}$
2	$\xi_{0.65} = \{(-A, 0.175), (0, 0.65), (A, 0.175)\}$
5	$\xi_{0.8} = \{(-A, 0.1), (0, 0.8), (A, 0.1)\}$
10	$\xi_{0.9} = \{(-A, 0.05), (0, 0.9), (A, 0.05)\}$

TABLE 19: List of optimal designs for different A under $\alpha = 1$ case

Example 20. To compare optimal designs to other designs for different values of A, in Table 20, each row records the sum of mean square error and its standard error for different designs. Row of $\xi_{Optimal}$ records the results from optimal designs from Table 19. One can observe that across different A values, optimal design performs the best since one can observe that their

$Sum(MSE)$ is two standard deviations away from the other designs'. Table 21 to Table 23 contain the more detailed results from each of the settings from Table 20.

Design	$A = 1, \xi_{0.5}$	$A = 2, \xi_{0.65}$	$A = 5, \xi_{0.8}$
$\xi_{Optimal}$	0.00229 (0.00009)	0.00071 (0.00003)	0.00039 (0.00003)
ξ_{5pt}	0.00484 (0.00020)	0.00176 (0.00011)	0.00107 (0.00007)
ξ_{10pt}	0.00486 (0.00022)	0.00110 (0.00007)	0.00084 (0.00006)
ξ_{15pt}	0.00567 (0.00028)	0.00122 (0.00006)	0.00084 (0.00006)

TABLE 20: Simulation Result: Designs For Nonregular Quadratic Regression

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{MSE(\theta_2)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.00054 (0.00004)	0.00059 (0.00004)	0.00116 (0.00007)	0.00229 (0.00009)	$\xi_{0.5}$
0.00089 (0.00006)	0.00083 (0.00005)	0.00311 (0.00019)	0.00484 (0.00020)	ξ_{5pt}
0.00080 (0.00005)	0.00076 (0.00005)	0.00331 (0.00020)	0.00486 (0.00022)	ξ_{10pt}
0.00079 (0.00005)	0.00091 (0.00007)	0.00397 (0.00027)	0.00567 (0.00028)	ξ_{15pt}

TABLE 21: $A = 1, \alpha = 1$

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{MSE(\theta_2)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.00029 (0.00002)	0.00028 (0.00002)	0.00015 (0.00001)	0.00071 (0.00003)	$\xi_{0.65}$
0.00121 (0.00011)	0.00028 (0.00002)	0.00027 (0.00003)	0.00176 (0.00011)	ξ_{5pt}
0.00075 (0.00006)	0.00017 (0.00001)	0.00019 (0.00001)	0.00110 (0.00007)	ξ_{10pt}
0.00080 (0.00006)	0.00019 (0.00001)	0.00022 (0.00001)	0.00122 (0.00006)	ξ_{15pt}

TABLE 22: $A = 2, \alpha = 1$

$\boldsymbol{MSE(\theta_0)}$	$\boldsymbol{MSE(\theta_1)}$	$\boldsymbol{MSE(\theta_2)}$	$\boldsymbol{Sum(MSE)}$	**Design**
0.00022 (0.00002)	0.00015 (0.00005)	1e-05 (0)	0.00039 (0.00002)	$\xi_{0.8}$
0.00103 (0.00007)	0.00003 (0)	1e-05 (0)	0.00107 (0.00007)	ξ_{5pt}
0.00080 (0.00006)	0.00003 (0)	1e-05 (0)	0.00084 (0.00006)	ξ_{10pt}
0.00078 (0.00006)	0.00003 (0)	1e-05 (0)	0.00082 (0.00006)	ξ_{15pt}

TABLE 23: $A = 5, \alpha = 1$

3.4.3 Some results for $\alpha \in (1, 2)$ case

Table 24 lists the likely but unproven optimal designs for nonregular quadratic regression models with different values of A, α. These are obtained by numerical search, following the

steps from Section 3.4.1. A simulation is conducted to compare designs from Table 24 and 5-, 10-, and 15-point uniform designs. The sum of mean square errors from the simulation and their standard error are recorded in Table 25. The simulation settings are the same as the ones for $\alpha = 1$ case, i.e. $N = 80, M = 1000, \theta = (2, 4, 0.8)$.

A	α	Likely candidates for the optimal design: $\xi^{(A,\alpha)}$
1	1.5	$\xi_{0.6} = \{(-A, 0.15), (0, 0.6), (A, 0.15)\}$
2	1.5	$\xi_{0.75} = \{(-A, \frac{0.25}{2}), (0, 0.75), (A, \frac{0.25}{2})\}$
2	1.9	$\xi_{0.8} = \{(-A, 0.1), (0, 0.8), (A, 0.1)\}$
5	1.8	$\xi_{0.95} = \{(-A, 0.025), (0, 0.95), (A, 0.025)\}$

TABLE 24: Likely but unproven optimal designs for different A,α cases

Design	$A = 1, \alpha = 1.5$	$A = 2, \alpha = 1.5$	$A = 2, \alpha = 1.9$	$A = 5, \alpha = 1.8$
$\xi^{(A,\alpha)}$	0.02213 (0.00062)	0.00922 (0.00028)	0.03091 (0.00073)	0.01852 (0.00053)
ξ_{5pt}	0.03836 (0.00129)	0.01411 (0.00053)	0.04027 (0.00132)	0.02831 (0.00114)
ξ_{10pt}	0.03749 (0.00142)	0.01206 (0.00045)	0.04397 (0.00143)	0.02834 (0.00118)
ξ_{15pt}	0.04092 (0.00148)	0.01229 (0.00044)	0.04001 (0.00129)	0.02593 (0.00115)

TABLE 25: Simulation result for design from Table 24

The first row of Table 25 contains the simulation results from likely candidates for optimal design from Table 24. Across different cases, the likely candidates for the optimal designs listed in Table 24 perform the best since their $Sum(MSE)$ are two standard deviations away from other designs', so the simulation results support the hypothesis that the optimal designs for the quadratic nonregular regression model with $\alpha \in (1, 2)$ are likely to belong to $\Xi_{3pt} = \{(\frac{1-w}{2}, -A), (w, 0), (\frac{1-w}{2}, A)\}$.

3.5 Summary of simulation results

By observing simulation results from the previous two sections, we see that for quadratic models, optimal design based on Hellinger information outperforms other designs. For linear models, when the boundary of experimental design variable space A is small, optimal design based on Hellinger information does outperform other designs in terms of sum of mean square error. However, as A becomes large, the optimal design is among the best designs, but not the one with the smallest sum of mean square error; however, in such cases, the risk from estimation from optimal design is not statistically different.

Optimal design of nonregular models is obtained from optimization of Hellinger information, which serves as a lower bound of a risk bound for arbitrary estimators. The simulation thus confirms Theorem 3.

Given the absence of efficient estimators for nonregular regression models and the lack of closed-form expression of the asymptotic variance-covariance structure of Smith's estimator (Smith, 1994), this approach to optimal design based on Hellinger information is the best available.

3.6 Appendix

3.6.1 Proof of Lemma 2

For a given design $\xi = \{(w_i, x_i), i = 1, ...k\}$, let $\bar{u} = \underset{u:\|u\|_2=1}{\operatorname{argmin}} \sum_{i=1}^{k} w_i |\sum_{j=0}^{p+1} x_i^j u_{j+1}|^\alpha$, then notice that for any j, $u_{j+1}(x)^j = (-1)^j u_{j+1}(-x)^j$.

Therefore,

$$\sum_i^k w_i |\sum_{j=0}^{p+1} x_i^j \bar{u}_{j+1}|^\alpha = \sum_i^k w_i |\sum_{j=0}^{p+1} (-x_i)^j (-1)^j \bar{u}_{j+1}|^\alpha \geq \min_{u:\|u\|_2=1} \sum_i^k w_i |\sum_{j=0}^{p+1} (-x_i)^j u_{j+1}|^\alpha. \quad (3.5)$$

Similarly, let $u^* = \underset{u:\|u\|_2=1}{\operatorname{argmin}} \sum_{i=1}^{k} w_i |\sum_{j=0}^{p+1} (-x_i)^j u_{j+1}|^\alpha$, then

$$\sum_i^k w_i |\sum_{j=0}^{p+1} (-x_i)^j u_{j+1}^*|^\alpha = \sum_i^k w_i |\sum_{j=0}^{p+1} x_i^j [(-1)^j u_{j+1}^*]|^\alpha \geq \min_{u:\|u\|_2=1} \sum_{i=1}^k w_i |\sum_{j=0}^{p+1} x_i^j u_{j+1}|^\alpha. \quad (3.6)$$

Therefore, by the definition of u^* and $\bar{u}$, combining (Equation 3.5) and (Equation 3.6),

$$\min_{u:\|u\|_2=1} \sum_i^k w_i |\sum_{j=0}^{p+1} x_i^j u_{j+1}|^\alpha = \min_{u:\|u\|_2=1} \sum_i^k w_i |\sum_{j=0}^{p+1} (-x_i)^j u_{j+1}|^\alpha. \quad (3.7)$$

Given design $\xi \in \Xi$, one can form a symmetric design, ξ^*, by an equal mixture of design ξ and $\xi^- = \{(w_i, -x_i), i = 1, ...k\}$. Let $\xi^* = \frac{1}{2}\xi + \frac{1}{2}\xi^-$, then

$$\begin{aligned}
\min_{u:\|u\|_2=1} J_{\xi^*}(\theta, u) &= \min_{u:\|u\|_2=1} (\frac{1}{2} \sum_i^k w_i |\sum_{j=0}^{p+1} x_i^j u_{j+1}|^\alpha + \frac{1}{2} \sum_i^k w_i |\sum_{j=0}^{p+1} (-x_i)^j u_{j+1}|^\alpha) \\
&\geq \min_{u:\|u\|_2=1} (\frac{1}{2} \sum_i^k w_i |\sum_{j=0}^{p+1} x_i^j u_{j+1}|^\alpha) + \min_{u:\|u\|_2=1} (\frac{1}{2} \sum_i^k w_i |\sum_{j=0}^{p+1} (-x_i)^j u_{j+1}|^\alpha).
\end{aligned}$$

Based on (Equation 3.7), the above inequality can be written as

$$\min_{u:\|u\|_2=1} J_{\xi^*}(\theta,u) > \min_{u:\|u\|_2=1} \sum_{i}^{k} w_i |\sum_{j=0}^{p+1} x_i^j u_{j+1}|^\alpha.$$

This means that for any design $\xi \in \Xi$, there is a symmetric design $\xi^* \in \Xi^*$ such that

$$\min_{u:\|u\|_2=1} J_{\xi^*}(\theta,u) \geq \min_{u:\|u\|_2=1} J_{\xi}(\theta,u),$$

which means,

$$\underset{\xi^*\in\Xi^*}{max} \min_{u:\|u\|_2=1} J_{\xi^*}(\theta,u) \geq \underset{\xi\in\Xi}{max} \min_{u:\|u\|_2=1} J_{\xi}(\theta,u),$$

and since $\Xi^* \subset \Xi$,

$$\underset{\xi^*\in\Xi^*}{max} \min_{u:\|u\|_2=1} J_{\xi^*}(\theta,u) = \underset{\xi\in\Xi}{max} \min_{u:\|u\|_2=1} J_{\xi}(\theta,u).$$

3.6.2 Proof of Lemma 3

Based on model (Equation 3.1) with $g(\theta,x) = \theta_0 + \theta_1 x$,

$$J_{\{(0.5,\pm A)\}}(\theta,u) = 0.5(|u_1 + u_2 A|^\alpha + |u_1 - u_2 A|^\alpha), \text{ and}$$

$$J_{\xi^*}(\theta,u) = \sum_{1}^{m} w_i(|u_1 + u_2 x_i|^\alpha + |u_1 - u_2 x_i|^\alpha).$$

Any unit vector (except $u = (1, 0)$)[1] can be written in the following format:

$$u = \pm(\pm\frac{B}{\sqrt{1+B^2}}, \frac{1}{\sqrt{1+B^2}}), B \in [0, \infty). \tag{3.8}$$

Notice that no matter what choices of sign combination of u_1, u_2 is given,

$$|u_1 + u_2 x_i|^\alpha + |u_1 - u_2 x_i|^\alpha = (1+B^2)^{-0.5\,\alpha}(|B + x_i|^\alpha + |B - x_i|^\alpha).$$

First we check that for any given unit-vector u, $J_{\{(0.5,\pm A)\}}(\theta, u) - J_{\xi^*}(\theta, u)$ is non-negative. Recall that $\sum_1^m w_i = 0.5$,

$$\begin{aligned}
&J_{\{(0.5,\pm A)\}}(\theta, u) - J_{\xi^*}(\theta, u) \\
=&0.5(|u_1 + u_2 A|^\alpha + |u_1 - u_2 A|^\alpha) - \sum_1^m w_i(|u_1 + u_2 x_i|^\alpha + |u_1 - u_2 x_i|^\alpha) \\
=&\sum_1^m w_i(|u_1 + u_2 A|^\alpha + |u_1 - u_2 A|^\alpha - |u_1 + u_2 x_i|^\alpha - |u_1 - u_2 x_i|^\alpha) \\
=&(1+B^2)^{-0.5\,\alpha}\sum_1^m w_i(|B + A|^\alpha + |B - A|^\alpha - (|B + x_i|^\alpha + |B - x_i|^\alpha)).
\end{aligned}$$

Based on the expression above, for any ξ^*, to see if $J_{\{(0.5,\pm A)\}}(\theta, u) - J_{\xi^*}(\theta, u)$ is non-negative or not boils down to checking the sign of

$$\sum_1^m w_i(|B + A|^\alpha + |B - A|^\alpha - (|B + x_i|^\alpha + |B - x_i|^\alpha))$$

---

[1]The case for $u = (1, 0)$ can be ignored, since $J_\xi(\theta, (1, 0))$ are the same for all $\xi \in \Xi$.

for any $0 \leq x_1, .., x_m \leq A$ and $B \in [0, \infty)$.

The following shows that $|B+A|^\alpha + |B-A|^\alpha - (|B+x_i|^\alpha + |B-x_i|^\alpha)$ is non-negative for all possible cases defined by relationships between x_i, A, B in location: $x_i \leq A \leq B$, $B \leq x_i \leq A$ and $x_i \leq B \leq A$.

- Case 1, $0 \leq x_i \leq A \leq B$,

  $$|B+A|^\alpha + |B-A|^\alpha - (|B+x_i|^\alpha + |B-x_i|^\alpha) = (B+A)^\alpha + (B-A)^\alpha - (B+x_i)^\alpha - (B-x_i)^\alpha$$

  When $\alpha = 1$,

  $$(B+A) + (B-A) - (B+x_i) - (B-x_i) = 2B - B - B = 0.$$

  When $\alpha > 1$, function $f_1(x) = (B+x)^\alpha + (B-x)^\alpha$ is an increasing function, since its first derivative is always positive when $B > x$,

  $$\frac{\partial f_1(x)}{\partial x} = \alpha[(B+x)^{\alpha-1} - (B-x)^{\alpha-1}] > 0.$$

  Then, for any $x_i, 0 \leq x_i \leq A$, $f_1(A) - f_1(x_i) \geq 0$, i.e.

  $$(B+A)^\alpha + (B-A)^\alpha - (B+x_i)^\alpha - (B-x_i)^\alpha \geq 0, \text{ for all } i = 1, .., m.$$

- Case 2: $0 \leq B \leq x_i \leq A$

$$|B+A|^\alpha + |B-A|^\alpha - (|B+x_i|^\alpha + |B-x_i|^\alpha) = (B+A)^\alpha + (A-B)^\alpha - (B+x_i)^\alpha - (x_i-B)^\alpha$$

When $\alpha = 1$,

$$(B+A) + (A-B) - (B+x_i) - (x_i-B) = 2A - 2x_i \geq 0.$$

When $\alpha > 1$, function $f_2(x) = (B+x)^\alpha + (x-B)^\alpha, 0 \leq B < x$, is an increasing function since its first derivative is always positive,

$$\frac{\partial f_2(x)}{\partial x} = \alpha[(B+x)^{\alpha-1} + (x-B)^{\alpha-1}] > 0.$$

Since $x_i \leq A, f_2(A) - f_2(x_i) \geq 0$, for all i,

$$(B+A)^\alpha + (A-B)^\alpha - (B+x_i)^\alpha - (x_i-B)^\alpha > 0, i = 1, ..., m.$$

- Case 3 When $0 \leq x_i \leq B \leq A$,

$$
\begin{aligned}
&|B+A|^{\alpha}+|B-A|^{\alpha}-(|B+x_i|^{\alpha}+|B-x_i|^{\alpha}) \\
=&(B+A)^{\alpha}+(A-B)^{\alpha}-(B+x_i)^{\alpha}-(B-x_i)^{\alpha} \\
=&(B+A)^{\alpha}-(B+x_i)^{\alpha}+(A-B)^{\alpha}-(B-x_i)^{\alpha}. \qquad (3.9)
\end{aligned}
$$

When $\alpha = 1$, $(B+A)-(B+x_i)+(A-B)-(B-x_i) = 2A-2B \geq 0$.

When $\alpha > 1$, if $A-B \geq B-x_i \geq 0$, then $(A-B)^{\alpha}-(B-x_i)^{\alpha} \geq 0$, so (Equation 3.9) is non-negative.

When $\alpha > 1$, if $0 \leq A-B < B-x_i$, then $(A-B)^{\alpha}-(B-x_i)^{\alpha} < 0$. Let $A-B = d_m, B-x_i = d_i$. Notice that this assumption means $0 \leq d_m < d_i$. Set

$$B+x_i = W, \text{ then } B+A = x_i+d_i+B+d_m = W+d_i+d_m.$$

Consider $f_3(x) = (x+y)^{\alpha}-x^{\alpha}-y^{\alpha}, y > 0, x \geq 0$, $f_3(x)$ is an increasing function, as its first derivative is positive,

$$f_3'(x) = \alpha(x+y)^{\alpha-1}-\alpha x^{\alpha-1} > 0.$$

Also notice that $f_3(0) = 0$, so $f_3(x)$ is a non-negative function.

Therefore, due to $w > 0, d_i > 0$,

$$(W + d_i + d_m)^\alpha - (W)^\alpha - (d_i + d_m)^\alpha > 0 \text{ and } (d_i + d_m)^\alpha - d_i^\alpha - d_m^\alpha > 0.$$

Therefore, when $0 \leq A - B < B - x_i$

$$\begin{aligned}
&(B + A)^\alpha - (B + x_i)^\alpha + (A - B)^\alpha - (B - x_i)^\alpha \\
=&(W + d_i + d_m)^\alpha - (W)^\alpha + (d_m)^\alpha - (d_i)^\alpha \\
>&(d_i + d_m)^\alpha + (d_m)^\alpha - (d_i)^\alpha \\
>&d_i^\alpha + d_m^\alpha + (d_m)^\alpha - (d_i)^\alpha \\
\geq&0.
\end{aligned}$$

In summary of all three cases, no matter where B is in relation to x_i and A,

$$|B + A|^\alpha + |B - A|^\alpha - (|B + x_i|^\alpha + |B - x_i|^\alpha) \geq 0 \text{ for all i=1,...,m.}$$

Therefore, for any given u, and symmetric design $\xi^* = \{(w_i, -x_i), (w_i, x_i), i = 1, .., m\}$,

$$J_{\{(0.5,\pm A)\}}(\theta, u) \geq J_{\xi^*}(\theta, u).$$

Let $\tilde{u} = \underset{u:\|u\|_2=1}{\operatorname{argmin}} J_{(0.5,\pm A)}(\theta, u)$, then following from the above conclusion, for any ξ^*,

$$\min_{u:\|u\|_2=1} J_{(0.5,\pm A)}(\theta, u) = J_{(0.5,\pm A)}(\theta, \tilde{u}) \geq J_{\xi^*}(\theta, \tilde{u}) \geq \min_{u:\|u\|_2=1} J_{\xi^*}(\theta, \tilde{u}).$$

Consequently,

$$\min_{u:\|u\|_2=1} J_{(0.5,\pm A)}(\theta, u) \geq \max_{\xi^* \in \Xi^*} \min_{u:\|u\|_2=1} J_{\xi^*}(\theta, u).$$

3.6.3 Proof of Theorem 4

Recall that optimal design is defined as the design that maximizes the Hellinger information, $\underset{\xi \in \Xi}{\operatorname{argmax}} \min_{u:\|u\|_2=1} J_\xi(\theta, u)$. Lemma 1 says that the optimal design has to be a symmetric design,

$$\underset{\xi^* \in \Xi^*}{\operatorname{argmax}} \min_{u:\|u\|_2=1} J_\xi(\theta, u) = \underset{\xi \in \Xi}{\operatorname{argmax}} \min_{u:\|u\|_2=1} J_\xi(\theta, u).$$

Lemma 2 says that the best design among symmetric designs is the symmetric two-point design on the boundary, i.e.,

$$\{(0.5, -A), (0.5, A)\} = \underset{\xi^* \in \Xi^*}{\operatorname{argmax}} \min_{u:|u|=1} J_\xi(\theta, u).$$

Combining both lemmas,

$$\{(0.5, -A), (0.5, A)\} = \underset{\xi \in \Xi}{\operatorname{argmax}} \min_{u:|u|=1} J_\xi(\theta, u).$$

3.6.4 Proof of Theorem 5

Lemma (2) says that the optimal design for quadratic model must be a symmetric design, so here we only need to search among the collection of symmetric designs.

Given any symmetric design $\xi^* = \{(w_1, -x_1), ...(w_k, -x_m), (w_1, x_1), ...(w_m, x_m)\}$ and direction vector u, the Hellinger information of ξ^* in the direction of $u = (u_1, u_2, u_3)$ has the expression of

$$J_{\xi^*}(\theta, u) = \sum_1^m w_i(|u_1 + u_2 x_i + u_3 x_i^2|^\alpha + |u_1 + u_2(-x_i) + u_3 x_i^2|^\alpha).$$

For simplicity, denote $f_u(x) = u_1 + u_2 x + u_3 x^2$, then when $\alpha = 1$ the above becomes

$$J_{\xi^*}(\theta, u) = \sum_1^m w_i(|f_u(x_i)| + |f_u(-x_i)|).$$

First, we want to show that there exist $r_i \in [0, 1]$ such that the following relation is true, for all $x_i \in [-A, A]$,

$$2r_i|f_u(0)| + (1 - r_i)|f_u(A)| + (1 - r_i)|f_u(-A)| > |f_u(x_i)| + |f_u(-x_i)|. \tag{3.10}$$

Notice that $|f_u(x)| = |f_{-u}(x)|$, i.e. $|u_1 + u_2 x + u_3 x^2| = |-u_1 - u_2 x - u_3 x^2|$. Therefore, for every given $\bar{u}$ with $\bar{u}_3 < 0$, i.e. when $f_{\bar{u}}(x)$ is concave down, there is a $\dot{u} = -\bar{u}$ such that $|f_{\dot{u}}(x)| = |f_{\bar{u}}(x)|$, and $f_{\dot{u}}(x)$ is convex. Thus, for simplicity, the following only shows that (Equation 3.10) is true for $f_u(x)$ with $u_3 > 0$, i.e., only when $f_u(x)$ is convex. There are seven

cases based on the locations of x-intercepts of $f_u(x)$, and for each case, (Equation 3.10) can be shown to be true. Here we only consider cases of u such that $u_3 \neq 0$, since the case for $u_3 = 0$ is equivalent to the linear regression case. In the rest of the proof, for simplicity, let $f(x) \equiv f_u(x)$.

By convexity, if $f(x_i) > 0$ over $[0, B]$ for some $B > 0$ and $x_i \in [-B, B]$, and then there is a $r_i \in (0, 1)$, such that $x_i = r_i 0 + (1 - r_i)B$ and,

$$r_i f(0) + (1 - r_i) f(B) > f(x_i) \text{ and } r_i f(0) + (1 - r_i) f(-B) > f(-x_i),$$

then

$$2r_i |f(0)| + (1 - r_i)|f(B)| + (1 - r_i)|f(-B)| > |f(x_i)| + |f(-x_i)|. \tag{3.11}$$

Given direction vector u and design point location $-x_i, x_i$, with $x_i > 0$ and the assumption that $u_3 > 0$, there are seven cases that describe the possible relationships between $-x_i, x_i$ and the left, right roots of $f(x)$, $x_L < x_R$.

- Case 1 $x_i < x_L, x_R$,
- Case 2 $x_L, x_R < -x_i$
- Case 3: $-x_i \leq x_L, x_R \leq x_i$
- Case 4: $x_L \leq -x_i, x_i \leq x_R$
- Case 5: $-x_i \leq x_L \leq x_i \leq x_R$
- Case 6: $x_L \leq -x_i \leq x_R \leq x_i$
- Case 7: There are at most one root for $f(x)$, i.e. $f(x) > 0$ for all $x \in R$

The following goes through these cases and shows that (Equation 3.10) is true for each of them. Notice that cases 1 and 2 are equivalent, and that cases 5 and case 6 are equivalent.

- In case 1 both roots are above x_i, so there are two possible ways that this can happen regarding the given value of A:
  - 1.1) The left root x_L is above A, i.e. $A \leq x_L$. This implies that $f(x_i) > 0$ over $[-A, A]$, then by the argument of convexity in (Equation 3.11), then (Equation 3.10) is true.
  - 1.2) The left root x_L is below A, i.e. $x_L < A$.

    Under 1.2), $f(-x_i), f(x_i), f(-A) > 0$, which implies that

$$|f(-x_i)| + |f(x_i)| = 2u_1 + 2u_3x^2, \text{ and } f(-A) = u_1 - u_2A + u_3A^2. \tag{3.12}$$

    If A is smaller than right root, $A < x_R$, then $f(A) < 0$, so

    $|f(A)| = -u_1 - u_2A - u_3A^2 > 0$, and $-u_2A > u_1 + u_3A^2$. Then with (Equation 3.12),

$$|f(A)| + |f(-A)| = -2u_2A > 2u_1 + 2u_3A^2 > 2u_1 + 2u_3x_i^2 = |f(-x_i)| + |f(x_i)|.$$

    If A is larger than right root, $A > x_R$, then $f(A) > 0$, so

    $|f(A)| = u_1 + u_2A + u_3A^2 > 0$. Then with (Equation 3.12),

$$|f(A)| + |f(-A)| = 2u_1 + 2u_3A^2 > 2u_1 + 2u_3x_i^2 = |f(-x_i)| + |f(x_i)|.$$

Then for 1.2) one can find a ratio r_A such that $r_A(|f(A)|+f(-A)) > f(x_i)+f(-x_i)$, let $r_i = 1 - r_A$, then (Equation 3.10) is true, i.e.

$$2r_i|f(0)| + (1-r_i)|f(A)| + (1-r_i)|f(-A)| > |f(x_i)| + |f(-x_i)|.$$

- Case 3: $-x_i \leq x_L, x_R \leq x_i$, is the case of both roots of $f(x)$ are in $[-x_i, x_i]$, then $f(x)$ would be positive and increasing over $[x_i, A]$, while positive and decreasing over $[-A, -x_i]$, i.e.

$$f(A) > f(x_i) > 0, \quad f(-A) > f(-x_i) > 0,$$

Let $r_i = 1 - r_A$, then, under $\alpha = 1$, (Equation 3.10) is true, i.e.,

$$2r_i|f(0)| + (1-r_i)|f(A)| + (1-r_i)|f(-A)| > |f(x_i)| + |f(-x_i)|.$$

- Case 4: $x_L \leq -x_i, x_i \leq x_R$. In this case, $f(x) \leq 0$ over $[-x_i, x_i]$, which means $|f(x)| = -f(x) = -u_1 - u_2 x - u_3 x^2$ is concave over $[-x_i, x_i]$. Thus, $|f(0)| > \frac{1}{2}|f(x_i)| + \frac{1}{2}|f(-x_i)|$, consequently, (Equation 3.10) holds.

- Case 5: $-x_i \leq x_L \leq x_i \leq x_R$ and case 6: $x_L \leq -x_i \leq x_R \leq x_i$.

  Since case 6 is the symmetrical to case 5, here we only discuss case 5.

  First, the assumption of case 5, $-x_i \leq x_L \leq x_i \leq x_R$, implies that $\frac{-u_2}{2u_3} = \frac{x_L+x_R}{2} > \frac{-x_i+x_i}{2} = 0$, i.e. $u_2 < 0$.

  Also notice that $-x_i \leq x_L$ implies that $0 < f(-x_i) < f(-A)$ and

$$|f(-A)| = u_1 - u_2 A + u_3 A^2, \quad |f(-x_i)| = u_1 - u_2 x_i + u_3 x_i^2 \tag{3.13}$$

Based on the set up of case 5, and the possible relations of given A and direction u, the expression of $f(x_i)$ and $f(A)$ depends on the following two sub-cases,

– The right boundary A is below right intercept, i.e. $A < x_R$, i.e. $f(x_i) < f(A) < 0$, then

$$|f(A)| = -f(x_i) = -u_1 - u_2 A - u_3 A^2, \text{ and } |f(x_i)| = -f(x_i) = -u_1 - u_2 x_i - u_3 x_i^2.$$

Therefore, with the fact that $-u_2 > 0$, $A \geq x_i$, and (Equation 3.13), we have

$$\begin{aligned} & |f(-A)| + |f(A)| - |f(-x_i)| - |f(x_i)| \\ =& -2u_2 A + 2u_2 x_i \\ =& -2u_2(A - x_i) \\ \geq& 0. \end{aligned}$$

– The right boundary A is above right intercept, i.e. $x_R < A$ which implies that $f(x_i) < 0 < f(A)$,

$$|f(A)| = u_1 + u_2 A + u_3 A^2, \text{ and } |f(x_i)| = -f(x_i) = -u_1 - u_2 x_i - u_3 x_i^2.$$

Therefore, with $-u_2 > 0$, $A \geq x_i$, and (Equation 3.13)

$$\begin{aligned}
&|f(-A)| + |f(A)| - |f(-x_i)| - |f(x_i)| \\
=&2u_1 + 2u_3A^2 + 2u_2x_i \\
=&2(u_1 + u_2x_i + u_3A^2) \\
=&2(u_1 + u_2A + u_3A^2) - u_2(A - x_i) \\
=&2|f(A)| - u_2(A - x_i) \\
\geq&0.
\end{aligned}$$

Combining these two sub-cases, we can conclude that under case 5,

$$|f(-A)| + |f(A)| \geq |f(-x_i)| + |f(x_i)|.$$

Then one can find a ratio r_A such that $r_A(|f(A)|+f(-A)) > f(x_i)+f(-x_i)$, let $r_i = 1-r_A$, then (Equation 3.10) is true

$$2r_i|f(0)| + (1 - r_i)|f(A)| + (1 - r_i)|f(-A)| > |f(x_i)| + |f(-x_i)|.$$

- Case 7: There is at most one root, which means, $f(x) > 0$ for all $x \in [-A, A]$. Thus, by argument in (Equation 3.11), implies (Equation 3.10).

In summary of these seven cases, (Equation 3.10) holds. For a given w_i, after multiplying w_i on both side of the inequality of (Equation 3.10), we have

$$2w_i r_i|f(0)| + w_i(1-r_i)|f(A)| + w_i(1-r_i)|f(-A)| \geq w_i(|f(x_i)| + |f(-x_i)|). \tag{3.14}$$

Let $w = \sum_{i=1}^{m}(1-r_i)w_i$, by the fact that $\sum_{i=1}^{m} w_i = 0.5, 1-2w = \sum_{i=1}^{m} 2w_i r_i$. We can denote a three point symmetric design based on the left hand side of (Equation 3.14) as

$$\xi_w = \{(w,-A),(1-2w,0),(w,A)\}, 0 \leq w \leq 0.5.$$

Hellinger information based on design ξ_w in the direction of a given u has the expression

$$J_{\xi_w}(\theta,u) = (1-2w)|f(0)| + w|f(A)| + w|f(-A)|.$$

Thus, based on (Equation 3.14), for any u, for any symmetric design ξ^*, there is a w, such that

$$J_{\xi_w}(\theta,u) \geq J_{\xi^*}(\theta,u).$$

By the exact same argument and Lemma 1, the conclusion of this theorem holds.

CHAPTER 4

SUMMARY AND DISCUSSION

This project seeks to address the problem that Fisher information does not exist for nonregular models, which implies that, in the context of optimal design, the object of optimization is absent. Our proposed approach introduces and defines an alternative measure of information, namely, Hellinger information, to be used as the object of optimization in optimal design for nonregular models. The theoretical foundation for this approach is based on the Hellinger information inequality presented in Theorem 3, which shows that, when Hellinger information exists, it is proportional to a lower bound of risk for any estimators. Furthermore, the minimum eigenvalue of Fisher information can be viewed as a special case of Hellinger information under regularity conditions. Based on our proposed approach using Hellinger information, we derived optimal designs for some nonregular models over different simulations, with different parameters. The results of these simulations support that our approach is valid in optimal design problems for nonregular models.

This thesis is a first attempt at developing a general approach to optimal design of experiment when regularity conditions do not apply. A number of questions remain that cannot be covered sufficiently in this text. Below are several directions for further work based on this thesis.

- In Chapter 3, optimal design results were presented for polynomial nonregular regressions up to degree two. One line of further research would be to explore how to obtain optimal design for different functions of $g(\theta;(x))$ in the model (Equation 3.1).

  For example, given a nonregular regression model with $g(\theta;x) = e^{\theta^\top \mathbf{x}}$, and exponential distributed error, the optimal design based on Definition 10, would be

$$\xi_{opt.} = \underset{\xi}{\operatorname{argmax}} \min_{u:\|u\|_2=1} \sum_{i=1}^{m} w_i \left| e^{\theta^\top \mathbf{x}_i} \sum_{j=1}^{d} u_j x_{i,j} \right|.$$

  Notice that, because the parameter of interest appears in the expression above, one can use the local optimal design approach for this problem.

- More work is needed in order to determine what kind of estimator is efficient with respect to the Hellinger information inequality presented in Chapter 2. A more rigorous understanding of the Hellinger information inequality and its relationship to the available estimators would be useful.

- As discussed in Chapter 2, Section 2.4, Fisher information also plays the role of being a non-informative prior in Bayesian models. While (Shemyakin, 2014) considered Hellinger information a non-informative prior, he did not adequately justify this use of Hellinger information. (Shemyakin, 2014) only pointed out that, for models with one-dimensional parameter, Fisher information is a special case of Hellinger information. It would therefore be worthwhile to investigate whether and how Hellinger information, with or without direction, can be used as a non-informative prior in Bayesian inference.

CITED LITERATURE

Allen, A. O.: Probability, statistics, and queueing theory. Academic Press, 2014.

Bartkut-Norkūnien, V. and Sakalauskas, L.: Estimation of the three-parameter weibull distribution with applications to large-scale data sets. In The XIIIth International Conference "Applied Stochastic Models and Data Analysis 2009" Selected papers, 2009.

Bartolucci, A. A., Singh, K. P., Bartolucci, A. D., and Bae, S.: Applying medical survival data to estimate the three-parameter weibull distribution by the method of probability-weighted moments. Mathematics and Computers in Simulation, 48(4):385–392, 1999.

Cheng, R. and Amin, N.: Maximum product of spacings estimation with application to the lognormal distribution (mathematical report 79-1). Cardiff: University of Wales IST, 1979.

Chernozhukov, V. and Du, S.: Extremal quantiles and value-at-risk. MIT Department of Economics Working Paper, Jan 2001. http://dx.doi.org/10.2139/ssrn.956433.

Chernozhukov, V. and Hong, H.: Likelihood estimation and inference in a class of nonregular econometric models. Econometrica, 72(5):1445–1480, 2004.

Cousineau, D.: Fitting the three-parameter Weibull distribution: Review and evaluation of existing and new methods. IEEE Transactions on Dielectrics and Electrical Insulation, 16(1):281–288, 2009.

Cousineau, D., Brown, S., and Heathcote, A.: Fitting distributions using maximum likelihood: Methods and packages. Behavior Research Methods, 36(4):742–756, 2004.

De la Garza, A.: Spacing of information in polynomial regression. The Annals of Mathematical Statistics, 25(1):123–130, 1954.

DePriest, D. J.: Using the singly truncated normal distribution to analyze satellite data. Communications in Statistics-Theory and Methods, 12(3):263–272, 1983.

Dette, H., Melas, V. B., et al.: A note on the de la Garza phenomenon for locally optimal designs. The Annals of Statistics, 39(2):1266–1281, 2011.

Donald, S. G. and Paarsch, H. J.: Superconsistent estimation and inference in structural econometric models using extreme order statistics. Journal of Econometrics, 109(2):305–340, 2002.

Finney, D. and Varley, G.: An example of the truncated Poisson distribution. Biometrics, 11(3):387–394, 1955.

Ford, I., Torsney, B., and Wu, C. J.: The use of a canonical form in the construction of locally optimal designs for non-linear problems. Journal of the Royal Statistical Society, Series B (Methodological), 54(2):569–583, 1992.

Fu, H.: Methods for Fitting Truncated Weibull Distributions to Logistic Models. Doctoral dissertation, Dissertation, Duisburg, Essen, Universität Duisburg-Essen, 2016.

Hall, P., Van Keilegom, I., et al.: Nonparametric regression when errors are positioned at end-points. Bernoulli, 15(3):614–633, 2009.

Heathcote, A. and Brown, S.: Reply to Speckman and Rouder: A theoretical basis for qml. Psychonomic Bulletin & Review, 11(3):577–578, 2004.

Heathcote, A., Brown, S., and Cousineau, D.: Qmpe: Estimating lognormal, Wald, and Weibull RT distributions with a parameter-dependent lower bound. Behavior Research Methods, 36(2):277–290, 2004.

Heathcote, A., Brown, S., and Mewhort, D.: Quantile maximum likelihood estimation of response time distributions. Psychonomic bulletin & review, 9(2):394–401, 2002.

Hirano, K. and Porter, J.: Efficiency in asymptotic shift experiments. Technical report, mimeo, University of Miami and Harvard University, 2003.

Hirose, H.: Bias correction for the maximum likelihood estimates in the two-parameter weibull distribution. IEEE Transactions on Dielectrics and Electrical Insulation, 6(1):66–68, 1999.

Hirose, H. and Lai, T. L.: Inference from grouped data in three-parameter weibull models with applications to breakdown-voltage experiments. Technometrics, 39(2):199–210, 1997.

Ibragimov, I. and Hasminskii, R.: Statistical estimation: Asymptotic theory. Springer, 1981.

Jacquelin, J.: Generalization of the method of maximum likelihood (insulation testing). IEEE Transactions on Electrical Insulation, 28(1):65–72, 1993.

Kagel, J. H. and Levin, D.: Implementing efficient multi-object auction institutions: An experimental study of the performance of boundedly rational agents. Games and Economic Behavior, 66(1):221–237, 2009.

Kiefer, J.: General equivalence theory for optimum designs (approximate theory). The annals of Statistics, 2(5):849–879, 1974.

Koenker, R. and Hallock, K.: Quantile regression: An introduction. Journal of Economic Perspectives, 15(4):43–56, 2001.

Le Cam, L. et al.: Limits of experiments. In Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics. The Regents of the University of California, 1972.

Lehmann, E. L. and Casella, G.: Theory of point estimation; 2nd ed.. Springer texts in statistics. New York, NY, Springer, 1998.

Lehmann, E.: Elements of large-sample theory. Springer, 1999.

Lehmann, E. L. and Casella, G.: Theory of Point Estimation. Springer Science & Business Media, 2006.

Ranneby, B.: The maximum spacing method. an estimation method related to the maximum likelihood method. Scandinavian Journal of Statistics, 11(2):93–112, 1984.

Shemyakin, A.: On information inequalities in the parametric estimation. Theory of Probability & its Applications, 37(1):89–91, 1993.

Shemyakin, A.: Hellinger distance and non-informative priors. Bayesian Analysis, 9(4):923–938, 2014.

Smith, R. L.: Maximum likelihood estimation in a class of nonregular cases. Biometrika, 72(1):67–90, 1985.

Smith, R. L.: Nonregular regression. Biometrika, 81(1):173–183, 1994.

Speckman, P. L. and Rouder, J. N.: A comment on Heathcote, Brown, and Mewhort's qmle method for response time distributions. Psychonomic Bulletin & Review, 11(3):574–576, 2004.

Stufken, J. and Yang, M.: Optimal designs for generalized linear models. Design and Analysis of Experiments, Special Designs and Applications, 3:137, 2012.

Umberger, W. J. and Feuz, D. M.: The usefulness of experimental auctions in determining consumers' willingness-to-pay for quality-differentiated products. Review of Agricultural Economics, 26(2):170–185, 2004.

Van der Vaart, A. W.: An asymptotic representation theorem. International Statistical Review/Revue Internationale de Statistique, 59(1):97–121, 1991.

Van der Vaart, A. W.: Asymptotic statistics, volume 3. Cambridge university press, 1998.

Woodroofe, M.: Maximum likelihood estimation of translation parameter of truncated distribution ii. The Annals of Statistics, 2(3):474–488, 1974.

Yang, M.: On the de la Garza phenomenon. The Annals of Statistics, 38(4):2499–2524, 2010.

Yang, M. and Stufken, J.: Support points of locally optimal designs for nonlinear models with two parameters. The Annals of Statistics, 37(1):518–541, 2009.

Yang, M., Stufken, J., et al.: Identifying locally optimal designs for nonlinear models: A simple extension with profound consequences. The Annals of Statistics, 40(3):1665–1681, 2012.

Zaninetti, L.: A right and left truncated gamma distribution with application to the stars. arXiv preprint:1401.0287, 2014.

Zhang, T. and Xie, M.: On the upper truncated Weibull distribution and its reliability implications. Reliability Engineering & System Safety, 96(1):194–200, 2011.

VITA

Education

- Ph.D.Mathematics, Concentration in Statistics, University of Illinois at Chicago, 2018
- BA, Mathematics and BA, Philosophy, Loyola University Chicago, 2010

Presentation

- October 2017 — Hellinger Information and Optimal Design for Nonregular Models (Poster)— The Design and Analysis of Experiments Conference, UCLA
- May 2016 — On Optimal Designs for Nonregular Models (Talk)— Spring Research Conference, Illinois Institute of Technology

Experience

- Sept. 2010 – Dec. 2017 — Teaching Assistant — University of Illinois at Chicago
- May 2015 – Aug. 2015 — Statistics Intern — Helomics Corporation
- Aug 2015 – Oct. 2016 — Statistics Intern — Stat4ward LLC