On optimal designs for non-regular models

Yi Lin∗ Ryan Martin† Min Yang∗
January 29, 2019

Abstract

Classically, Fisher information is the relevant object in defining optimal experimental designs. However, for models that lack certain regularity, the Fisher information does not exist and, hence, there is no notion of design optimality available in the literature. This article seeks to fill the gap by proposing a so-called Hellinger information, which generalizes Fisher information in the sense that the two measures agree in regular problems, but the former also exists for certain types of non-regular problems. We derive a Hellinger information inequality, showing that 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 non-regular optimal design problems. Hellinger optimal designs are derived for several non-regular regression problems, with numerical results empirically demonstrating the efficiency of these designs compared to alternatives.

Keywords and phrases: E-optimality; experimental design; Fisher information; Hellinger distance; information inequality.

1 Introduction

Optimal experimental design is a classical problem with substantial recent developments. For example, Biedermann et al. (2006), Dette et al. (2008), Feller et al. (2017), and Schorning et al. (2017) studied optimal designs for dose-response models; Dette et al. (2016) and Dette et al. (2017) investigated optimal designs for correlated observations; Dror and Steinberg (2006) and Gotwalt et al. (2009) studied robustness issues in optimal designs; López-Fidalgo et al. (2007), Waterhouse et al. (2008), Biedermann et al. (2009), Dette and Titoff (2009), and Dette et al. (2018) studied optimal discrimination designs; Biedermann et al. (2011) studied optimal design for additive partially nonlinear models; Yu (2011), Yang et al. (2013), Sagnol and Harman (2015), and Harman and Benková (2017) investigated algorithms for deriving optimal designs; and Yang and Stufken (2009), Yang (2010), Dette and Melas (2011), Yang and Stufken (2012), and Dette and Schorning.
(2013) built a new theoretical framework for studying optimal designs. The focus of these developments has been exclusively on regular models that enjoy certain normal features asymptotically, such as generalized linear models. However, certain non-regular models may be appropriate in practical applications (e.g., Chernozhukov and Hong 2004; Cousineau 2009; Hirose and Lai 1997). In particular, Smith (1994) describes a class of non-regular linear regression models,

\[ y = x^\top \theta + \varepsilon, \]

where the error \( \varepsilon \) is non-negative, which implies a non-regular model for \( y \), given \( x \), since its distribution has \( \theta \)-dependent support. Such models are useful if the goal is to study extremes; for example, \( x^\top \theta \) might represent the lower bound on remission time when a patient is subjected to treatment settings described by the vector \( x \). To date, there is no literature on optimal designs for cases like this, and the goal of this paper is to fill this gap by developing an approach to optimal design in non-regular problems.

Towards formulating a design problem in a non-regular model, the first obstacle is that the Fisher information matrix—the fundamental object in the classical optimal design context—does not exist. To overcome this, we draw inspiration from recent work on the development of non-informative priors in the Bayesian context, thereby backtracking the path taken by Lindley (1956) and Bernardo (1979) from information in an experiment to non-informative priors. In particular, Shemyakin (2014) proposes an alternative to Fisher information and generalizes the non-informative prior construction of Jeffreys. An important feature of the Fisher information is how it describes the local behavior of the Hellinger distance (see Section 2), leading to its connection to estimator quality via the information inequality. Unfortunately, the role that Shemyakin’s information plays in the local approximation of Hellinger distance for multi-parameter models remains unclear; see Remark 2. Since a connection to the quality of estimators is essential to our efforts to define a meaningful notion of optimal designs, we take an alternative approach where the focus is on a local approximation of Hellinger distance.

We start by looking at the local behavior of the squared Hellinger distance between models \( P_{\theta} \) and \( P_{\vartheta} \), for \( \vartheta \approx \theta \). In the regular cases, there is a local quadratic approximation to the squared distance and the Fisher information matrix appears in the approximating quadratic form. In non-regular problems, by definition, the squared Hellinger distance is not locally quadratic, so there is no reason to expect that an “information matrix” can be extracted from this approximation. In fact, not being differentiable in quadratic mean implies that the Hellinger distance is continuous at \( \theta \), but not differentiable, so important features of the local approximation will generally depend on both the magnitude and the direction of the departure of \( \vartheta \) from \( \theta \). From the local Hellinger distance approximation for a given direction, we define a direction-dependent Hellinger information, which is additive like Fisher information for independent data, and establish a corresponding information inequality that suitably lower-bounds the risk function of an arbitrary estimator along that direction. The direction-dependence is removed via profiling, and the result is a locally minimax lower bound on the risk of arbitrary estimators, which is inversely related to our direction-free Hellinger information. Therefore, just like in the familiar Cramér–Rao inequality for regular models, larger Hellinger information means a smaller lower bound and, consequently, better estimation in terms of risk.

The established connection between our Hellinger information for non-regular models
and the quality of estimators provides a natural path to approach the optimal design problem. In particular, our Hellinger information depends on the design, so we define the optimal design as one that maximizes the Hellinger information. The intuition, just like in the regular case, is that maximizing the information minimizes the lower bound on the risk, thereby leading to improved estimation. If the model happens to be regular, then our proposed optimal design corresponds to the classical E-optimal design that maximizes the minimum eigenvalue of the Fisher information matrix, so the new approach at least has intuitive appeal. After formally defining the notion of optimal design in this context, we develop some novel theoretical results, in particular a complete class theorem for symmetric designs in the context of non-regular polynomial regression. This theorem, along with some special cases presented in Propositions 4–5, suggests the potential for a line of developments parallel to that for regular models.

The remainder of the paper is organized as follows. Section 2 sets our notation and briefly reviews the Fisher information and its properties under regularity conditions. We relax those regularity conditions in Section 3 and develop a notion of Hellinger information for certain non-regular models. The main result of the paper, Theorem 1, establishes a connection between this Hellinger information and the quality of estimators, thus paving the way for a framework of optimal designs for non-regular models in Section 4. Some specific non-regular regression models are considered in Section 5, and we derive some analytical optimality results and some numerical demonstrations of the improved efficiency of the optimal designs over other designs. Some concluding remarks are given in Section 6 and proofs of the two main theorems are presented in Appendix A; the remaining details are given in the Supplementary Material (Lin et al. 2018).

2 Review of information in regular models

The proposed model assumes that the $Y$-valued observations $Y_1, \ldots, Y_n$ are independent, and the marginal distribution of $Y_i$ is $P_{i,\theta}$, where $\theta$ is a fixed and unknown parameter in $\Theta \subseteq \mathbb{R}^d$. For example, $P_{i,\theta}$ might be a distribution that depends on both the parameter $\theta$ and a fixed covariate vector $x_i$. We will further assume that, for each $i = 1, \ldots, n$, $P_{i,\theta}$ has a density $p_{i,\theta}$ with respect to a fixed dominating $\sigma$-finite measure $\mu$. When the index $i$ is not important, and there is no risk of confusion, we will drop the index and write simply $p_{\theta}$ for the density function with respect to $\mu$.

It is common to assume that the model is regular in the sense that $\theta \mapsto p_{i,\theta}(y)$ is smooth for each $y$, and that $\theta$-derivatives of expectations can be evaluated by interchanging differentiation and integration. For example, under conditions (6.6) in Lehmann and Casella (1998), one can define the $d \times d$ Fisher information matrix $I_i(\theta)$, whose $(k,\ell)$ entry is given by

$$E_\theta \left\{ \frac{\partial}{\partial \theta_k} \log p_{i,\theta}(Y_i) \cdot \frac{\partial}{\partial \theta_\ell} \log p_{i,\theta}(Y_i) \right\}, \quad k, \ell = 1, \ldots, d. \quad (1)$$

The Fisher information matrix can be defined in broader generality for families of distributions with a differentiability in quadratic mean property (e.g., Pollard 1997, van der Vaart 1998). That is, assume that there exists a function $\ell_\theta$, typically the gradient of
log \( p_\theta \), taking values in \( \mathbb{R}^d \), such that

\[
\int \left( \mathbf{1}^{1/2} - \mathbf{1}^{1/2} - \frac{1}{2} \varepsilon^\top \ell_\theta \mathbf{1}^{1/2} \right)^2 \, d\mu = o(\|\varepsilon\|^2), \quad \varepsilon \to 0,
\]

where \( \| \cdot \| \) denotes the \( \ell_2 \)-norm. Then the Fisher information matrix exists and is given by the formula

\[
I(\theta) = \int \ell_\theta \ell_\theta^\top \, p_\theta \, d\mu.
\]

If we let \( H(P_\theta, P_\vartheta) \) denote the Hellinger distance and define \( h \) as

\[
h(\theta; \vartheta) \equiv H_2^2(P_\theta, P_\vartheta) := \int \left( \frac{1}{2} \mathbf{1}^{1/2} - \frac{1}{2} \vartheta^{1/2} \right)^2 \, d\mu = 2 - 2 \int (p_\theta p_\vartheta)^{1/2} \, d\mu,
\]

then the above condition amounts to \( h \) being locally quadratic:

\[
h(\theta; \theta + \varepsilon) = \frac{1}{4} \varepsilon^\top I(\theta) \varepsilon + o(\|\varepsilon\|^2).
\]

Therefore, a model is regular if the squared Hellinger distance is locally approximately quadratic, with the Fisher information matrix characterizing that quadratic approximation. This is the description of Fisher information that we will attempt to extend to the non-regular case below.

Recall, also, that Fisher information is additive under independence. That is, if \( Y_1, \ldots, Y_n \) are independent, with \( Y_i \sim p_{i, \theta} \), regular as above for each \( i \), then the Fisher information in the sample of size \( n \) satisfies

\[
\mathcal{I}_n(\theta) = \sum_{i=1}^n I_i(\theta),
\]

where \( I_i(\theta) \) is the Fisher information matrix in \([1]\) based on \( p_{i, \theta} \) alone. This property has a nice interpretation: larger samples have more information.

Under differentiability in quadratic mean, one can prove an information inequality which states that, for any unbiased estimator \( T = T(Y_1, \ldots, Y_n) \) of \( m(\theta) = E_\theta(T) \in \mathbb{R} \) with finite second moment, the variance is lower-bounded and satisfies

\[
V_\theta(T) \geq \dot{m}(\theta)^\top \mathcal{I}_n(\theta)^{-1} \dot{m}(\theta),
\]

where \( \dot{m}(\theta) \) is the gradient of \( m(\theta) \); see [Pollard (2005)]. The information inequality above, and its various extensions, establishes a fundamental connection between the quality of an estimator—in this case, the variance of an unbiased estimator—and the Fisher information matrix. This connection has been essential to the development of optimal design theory and practice since the quality of an estimator can be “optimized” by choosing a design that makes the quadratic form in the lower bound as small as possible, or the Fisher information as large as possible.

Finally, differentiability in quadratic mean implies local asymptotic normality (e.g., [van der Vaart 1998], Theorem 7.2) which is almost all one needs to show that maximum likelihood estimators are efficient in the sense that they attain the information inequality lower bound (e.g., [van der Vaart 1998], Theorem 7.12). Therefore, in sufficiently regular problems, there is a general procedure for constructing high-quality estimators, and that the quality of such estimators is controlled by the Fisher information matrix. The remainder of this paper is concerned with non-regular cases and, unfortunately, these differ.
from their regular counterparts in several fundamental ways. First, the Fisher information is not well-defined in non-regular cases, so we have no general way of measuring the quality of estimators. Second, one cannot rely on maximum likelihood for constructing good estimators. For example, Le Cam writes (see van der Vaart [2002, p. 674])

The author is firmly convinced that a recourse to maximum likelihood is justifiable only when one is dealing with families of distributions that are extremely regular. The cases in which maximum likelihood estimates are readily obtainable and have been proved to have good properties are extremely restricted.

Therefore, to achieve our goals, we need a measure of information that is flexible enough to handle non-regular problems and is connected to estimation quality in general, but does not depend on a particular estimator. The Hellinger information, defined in Section 3.1, will meet these criteria and will provide a basis for defining optimal designs in non-regular problems.

3 Information in non-regular models

3.1 Definition and basic properties

To start, we consider the scalar case with \( d = 1 \). Suppose that there exists a constant \( \alpha \in (0, 2] \) such that, for each \( \theta \), the limit \( J(\theta) = \lim_{\epsilon \to 0} |\epsilon|^{-\alpha} h(\theta; \theta + \epsilon) \) exists, is finite, and non-zero. If such an \( \alpha \) exists, then it must be unique; but there are cases where existence fails, e.g., when \( \theta \) is not identifiable, so that \( h(\theta, \theta + \epsilon) \equiv 0 \) for all sufficiently small \( \epsilon \).

The case \( \alpha = 2 \) corresponds to differentiable in quadratic mean and, hence, “regular,” while \( \alpha \in (0, 2) \) corresponds to “non-regular.” Differentiability of \( \vartheta \mapsto H(P_{\theta}, P_{\vartheta}) \) or lack thereof determines a model’s regularity, so the largest value \( \alpha \) can take is 2; otherwise, the limit is infinite. From the above limit, there is a local approximation,

\[
h(\theta; \vartheta) = J(\theta)|\theta - \vartheta|^{\alpha} + o(|\theta - \vartheta|^{\alpha}). \tag{2}
\]

This resembles the local Hölder condition considered in Ibragimov and Hasminskii [1981, Section I.6]. We call \( \alpha \) the index of regularity and \( J(\theta) \) the Hellinger information. Of course, if \( \alpha = 2 \), then \( J(\theta) \) is proportional to \( I(\theta) \), the Fisher information. Next are a few quick examples, all with \( \alpha = 1 \).

- If \( P_{\theta} = \text{Unif}(0, \theta) \), \( \theta > 0 \), then \( J(\theta) = \theta^{-1} \).
- If \( P_{\theta} = \text{Unif}(\theta^{-1}, \theta) \), \( \theta > 1 \), then \( J(\theta) = (\theta^2 + 1)\{\theta(\theta^2 - 1)\}^{-1} \).
- If \( P_{\theta} = \text{Unif}(\theta; \theta^2) \), \( \theta > 1 \), then \( J(\theta) = (2\theta + 1)\{\theta(\theta - 1)\}^{-1} \).

A class of non-regular models of particular interest to us here are those in Smith [1994] based on location shifts of distributions supported on the positive half-line. Consider a density \( p_0 \) on \((0, \infty)\) that satisfies

\[
p_0(y) = \beta y^{\beta - 1}, \quad \text{as } y \to 0, \tag{3}
\]
where \( \beta \geq 1 \) and \( c = c(\beta) \in (0, \infty) \). For example, the gamma and Weibull families, with shape parameter \( \beta \) and scale \( \sigma \), have \( c = \{\beta \sigma^\beta \Gamma(\beta)\}^{-1} \) and \( c = \sigma^{-\beta} \), respectively. The next result identifies the regularity index \( \alpha \) and the Hellinger information \( J(\theta) \) for this class of location parameter problems, with \( p_0(y) = p_0(y - \theta) \). It shows that \( \alpha \) need not be an integer and the Hellinger information, like Fisher’s, is constant in location models. When \( \beta \geq 2 \), the model is regular—with \( \alpha = 2 \) and the Fisher information defined as usual—so we focus here on the non-regular case with \( \beta \in [1, 2) \).

**Proposition 1.** Let \( p_0 \) satisfy (3) with \( \beta \in [1, 2) \). If, for some \( \Delta > 0 \),
\[
\int_{\Delta}^{\infty} \left( \frac{d}{dy} \log p_0(y) \right)^2 p_0(y) dy < \infty,
\]
then \( \alpha = \beta \) and \( J(\theta) \equiv c\{1 + \beta r(\beta)\} \), where \( c \) is as in (3) and
\[
r(\beta) = \int_{0}^{\infty} \{(w + 1)^{(\beta-1)/2} - w^{(\beta-1)/2}\}^2 dw.
\]

**Proof.** See Section S2.1 in the Supplementary Material.

Ibragimov and Hasminskii (1981, Theorem VI.1.1) show that, in this case, \( h(\theta; \theta + \varepsilon) = O(|\varepsilon|^{\beta}) \) as \( \varepsilon \to 0 \), but they do not identify \( J(\theta) \). Similar results have appeared elsewhere in the literature on non-regular models; our condition (4) is basically the same as Condition \( C_5 \) in Woodroofe (1974), which is basically the same as Assumption 9 in Smith (1985).

Turning to the general, non-regular multi-parameter case, where \( \Theta \) is an open subset of \( \mathbb{R}^d \), defining Hellinger information requires some additional effort. In particular, non-regularity implies that the familiar local quadratic approximation of \( h \) fails, so we should not expect to have an “information matrix” to describe the local behavior in such cases. In fact, \( h(\theta; \vartheta) \) depends locally on the direction along which \( \vartheta \) approaches \( \theta \), so there is no “direction-free” summary of the local structure and, hence, no “information matrix”; see Remark 2. But this lack of a convenient quadratic approximation need not stop us from defining a suitable Hellinger information.

**Definition 1.** Let \( \Theta \) be an open subset of \( \mathbb{R}^d \), for \( d \geq 1 \), and let \( u \) denote a generic direction, a \( d \)-vector with \( \|u\| = 1 \). 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 \( \infty \):
\[
\lim_{\varepsilon \to 0} \frac{h(\theta; \theta + \varepsilon u)}{|\varepsilon|^\alpha} = J(\theta; u).
\]

Then, the following local approximation holds:
\[
h(\theta; \theta + \varepsilon u) = J(\theta; u) |\varepsilon|^\alpha + o(|\varepsilon|^\alpha), \quad \varepsilon \to 0.
\]

This defines the index of regularity \( \alpha \) and the Hellinger information \( J(\theta; u) \) at \( \theta \) in the direction of \( u \).
Since the approximation (7) is in terms of $|\varepsilon|$, it follows that $J(\theta; u) = J(\theta; -u)$, so $J(\theta; u)$ really only depends on the line defined by $u$. If $d = 1$, then there is only one line, i.e., $u = \pm 1$, hence, for the scalar case, we can drop the $u$ argument entirely and write $J(\theta)$ as described above. It is also worth pointing out that Definition 1 assumes that a single index $\alpha$ suffices to describe the regularity of a model with a $d$-dimensional parameter. This is appropriate for the kinds of regression models we have in mind here, but can be a limitation in other cases; see Remark 1 below.

As a quick example, let $P_\theta = \text{Unif}(\theta_1, \theta_1 + \theta_2)$, where $\theta_1 \in \mathbb{R}$ and $\theta_2 > 0$. In this form, $\theta_1$ and $\theta_2$ are location and scale parameters, respectively. If $u = (u_1, u_2)$ is a generic vector on the unit circle, then $J(\theta; u) = \theta_2^{-1} g(u)$, where $g(u)$ has a form which is slightly too complicated to present here; see Section 5.1 in the Supplementary Material. This expression agrees with the familiar properties of Fisher information for location-scale models.

Although we do not define an “information matrix” in the non-regular case (see Remark 2), when the model is regular, i.e., when $\alpha = 2$, there are still some connections between our Hellinger information and the familiar Fisher information. In particular, $J(\theta; u)$ is a quadratic form involving the Fisher information $I(\theta)$ and the direction $u$. This gives an alternative explanation of how the regular models admit a separation of the dependence on $\theta$ and on the direction $u$ of departure from $\theta$.

**Proposition 2.** For a regular model, with $\alpha = 2$, if $I(\theta)$ denotes the $d \times d$ Fisher information matrix, then $J(\theta; u) = \frac{1}{2} u^T I(\theta) u$.

Another useful and familiar feature of Fisher information that also holds for Hellinger information is the reparametrization formula (Proposition 3), which comes in handy for regression problems where the natural parameter is expressed as a function of covariates and another parameter.

### 3.2 Hellinger information inequality

We now return to our original setup where $Y_1, \ldots, Y_n$ are independent, but not necessarily identically distributed, with $Y_i \sim P_{i,\theta}$, $i = 1, \ldots, n$, and $\theta$ is an unknown parameter taking values in an open subset $\Theta$ of $\mathbb{R}^d$ for some $d \geq 1$. Let $P_\theta$ denote the joint distribution of $Y^n = (Y_1, \ldots, Y_n)$. Motivated by the regression problems below, we assume that each $P_{i,\theta}$ has the same index of regularity, $\alpha \in (0, 2]$. Following our intuition from the regular case, define the Hellinger information at $\theta$, in the direction of $u$, based on the sample of size $n$, as

$$H_n(\theta; u) = \sum_{i=1}^{n} J_i(\theta; u),$$

where $J_i(\theta; u)$ is the Hellinger information based on $P_{i,\theta}$ as described above. See Remark 3 for more on this additivity property. Theorem 4 below will establish a suitable connection between $H_n(\theta; u)$ and the quality of an estimator, and this will provide the necessary foundation for defining optimal designs for non-regular models.

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

$$R_{\psi}(T_n, \theta) = \mathbb{E}_\theta \|T_n - \psi(\theta)\|^2,$$
the $q$-vector version of mean square error, where expectation, $E_{\theta}$, is with respect to $P_{\theta}$. This covers the case where $\psi(\theta) = \theta$ and $q = d$, so that interest is in the full parameter $\theta$, and the case where $\psi(\theta)$ is a single component of $\theta$ and $q = 1$, as well as other intermediate cases. Next is the aforementioned lower bound on the risk in terms of the total Hellinger information.

**Theorem 1.** Let $Y^n = (Y_1, \ldots, Y_n)$ consist of independent observations with $Y_i \sim P_{i,\theta}$, $i = 1, \ldots, n$. Let $\alpha \in (0, 2]$ denote the common index of regularity, and $J_n(\theta; u)$ the corresponding Hellinger information in (8). Let $\psi : \Theta \to \mathbb{R}^q$ be a differentiable function with full-rank $q \times d$ derivative matrix $D\psi(\theta)$, and let $T_n = T(Y^n)$ be any estimator of $\psi(\theta)$ with risk function defined in (9). If $\varepsilon_{n,u} = \{3J_n(\theta; u)\}^{-1/\alpha}$, and

$$\lim_{n \to \infty} \inf_{u} [n^{-1}J_n(\theta; u)] > 0,$$

then, for all large $n$,

$$\inf_{T_n} \sup_{\theta \in B_n(\theta)} R_{\psi}(T_n, \vartheta) \geq \left[ \inf_{u} \{\|D\psi(\theta) u\|^{-\alpha} J_n(\theta; u)\} \right]^{-2/\alpha},$$

where $B_n(\theta) \subset \Theta$ is the region whose boundary is determined by the union of $\{\theta + \varepsilon_{n,u}u\}$ over all directions $u$.

**Proof.** See Appendix A.1.

Two very brief comments: first, the universal constant hidden in “$\geq$” is known and given in the proof; second, there is nothing special about “3” in the definition of $\varepsilon_{n,u}$, any number strictly greater than 2 would suffice.

Some additional comments about the interpretation of Theorem 1 are in order. First, the reason for taking supremum over a small “neighborhood” of $\theta$ is that a lucky choice of $T_n \equiv \psi(\theta)$ has excellent performance at $\theta$, but poor performance at a nearby $\vartheta$. The theorem basically says that, if one looks at a locally uniform measure of risk, which prevents “cheating” towards or luck at a particular $\theta$, then one cannot have smaller risk than that in the lower bound (11). The classical Cramér–Rao lower bound uses unbiasedness of the estimator to prevent this kind of cheating/luck.

To assess the sharpness of the bound in (11) when regularity conditions do not apply, consider the case where $q = 1$, so that $\psi(\theta)$ is a scalar function. For the rate, if we consider the identically independently distributed case, so that $J_n(\theta; u) = nJ_1(\theta; u)$, then it follows that the lower bound is of order $n^{-2/\alpha}$, which agrees with the known minimax rate for estimators in non-regular models (Ibragimov and Hasminskii [1981] Sec. I.5). Therefore, the bound cannot be improved in terms of dependence on the sample size. To assess the quality of the lower bound in terms of its dependence on $\theta$, if the observations come from $\text{Unif}(0, \theta)$, 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

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

Asymptotically, this expression is proportional to $\theta^2 n^{-2}$, which agrees with our lower bound. Therefore, up to universal constants, the bound in Theorem 1 is sharp. Whether
there exists an estimator that can attain the bound exactly or asymptotically is unclear in general; see Remark 4.

It is worth stating the special case where $\alpha = 2$ as a corollary to Theorem 1. This reveals some connection to the classical Cramér–Rao bound, even though we do not have access to an information matrix, and demonstrates the generality of our result.

**Corollary 1.** When $\alpha = 2$, if $\psi : \Theta \to \mathbb{R}^q$ has $q \times d$ derivative matrix $D_\psi(\theta)$ of rank $q \leq d$, and $I_n(\theta)$ is the positive definite $d \times d$ Fisher information matrix, then the lower bound in (11) is proportional to

$$
\lambda_{\text{max}} \{ D_\psi(\theta) I_n(\theta)^{-1} D_\psi(\theta)^\top \},
$$

where $\lambda_{\text{max}}(A)$ denotes the maximal eigenvalue of a matrix $A$.

**Proof.** See Section S2.2 in the Supplementary Material.

For comparison to the classical setting, if we take $\psi(\theta) = \theta$, then the expression in the above display simplifies to

$$
\lambda_{\text{max}} \{ I_n(\theta)^{-1} \} = \lambda_{\text{min}}^{-1} \{ I_n(\theta) \}. \tag{12}
$$

Wanting the information matrix to have a large minimal eigenvalue is a familiar concept in the classical optimal design theory; see Section 4.

This and the previous subsection, along with the remarks in Section 3.3, establish some important properties and insights concerning our proposed Hellinger information. A difficulty that has not yet been addressed is the dependence of $J_n(\theta; u)$ on the arbitrary direction $u$. However, the lower bound in (11) is free of a direction, so it makes sense to formulate a direction-free Hellinger information based on that. For a non-regular model as formulated above, with index of regularity $\alpha \in (0, 2]$, we set the direction-free Hellinger information at $\theta$, for interest parameter $\psi(\theta)$, as

$$
J_n^\psi(\theta) = \inf_u \{ \| D_\psi(\theta) u \|^{-\alpha} J_n(\theta; u) \}. \tag{13}
$$

In the special case where $\psi(\theta) = \theta$, this simplifies to

$$
J_n(\theta) = \inf_u J_n(\theta; u). \tag{14}
$$

Moreover, in the regular case with $\alpha = 2$, it follows from Corollary 1 and, in particular, (12), that $J_n(\theta)$ above is (proportional to) the smallest eigenvalue of the Fisher information matrix. Therefore, definition (13) seems very reasonable; more details are presented in Section 4.

### 3.3 Technical remarks

**Remark 1.** Definition 1 does not allow $\alpha$ to depend on $u$, so each component of $\theta$, treated individually, must have the same index of regularity. To see this, consider an exponential distribution with location and rate parameters $\theta_1$ and $\theta_2$, respectively. If $\theta_1$ was fixed and only $\theta_2$ was unknown, then it is a regular problem and the above definition would hold with $\alpha = 2$. Similarly, if $\theta_2$ was fixed and only $\theta_1$ was unknown, then the definition
holds with $\alpha = 1$ according to Proposition 1. However, if both $\theta_1$ and $\theta_2$ are unknown, then the model does not satisfy the conditions of Definition 1. 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; likewise, if $\alpha = 2$, then $J(\theta; u)$ is in $(0, \infty)$ but $J(\theta; u')$ is infinite. Therefore, the above definition cannot accommodate situations where the components of $\theta$, treated individually, would have different regularity indices. But the design applications we have in mind in this paper fit naturally within a setting where all components have the same regularity; the more general case will be considered elsewhere.

Remark 2. Our definition of Hellinger information coincides with that in Shemyakin (2014) for one-parameter models, but our perspectives differ when it comes to multi-parameter models. Shemyakin defines a “Hellinger information matrix” for non-regular problems, which seems to contradict our above claim that no such matrix is available, so some more detailed comments are necessary. Shemyakin makes no claim that his information matrix is related to the local behavior of $h$, and we are unable to conclude definitively whether it is or is not. We do know, however, that $\vartheta \mapsto h(\theta, \vartheta)$ is “bowl-shaped” (though not smooth) at each $\theta$, so if such a matrix could describe the local behavior, then it ought to be non-negative definite. However, Shemyakin (2014, p. 931) admits that a general non-negative definiteness result has not been established for his information matrix. Without a non-negative definiteness result for his Hellinger information matrix, lower bounds like those in, e.g., Shemyakin (1991, 1992) may not be informative, and its use in defining optimal designs lacks justification.

Remark 3. In (8) we defined the Hellinger information in an independent sample of size $n$ as $J_n(\theta; u) = \sum_{i=1}^n J_i(\theta; u)$, the sum of the individual Hellinger information measures. This, however, is not a choice made by us, it is a consequence of the proof of Theorem 1. To see this, heuristically, start with the Hellinger distance between joint distributions $P^\theta_n$ and $P^\vartheta_n$, assuming independence. A straightforward calculation reveals

$$H^2(P^\theta_n, P^\vartheta_n) = 2 - 2 \prod_{i=1}^n \left\{ \int \{p_i,\theta(y_i)p_i,\vartheta(y_i)\}^{1/2} dy_i \right\}$$

$$= 2 - 2 \exp \left\{ \sum_{i=1}^n \log \left[ 1 - \frac{1}{2} H^2(P_i, P_i, \vartheta) \right] \right\}$$

Since $\log(1 + x) \approx x$ for $x \approx 0$, if $\vartheta$ is sufficiently close to $\theta$, then the exponent is approximately $-\frac{1}{2} \sum_{i=1}^n H^2(P_i, P_i, \vartheta)$ and then, by Taylor’s theorem applied to $x \mapsto e^{-x}$ at $x \approx 0$, we conclude that

$$H^2(P^\theta_n, P^\vartheta_n) \approx \sum_{i=1}^n H^2(P_i, P_i, \theta).$$

Therefore, a local approximation of the left-hand side is roughly equal to a sum of local approximations on the right-hand side, which leads to (8).

Remark 4. An important unanswered question in the above theory is if there are any estimators that are efficient in the sense that they attain the lower bound in Theorem 1 in some generality. In the simple $\text{Unif}(0, \theta)$ example above, we showed that the bound is asymptotically attained, up to universal constants, by the sample maximum. General
results about the rate of convergence in non-regular models are consistent with our lower bound, but, to our knowledge, more precise results concerning the asymptotic behavior of estimators in non-regular problems is limited to certain special cases. Our work here provides some motivation for further investigation of these asymptotic properties. Not having an estimator that provably attains the lower bound complicates our attempts to demonstrate the efficiency gains of our proposed optimal designs in Section 4, but a quality estimator is available in the applications we have in mind; see Section 5.3.

4 Optimal designs for non-regular models

4.1 Definition

The previous section built up a framework of information, based on a local approximation of the squared Hellinger distance, suitable for non-regular problems where Fisher information does not exist. Our motivation for building such a framework was to address the problem of optimal experimental designs in cases where the underlying statistical model is non-regular. This section defines what we mean by an optimal design for non-regular models, and provides some additional details about the Hellinger information that are particularly relevant to the design problem.

We start here with a slightly different setup than in the previous section, but quickly connect it back to the preceding. Let \( Y_1, \ldots, Y_n \) be independent observations, where \( Y_i \) has density function \( q_{\eta_i} \), for \( i = 1, \ldots, n \). That is, each \( Y_i \) has its own parameter \( \eta_i \), which we will assume is real-valued, as is typical in linear and generalized linear models. Then the design problem proceeds by expressing the unit-specific parameter \( \eta_i \) as a given function \( g(x_i, \theta) \) of a common parameter \( \theta \in \mathbb{R}^d \) and a vector of unit-specific covariates; here, of course, the covariates are constants that the investigator is able to set in any way he/she pleases, but preferably in a way that is “optimal” in some sense. By linking each \( \eta_i \) to a common \( \theta \), we obtain the setup from previous sections, i.e., \( Y_i \sim p_{i, \theta} \), independent, for \( i = 1, \ldots, n \).

The next result, stated in the context of \( n = 1 \), parallels a familiar one in the regular case for Fisher information. It aids in computing the Hellinger information under a reparametrization like the one described above.

**Proposition 3.** Let \( q_\eta \) be a density function depending on a scalar parameter \( \eta \), and suppose that the index of regularity is \( \alpha \in (0, 2] \) and the Hellinger information is \( \tilde{J}(\eta) \). Define a new density \( p_\theta \), for \( \theta \in \Theta \subseteq \mathbb{R}^d \), as \( q_{g(\theta)} \) where \( g : \Theta \rightarrow \mathbb{R} \) is a smooth function with non-vanishing gradient \( \dot{g} \). Then \( p_\theta \) also has index of regularity \( \alpha \), and the corresponding Hellinger information at \( \theta \), in the direction of \( u \), is

\[
J_n(\theta; u) = |\dot{g}(\theta)^\top u|^\alpha \tilde{J}(g(\theta)).
\]

**Proof.** See Section S2.3 in the Supplementary Material.

From the general theory in Section 3, if \( Y_1, \ldots, Y_n \) are independent, then under the assumptions in Proposition 3, i.e., \( Y_i \sim p_{i, \theta} = q_{g_i(\theta)} \), the Hellinger information at \( \theta \), in direction of \( u \), is

\[
J_n(\theta; u) = \sum_{i=1}^n |\dot{g}_i(\theta)^\top u|^\alpha \tilde{J}(g_i(\theta)).
\]
For the special case where \( g_i(\theta) = g(x_i, \theta) \) for covariates \( x_i \), it is clear that \( J_n(\theta; u) \) depends on \( x_1, \ldots, x_n \). For example, if \( Y_1, \ldots, Y_n \) are independent, with \( Y_i \sim g(x_i, \theta) + \Gamma(\beta, 1) \), where \( g(x, \theta) = \theta_0 + \sum_{k=1}^{p} \theta_k x_k^{k+1} \), then it follows from Propositions 1 and 3 that

\[
J_n(\theta; u) = 1 + \frac{\beta r(\beta)}{\beta \Gamma(\beta)} \sum_{i=1}^{n} \left| \sum_{k=0}^{p} x_k^k u_{k+1} \right|^\beta.
\]

The Hellinger information’s dependence on the covariates \((x_1, \ldots, x_n)\) is what makes our theory of optimal design possible.

In what follows, we focus exclusively on the case of \( \psi(\theta) = \theta \), and the direction-free definition of Hellinger distance in (14), though this is only for simplicity. The same derivations can be carried out with any specific interest parameter \( \psi(\theta) \) in mind.

Following the now-standard approximate design theory put forth by Kiefer (1974), let \( \xi \) denote a discrete probability measure defined on the design space—the space where the covariates \( x_i \) live—with at most \( m \) distinct atoms, representing the design itself. That is, the atoms of \( \xi \) represent the specific design points, and the probabilities correspond to the weights (more details below). Next, with a slight abuse of our previous notation, we write \( J_\xi(\theta; u) \) to indicate that the Hellinger information in the direction \( u \) depends on the design \( \xi \) through the specific covariate values. For example, given design \( \xi = \{(w_i, x_i) : i = 1, \ldots, m\} \), then \( J_\xi(\theta; u) = \sum_{i=1}^{m} w_i J_i(\theta; u) \), where \( J_i(\theta; u) \) is the Hellinger information in the direction \( u \) based on one observation taken at location \( x_i \). Following (14), the Hellinger information based on design \( \xi \) is defined as

\[
J_\xi(\theta) = \inf_u J_\xi(\theta; u).
\]

Naturally, the optimal design under this setup would be defined as the one that maximizes this measure of information.

**Definition 2.** Under the non-regular model setup presented above, the optimal design \( \xi^* \) is one which maximizes the Hellinger information, i.e.,

\[
\xi^* = \arg\max_\xi J_\xi(\theta).
\]

For comparison to the classical design theory, property (12) implies that our optimal design in Definition 2 under a regular model, corresponds to an E-optimal design, one that maximizes the minimum eigenvalue of the Fisher information matrix. For the non-regular case, however, we do not have an information matrix, so it is not clear if other common notions of optimality, such as A- and D-optimality, have any meaning. For example, non-regularity will cause sampling distributions of estimators to be non-ellipsoidal, so we cannot expect the determinant of some information matrix to correspond to the volume of a confidence ellipsoid.

Definition 2 formulates a new class of optimal design problems, deserving further attention. As discussed briefly in Section 1, there is now a substantial literature on theory and computation related to the optimal design problem in regular cases, and we hope that this paper stimulates a parallel line of work with similar developments for non-regular cases. There are some similarities to the regular case, in particular, the
Hellinger information is non-negative and additive like Fisher information. Also, the map \( \xi \mapsto J_\xi(\theta) \) is concave for fixed \( \theta \), i.e., for any two designs \( \xi \) and \( \xi' \) and any \( w \in [0, 1] \),

\[
J_{w\xi+(1-w)\xi'}(\theta) \geq wJ_\xi(\theta) + (1-w)J_{\xi'}(\theta),
\]

which is important for numerical and/or analytical solution of the optimal design problem. The following gives some first results along these lines.

4.2 A general result for non-regular polynomial models

Motivated by the setup in [Smith (1994)], we consider a non-regular model of the form

\[
y_i = g(x_i, \theta) + \varepsilon_i, \quad i = 1, \ldots, n,
\]

where \( x_i \) are scalars, \( g(x, \theta) = \theta_0 + \sum_{k=1}^{p} \theta_k x^k \) is a degree-\( p \) polynomial, \( \theta \in \mathbb{R}^d \), with \( d = p + 1 \), is an unknown parameter, and \( \varepsilon_i \) are independent and identically distributed with density \( p_0 \) given in (3) and known shape parameter \( \alpha \in [1, 2) \). As is customary (e.g., Koenker and Hallock 2001), we will insist that the design points be centered at the origin, which puts a constraint on the design itself. In particular, we will consider the space of designs \( \xi \) given by

\[
\Xi = \{ \xi = (w_i, x_i) : \sum_i w_i x_i = 0, x_i \in [-A, A] \},
\]

i.e., designs on \([-A, A]\) that are “balanced” in the sense that the mean \( x \) value is 0, where \( A > 0 \) is fixed and known.

The following result shows that, among balanced designs, the subclass of symmetric designs is complete in the sense that the maximum information over symmetric designs is the same as that over the larger class of balanced designs. This implies that the search for an optimal design can be simplified by restricting it to the smaller class of symmetric designs.

**Theorem 2.** Let \( \Xi_{\text{sym}} \subset \Xi \) denote the set of all balanced designs that are also symmetric in the sense that if \( x \) is a design point, then it assigns equal weight to both \( x \) and \(-x\). Then

\[
\max_{\xi \in \Xi_{\text{sym}}} J_\xi(\theta) = \max_{\xi \in \Xi} J_\xi(\theta).
\]

**Proof.** See Appendix A.2.

The next section applies this general result to identify optimal designs in some special cases of the non-regular polynomial regression model above. The two results, Propositions 4 and 5, suggest that there is a de la Garza phenomenon (e.g., de la Garza 1954) in the non-regular case as well, which would be an interesting theoretical topic to pursue in future work.

5 Optimal designs for some non-regular regression models

In this section, we apply the general result in Theorem 2 to identify optimal designs in two important special cases of the polynomial model, namely, linear and quadratic.
Throughout we assume the model stated in (16), namely, that the regression model has non-negative errors with distribution having density of the form (3), with known shape parameter $\alpha \in [1, 2)$.

5.1 Linear model

Consider the linear version of (16), where $g(x, \theta) = \theta_0 + \theta_1 x$. For linear models we have a strong intuition from the regular case as to what the optimal design might be. It turns out that the same result holds in the non-regular case as well.

**Proposition 4.** The optimal design $\xi^*$, according to Definition 2, for the non-regular linear regression model is the symmetric two-point design with weight $\frac{1}{2}$ on $x = \pm A$.

**Proof.** See Section S2.4 in the Supplementary Material.

5.2 Quadratic model

Consider a quadratic case where $g(x, \theta) = \theta_0 + \theta_1 x + \theta_2 x^2$. Here we restrict our attention to the case where the errors $\varepsilon_i$ in the model are exponential, $\alpha = 1$.

**Proposition 5.** For the quadratic model, with $\alpha = 1$ and the balanced design constraint, the optimal design $\xi^*$, according to Definition 2, is one with three distinct points $\{-A, 0, A\}$ with respective weights $\{\frac{1-\pi}{2}, \pi, \frac{1-\pi}{2}\}$ for some $\pi \in (0, 1)$.

**Proof.** See Section S2.5 in the Supplementary Material.

Although the proof of Proposition 5 holds only for the $\alpha = 1$ case, we expect that the result also holds for $\alpha \in [1, 2)$, and the numerical results in Figure 3 (b) support this conjecture. The practical importance is that it simplifies the search over $\Xi_{\text{sym}}$ to a search over the scalar $\pi \in [0, 1]$. The weight at point $\{0\}$ of the optimal design—or the likely optimal design for the case of $\alpha \in (1, 2]$—depends on the value of $A$ and $\alpha$. Based on Proposition 5 and the definition of Hellinger information, the optimal weight can be obtained by solving the optimization problem

$$\pi_A(\alpha) = \arg \max_{\pi \in [0, 1]} f(\pi),$$

(17)

where $f(\pi) = f_{\alpha, A}(\pi)$ is given by

$$f(\pi) = \min_{||u||=1} \left\{ \pi |u_1|^\alpha + \frac{1-\pi}{2} (|u_1 + Au_2 + A^2 u_3|^\alpha + |u_1 - Au_2 + A^2 u_3|^\alpha) \right\}.$$  

This search for the optimal weight, $\pi_A(\alpha)$, along with that over $u$ on the surface of the unit sphere, can be handled numerically.

Figure 4 shows $\alpha \mapsto \pi_A(\alpha)$ for several values of $A$. In particular, we see that the (likely) optimal designs put more weight on 0 as either $\alpha$ or $A$ increases. Our optimal designs for non-regular regression models have a similar format to their E-optimal counterparts in the regular case. That is, a regular E-optimal design for quadratic regression over $[-A, A]$ is given by

$$\{(\pm A, \frac{1-w_A}{2}), (0, w_A), (A, \frac{1-w_A}{2})\},$$

14
Figure 1: Optimal weight $\pi_A(\alpha)$ as a function of $\alpha$ for several $A$ values.

and, for $A$ in $\{1, 1.5, 2\}$, the corresponding values of $w_A$ are $\{0.6, 0.75, 0.81\}$. From Figure 1, as anticipated by Corollary 1, we observe that for $\alpha = 2$, $\pi_A(2)$ matches the weight $w_A$ of the corresponding regular E-optimal design. This is explained by Corollary 1; when $\alpha = 2$, optimal design under Hellinger information is the E-optimal design.

Henceforth, we call the regular E-optimal design counterpart of a non-regular model “regular-optimal.” For the non-regular linear model, based on Proposition 4, the optimal design coincides with the “regular-optimal” design. In the numerical results presented below, we compare optimal designs of non-regular quadratic models to their “regular-optimal” counterparts.

5.3 Numerical results

Here we show some numerical results to demonstrate the efficiency gain in using the proposed optimal designs over other reasonable designs. Recall our model is of the form (16) with non-negative errors having density (3), with known shape parameter $\alpha \in [1, 2)$.

One complication is that currently there are no results that identify an estimator whose risk attains the lower bound in Theorem 1. Consequently, we are currently unable to guarantee that minimizing this lower bound will result in improved estimation for any given estimator. But we do have a reasonable estimator, described next, and the results below do indicate that the design that minimizes the lower bound in Theorem 1 does indeed result in improved efficiency for this particular estimation.

For the class of non-regular polynomial regression problems in consideration here, Smith (1994) proposed an estimator based on solving a linear programming problem: choosing $(\theta_0, \ldots, \theta_p)$ such that $\theta_0$ is maximized subject to the condition that $y_i \geq \sum_{k=1}^{p} \theta_k x_i^k$ for each $i = 1, \ldots, n$. This estimator agrees with the maximum likelihood estimator in the case $\alpha = 1$, has a $O(n^{-1/\alpha})$ convergence rate, which matches the one given by the lower bound in (11), and can be readily computed using the quantreg package in R (Koenker 2013). Moreover, as Smith (1994, p. 174) argues, it is generally superior to maximum likelihood in non-regular cases. For these reasons, comparisons of designs based on this estimator ought to be informative.

Figure 2 presents simulation results on the quality of estimation for the Hellinger optimal design versus 5-, 10-, and 15-point uniform designs for the non-regular linear models, while Figure 3 presents simulation results comparing Hellinger optimal design versus 5-point uniform design and the regular-optimal design. The study proceeds as
follows. For each design space $[-A, A]$ and candidate design, the $n$-vector $y$ is simulated from the corresponding model, with the specified value of $\alpha$ and $\theta$, and then Smith’s estimator $\hat{\theta}$ is computed. Repeat this process 1000 times and compute the Monte Carlo estimate of the risk $R(\hat{\theta}, \theta)$ as usual. This risk is the sum of mean square errors for each component of the parameter vector.

Figure 2 shows that, under different regularity conditions, the optimal design from Proposition 4 is superior in terms of risk. In particular, it is significantly better in the estimation of the slope, $\theta_1$, whereas no design performs significantly better than the others in the estimation of the intercept. The results presented in Figure 3(a) are consistent with Proposition 5 in the case of $\alpha = 1$. In each case, the optimal design performs significantly better than both the 5-point uniform design and the regular-optimal design, despite the similarity of the optimal and regular-optimal designs in terms of weight at point 0. Similarly, Figure 3(b) supports our intuition that Proposition 5 can be extended to cases with $\alpha > 1$. 

Figure 2: Comparison for non-regular linear model, based on $n = 120$ and $\theta = (6, 0.5)$. 

(a) Linear model, $\alpha = 1$

(b) Linear model, $\alpha = 1.4$
(a) Left panel: \( \pi_1(1) = 0.5 \) and the regular-optimal design is \( \{(-1,0.2),(0,0.6),(1,0.2)\} \); Right panel: \( \pi_2(1) = 0.75 \) and the regular-optimal design is \( \{(-2,0.095),(0,0.81),(2,0.095)\} \)

(b) Left panel: \( \pi_{1.5}(1.1) = 0.6 \) and the regular-optimal design is \( \{(-1.5,0.125),(0,0.75),(1.5,0.125)\} \); Right panel: \( \pi_2(1.5) = 0.75 \) and the regular-optimal design is \( \{(-2,0.095),(0,0.81),(2,0.095)\} \)

Figure 3: Comparison for non-regular quadratic model, based on \( n = 120 \) and \( \theta = (2,4,0.8) \).
6 Conclusion

This paper aims to establish a framework for optimal design in the context of non-regular models where the Fisher information matrix does not exist. Towards this goal, we defined an alternative measure of information, based on a local approximation of the squared Hellinger distance between models, suitable for non-regular problems. The proposed Hellinger information has some close connection to the Fisher information when both exist and, more generally, the former has many of the familiar properties of the latter. In particular, in Theorem 1 we establish a parallel to the classical Cramér–Rao inequality which connects our proposed Hellinger information measure to the quality of estimators. This naturally leads to a notion of optimal designs in non-regular problems, i.e., the “optimal design” is one that minimizes the lower bound in Theorem 1.

The proposed optimal design framework introduces a new class of optimization problems to solve, what we have considered here is only the tip of the iceberg. However, the tools currently available in the optimal design literature for regular problems are expected to be useful here. For example, in a particular non-regular polynomial regression setting, we establish a theorem to simplify the numerical and/or analytical search for a particular optimal design, and we apply this general result in the linear and quadratic cases. Developing the theory and computational methods to handle more complex non-regular models, as well as identifying estimators that attain the lower bound \(11\), are interesting topics for future investigation.

Aside from creating a new class of design problems to be investigated, the developments here also shed light on how much our current understanding of design problems depends on the regularity of the models being considered. That is, beyond its value in helping us tackle specific cases in which regularity conditions do not apply, the study of non-regular problems also deepens our understanding of regularity itself and how it affects optimal design. For example, questions about the type of optimality criterion to consider (e.g., A- versus D- versus E-optimal) are apparently only relevant for those regular cases where the Fisher information matrix is exactly or approximately related to the dispersion matrix of an estimator. While this paper provides some important insights about non-regular models and corresponding optimal design problems, there is still much more to be done.

Acknowledgments

The authors are grateful for the helpful suggestions from the Associate Editor referees, and Professor Arkady Shemyakin. The authors also thank Mr. Zhiqiang Ye for pointing out a mistake in the statement and proof of Proposition 1 in a previous version.

A Proofs of theorems

A.1 Proof of Theorem 1

The proof requires a connection between Hellinger distance and risk of an estimator. This first step is based in part on Section I.6 of Ibragimov and Hasminskii (1981), although
our setup and conclusions are more general in certain ways. We summarize this in the following lemma, proved in the Supplementary Material.

**Lemma 1.** For data $Y \in \mathcal{Y}$, consider a model $P_\theta$, with $\mu$-density $p_\theta$, 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 $\psi$, the risk function $R_\psi(T, \theta)$ for the estimator $T$ satisfies

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

For the proof of Theorem 1, start with the squared Hellinger distance between joint distributions $P_\theta^n$ and $P_\vartheta^n$, given by

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

where $h_i(\theta; \vartheta) = H^2(P_i, \theta, P_i, \vartheta)$ is the squared Hellinger distance between individual components. If $\theta$ and $\vartheta$ 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

$$h_n^n(\theta; \vartheta) \leq -2\sum_{i=1}^{n} \log \left( 1 - \frac{h_i(\theta; \vartheta)}{2} \right) \leq 2\sum_{i=1}^{n} h_i(\theta; \vartheta). \quad (18)$$

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

$$h_n^n(\theta; \theta + \varepsilon u) \leq 2J_n(\theta; u) \varepsilon^\alpha + o(n\varepsilon^\alpha), \quad \varepsilon \to 0.$$

When we take $\varepsilon$ equal to $\varepsilon_{n,u} = \{3J_n(\theta; u)\}^{-1/\alpha}$, then we get

$$h_n^n(\theta; \theta + \varepsilon_{n,u} u) \leq \frac{3}{4} + o(1), \quad n \to \infty,$$

where the latter “$o(1)$” conclusion is justified by the assumption (10) about the rate of information accumulation. Therefore, for large enough $n$, with $\vartheta_{n,u} = \theta + \varepsilon_{n,u} u$, $h_n^n(\theta; \vartheta_{n,u}) \leq \frac{3}{4}$, it follows from the above lemma that

$$R_\psi(T_n, \theta) + R_\psi(T_n, \vartheta_{n,u}) \geq \frac{1}{16}\left\| \psi(\theta) - \psi(\vartheta_{n,u}) \right\|^2.$$

Since $\psi$ is differentiable, there is a Taylor approximation at $\theta$:

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

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

Plugging in the definition of \( \varepsilon_{n,u} \) establishes that
\[ R_\psi(T_n, \theta + \varepsilon_{n,u}) + R_\psi(T_n, \theta) \gtrsim \| D_\psi(\theta) u \| \, J_n(\theta; u) - \frac{2}{\alpha}. \]

Also, the constant that has been absorbed in \( \gtrsim \) is \( (32)^{-1}3^{-2/\alpha} \). Finally, the claim (11) follows from the above display and the general fact that, for a function \( f \) defined on a set \( A \), \( f(y_1) + f(y_2) \) is smaller than \( 2 \sup_A f(y) \).

### A.2 Proof of Theorem 2

Take any fixed design \( \xi = \{(w_m, x_m) : m = 1, \ldots, M\} \), and define a function
\[
L(u; x) = J_\xi(\theta; u) = \sum_{m=1}^{M} w_m \left| \sum_{k=0}^{p} x_m^k u_{k+1} \right|^\alpha.
\]

The \( L \) function does not depend on \( \theta \) because it is based on the information in a location parameter problem, but it does depend implicitly on the \( w \) component of the design \( \xi \). From the trivial identity,
\[ a x_m^k = a (-1)^k (-x_m)^k, \quad \text{any } a \in \mathbb{R}, \text{ any } m, \text{ and any } k, \]
it follows immediately that \( L(u; x) = L(v; -x) \), for any unit vector \( u \in \mathbb{R}^{p+1} \), where \( v_{k+1} = (-1)^k u_{k+1} \), \( k = 0, \ldots, p \). Since this new vector \( v \) is also a unit vector, we have
\[ \min_u L(u; x) = \min_v L(v; -x). \]

This implies that the reflected design \( \xi' \)—the one that replaces the original \( x_m \) in \( \xi \) with \(-x_m \), but keeps the same weights—satisfies \( J_\xi(\theta) = J_{\xi'}(\theta) \). Define the mixture design \( \xi^\dagger = \frac{1}{2} \xi + \frac{1}{2} \xi' \), which is symmetric by construction, and by concavity (15) satisfies
\[
J_{\xi^\dagger}(\theta) = \min_u \left\{ \frac{1}{2} J_\xi(\theta; u) + \frac{1}{2} J_{\xi'}(\theta; u) \right\} \\
\geq \frac{1}{2} \min_u J_\xi(\theta; u) + \frac{1}{2} \min_u J_{\xi'}(\theta; u).
\]

We showed above that the two terms in the lower bound are equal and, consequently, \( J_{\xi^\dagger}(\theta) \geq J_\xi(\theta) \). Therefore, for any design \( \xi \) there exists a symmetric design with Hellinger information at least as big; hence, symmetric designs form a complete class.

### References

Bernardo, J.-M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B*, 41(2):113–147. With discussion.

Biedermann, S., Dette, H., and Hoffmann, P. (2009). Constrained optimal discrimination designs for Fourier regression models. *Ann. Inst. Statist. Math.*, 61(1):143–157.
Biedermann, S., Dette, H., and Woods, D. C. (2011). Optimal design for additive partially nonlinear models. *Biometrika*, 98(2):449–458.

Biedermann, S., Dette, H., and Zhu, W. (2006). Optimal designs for dose-response models with restricted design spaces. *J. Amer. Statist. Assoc.*, 101(474):747–759.

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

Cousineau, D. (2009). 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.

de la Garza, A. (1954). Spacing of information in polynomial regression. *Ann. Math. Statistics*, 25:123–130.

Dette, H., Bretz, F., Pepelyshev, A., and Pinheiro, J. (2008). Optimal designs for dose-finding studies. *J. Amer. Statist. Assoc.*, 103(483):1225–1237.

Dette, H., Guchenko, R., Melas, V. B., and Wong, W. K. (2018). Optimal discrimination designs for semiparametric models. *Biometrika*, 105(1):185–197.

Dette, H., Konstantinou, M., and Zhigljavsky, A. (2017). A new approach to optimal designs for correlated observations. *Ann. Statist.*, 45(4):1579–1608.

Dette, H. and Melas, V. B. (2011). A note on the de la Garza phenomenon for locally optimal designs. *Ann. Statist.*, 39(2):1266–1281.

Dette, H., Pepelyshev, A., and Zhigljavsky, A. (2016). Optimal designs in regression with correlated errors. *Ann. Statist.*, 44(1):113–152.

Dette, H. and Schorning, K. (2013). Complete classes of designs for nonlinear regression models and principal representations of moment spaces. *Ann. Statist.*, 41(3):1260–1267.

Dette, H. and Titoff, S. (2009). Optimal discrimination designs. *Ann. Statist.*, 37(4):2056–2082.

Dror, H. A. and Steinberg, D. M. (2006). Robust experimental design for multivariate generalized linear models. *Technometrics*, 48(4):520–529.

Feller, C., Schorning, K., Dette, H., Bermann, G., and Bornkamp, B. (2017). Optimal designs for dose response curves with common parameters. *Ann. Statist.*, 45(5):2102–2132.

Gotwalt, C. M., Jones, B. A., and Steinberg, D. M. (2009). Fast computation of designs robust to parameter uncertainty for nonlinear settings. *Technometrics*, 51(1):88–95.

Harman, R. and Benková, E. (2017). Barycentric algorithm for computing $D$-optimal size- and cost-constrained designs of experiments. *Metrika*, 80(2):201–225.
Hirose, H. and Lai, T. L. (1997). Inference from grouped data in three-parameter Weibull models with applications to breakdown-voltage experiments. *Technometrics*, 39(2):199–210.

Ibragimov, I. A. and Hasminskii, R. Z. (1981). *Statistical Estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York-Berlin. Asymptotic theory, Translated from the Russian by Samuel Kotz.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Ann. Statist.*, 2:849–879.

Koenker, R. (2013). *quantreg: Quantile Regression*. R package version 5.05.

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

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition.

Lin, Y., Martin, R., and Yang, M. (2018). Supplement to “on optimal designs for non-regular models”. DOI...

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.*, 27:986–1005.

López-Fidalgo, J., Tommasi, C., and Trandafir, P. C. (2007). An optimal experimental design criterion for discriminating between non-normal models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(2):231–242.

Pollard, D. (1997). Another look at differentiability in quadratic mean. In *Festschrift for Lucien Le Cam*, pages 305–314. Springer, New York.

Pollard, D. (2005). *Asymptopia*. Chapter 6 on “Hellinger differentiability,” [http://www.stat.yale.edu/~pollard/Courses/607.spring05/handouts/DQM.pdf](http://www.stat.yale.edu/~pollard/Courses/607.spring05/handouts/DQM.pdf).

Sagnol, G. and Harman, R. (2015). Computing exact $D$-optimal designs by mixed integer second-order cone programming. *Ann. Statist.*, 43(5):2198–2224.

Schorning, K., Dette, H., Kettelhake, K., Wong, W. K., and Bretz, F. (2017). Optimal designs for active controlled dose-finding trials with efficacy-toxicity outcomes. *Biometrika*, 104(4):1003–1010.

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

Shemyakin, A. E. (1991). Multidimensional integral inequalities of Rao-Cramér type for parametric families with singularities. *Sibirsk. Mat. Zh.*, 32(4):204–215, 230.

Shemyakin, A. E. (1992). On information inequalities in parametric estimation theory. *Teor. Veroyatnost. i Primenen.*, 37(1):121–123.

22
Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1):67–90.

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

van der Vaart, A. (2002). The statistical work of Lucien Le Cam. *Ann. Statist.*, 30(3):631–682. Dedicated to the memory of Lucien Le Cam.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Waterhouse, T. H., Woods, D. C., Eccleston, J. A., and Lewis, S. M. (2008). Design selection criteria for discrimination/estimation for nested models and a binomial response. *J. Statist. Plann. Inference*, 138(1):132–144.

Woodroofe, M. (1974). Maximum likelihood estimation of translation parameter of truncated distribution. II. *Ann. Statist.*, 2:474–488.

Yang, M. (2010). On the de la Garza phenomenon. *Ann. Statist.*, 38(4):2499–2524.

Yang, M., Biedermann, S., and Tang, E. (2013). On optimal designs for nonlinear models: a general and efficient algorithm. *J. Amer. Statist. Assoc.*, 108(504):1411–1420.

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

Yang, M. and Stufken, J. (2012). Identifying locally optimal designs for nonlinear models: a simple extension with profound consequences. *Ann. Statist.*, 40(3):1665–1681.

Yu, Y. (2011). D-optimal designs via a cocktail algorithm. *Stat. Comput.*, 21(4):475–481.
S Supplementary material

S1. A multi-parameter example

As an illustrative example, consider the case where \( \theta = (\theta_1, \theta_2) \) is two-dimensional and \( P_\theta \) is \( \text{Unif}(\theta_1, \theta_1 + \theta_2) \), where \( \theta_1 \in \mathbb{R} \) and \( \theta_2 > 0 \). It is not difficult to show that

\[
h(\theta, \vartheta) = 2\left\{ 1 - \frac{(\theta_1 + \theta_2) \wedge (\vartheta_1 + \vartheta_2) - \theta_1 \vee \vartheta_1}{(\theta_2 \vartheta_2)^{1/2}} \right\}.
\]

Writing \( \vartheta = \theta + \varepsilon u \) for a unit vector \( u = (u_1, u_2) \), and by considering all the possible configurations of \( u \), it can be shown that \( \alpha = 1 \) and

\[
J(\theta; u) = \theta_2^{-1} g(u),
\]

where \( g(u) \) is a function that depends only on \( u \), not on \( \theta \). This expression is consistent with what we would expect from the familiar Fisher information, since \( \theta_1 \) and \( \theta_2 \) are, in this formulation, location and scale parameters, respectively. The function \( g(u) \) is not complicated, just that the expression varies depending on where on the unit circle \( u \) is. For example, if \( u \) is in the first or third quadrants, then \( g(u) = |2u_1 + u_2| \).

To derive the corresponding expressions for other values of \( u \), the second and fourth quadrants need to be split in half along the line \( y = -x \). Figure 4(a) shows a plot of \( \omega \mapsto J(\theta; u_\omega) \), where \( u_\omega = (\cos \omega, \sin \omega) \), as \( \omega \) varies over \((0, 2\pi)\), for several \( \theta \) values. Notice that the shape of the function does not depend on \( \theta \), only the scale, which means the value of \( u \) that minimizes \( J(\theta; u) \) does not depend on \( \theta \). This plot also reveals the symmetry with respect to reflections \( u \rightarrow -u \) through the origin.

To gain some intuition about the somewhat complicated lower bound established in Theorem 1, suppose that our goal is to estimate the scale parameter \( \theta_2 \), i.e., \( \psi(\theta) = \theta_2 \). Then the relevant Hellinger information is

\[
J^\psi(\theta; u) = \frac{J(\theta; u)}{|u_2|} = \frac{g(u)}{\theta_2 |u_2|}.
\]

Figure 4(b) plots \( \omega \mapsto J^\psi(\theta; u_\omega) \) for three different \( \theta \) values. The minimum value of these functions would be the relevant Hellinger information for estimating \( \theta_2 \), and the horizontal lines drawn there correspond to \( \theta_2^{-1} \). Given independent data \( Y_1, \ldots, Y_n \), an oracle who knows the value of \( \theta_1 \) and uses the maximum likelihood estimator of \( \theta_2 \) when \( \theta_1 \) is fixed and known, would have mean square error of the order \( n^{-2}\theta_2^2 \), which agrees with the theorem’s lower bound up to constants.

S2. Remaining proofs

S2.1. Proof of Proposition 1

Let \( f(y) \) be a function such that \( f(0) = c \) and the density \( p_0 \) in Equation (1) of the main text satisfies

\[
p_0(y) = \beta f(y) y^{\beta-1}, \quad y > 0.
\]
Without loss of generality, consider $\varepsilon > 0$ and small. For the squared Hellinger distance, we have

\[
h(\theta, \theta + \varepsilon) = \int_{-\infty}^{\infty} \left\{ p_{\theta+\varepsilon}^{1/2}(y) - p_\theta^{1/2}(y) \right\}^2 dy \\
= \int_{\theta}^{\theta+\varepsilon} p_\theta(y) dy + \int_{\theta+\varepsilon}^{\infty} \left\{ p_\theta^{1/2}(y) - p_{\theta+\varepsilon}^{1/2}(y) \right\}^2 dy \\
= P_\theta(\theta \leq Y \leq \theta + \varepsilon) + \int_{\theta+\varepsilon}^{\infty} \left\{ p_\theta^{1/2}(y) - p_{\theta+\varepsilon}^{1/2}(y) \right\}^2 dy.
\]

The first term equals $P_0(Z \leq \varepsilon)$ and it is easy to see that this is $f(0)\varepsilon^\beta + o(\varepsilon^\beta)$. For the second term, we make a change of variable, $z = y - \theta - \varepsilon$, so that

\[
\int_{\theta+\varepsilon}^{\infty} \left\{ p_\theta^{1/2}(y) - p_{\theta+\varepsilon}^{1/2}(y) \right\}^2 dy = \int_{0}^{\infty} \left\{ p_0^{1/2}(z + \varepsilon) - p_0^{1/2}(z) \right\}^2 dz.
\]

If we split up this latter integral as

\[
\int_{0}^{\Delta} \left\{ p_0^{1/2}(z + \varepsilon) - p_0^{1/2}(z) \right\}^2 dz + \int_{\Delta}^{\infty} \left\{ p_0^{1/2}(z + \varepsilon) - p_0^{1/2}(z) \right\}^2 dz,
\]

for $\Delta > 0$ as in the statement of the proposition, then the right-most integral is $O(\varepsilon^2) = o(\varepsilon^\beta)$ by Equation (4) in the main text and the dominated convergence theorem. It remains to investigate the left-most integral, which equals

\[
\beta \int_{0}^{\Delta} \left\{ f(z + \varepsilon)^{1/2}(z + \varepsilon)^{(\beta-1)/2} - f(z)^{1/2}z^{(\beta-1)/2} \right\}^2 dz.
\]

We proceed by adding and subtracting $f(z + \varepsilon)^{1/2}z^{(\beta-1)/2}$ inside $\{\cdots\}$, so that the new integrand looks like

\[
\{ f(z + \varepsilon)^{1/2}(z + \varepsilon)^{(\beta-1)/2} - f(z)^{1/2}z^{(\beta-1)/2} \}^2 = \sum_{i=1}^{3} I_i(z; \varepsilon),
\]

Figure 4: Plots of the Hellinger information for the two-dimensional uniform example. The black, red, and green lines correspond to $\theta_2 = 1$, $\theta_2 = 1.5$ and $\theta_2 = 0.75$, respectively.
where

\[ I_1(z; \varepsilon) = f(z + \varepsilon) \{ (z + \varepsilon)^{(\beta - 1)/2} - z^{(\beta - 1)/2} \}^2 \]
\[ I_2(z; \varepsilon) = \{ f(z + \varepsilon)^{1/2} - f(z)^{1/2} \}^2 z^{(\beta - 1)/2} \]
\[ I_3(z; \varepsilon) = 2f(z + \varepsilon)^{1/2} \{ f(z + \varepsilon)^{1/2} - f(z)^{1/2} \} \{ (z + \varepsilon)^{(\beta - 1)/2} - z^{(\beta - 1)/2} \}. \]

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

\[ f(z + \varepsilon)^{1/2} - f(z)^{1/2} \ll \varepsilon, \]

which implies that

\[ \int_0^\Delta I_2(z; \varepsilon) \, dz \ll \varepsilon^2 \int_0^\Delta z^{(\beta - 1)/2} \, dz \ll \varepsilon^2 = o(\varepsilon^\beta), \quad \varepsilon \to 0. \]

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

\[ \int_0^\Delta I_3(z; \varepsilon) \, dz \ll \varepsilon^2 \int_0^\Delta z^{-(3-\beta)/2} \, dz. \]

and, since the integral converges, the upper bound is \( O(\varepsilon^2) = o(\varepsilon^\beta) \) as \( \varepsilon \to 0 \). It remains to deal with the \( I_1 \) term, namely,

\[ \beta \int_0^\Delta f(z + \varepsilon) \{ (z + \varepsilon)^{b} - z^{b} \}^2 \, dz, \quad \text{where} \quad b = \frac{\beta - 1}{2}. \]

Make a change-of-variable, \( w = z/\varepsilon \), so that the above integral becomes

\[ \beta \varepsilon^\beta \int_0^{\Delta/\varepsilon} f(\varepsilon w + \varepsilon) \{ (w + 1)^{b} - w^{b} \}^2 \, dw. \]

By the mean value theorem, we have that \( \{ (w + 1)^{b} - w^{b} \}^2 \leq \min\{1, b^2 w^{2(b-1)}\} \), and, since \( 2(b - 1) = \beta - 3 < -1 \), this upper bound is integrable over \( w \in (0, \infty) \). Since \( f \) is also bounded, it follows from dominated convergence theorem that, as \( \varepsilon \to 0 \),

\[ \int_0^{\Delta/\varepsilon} f(\varepsilon w + \varepsilon) \{ (w + 1)^{b} - w^{b} \}^2 \, dw \to f(0) \int_0^\infty \{ (w + 1)^{b} - w^{b} \}^2 \, dw. \]

The integral on the right-hand side is to be recognized as \( r(\beta) \) in Equation (5) of the main text. Putting everything together, we have

\[ h(\theta, \theta + \varepsilon) = f(0) \{ 1 + \beta r(\beta) \} \varepsilon^\beta + o(\varepsilon^\beta), \]

which implies that the regularity index is \( \alpha = \beta \) and the Hellinger information is

\[ J(\theta) := \lim_{\varepsilon \to 0} \frac{h(\theta, \theta + \varepsilon)}{\varepsilon^\alpha} = f(0) \{ 1 + \beta r(\beta) \}. \]

26
S2.2. Proof of Corollary 1

For notational simplicity, write \( I = I_n(\theta) \) and \( D = D_\psi(\theta) \). Then, for the case \( \alpha = 2 \), it is easy to check that the lower bound in (9) from Theorem 1 is proportional to

\[
\sup_{u: \|u\| = 1} \frac{u^T D^T D u}{u^T I u}.
\]

Let \( I = ELE^T \) be the spectral decomposition of \( I \), and let \( M = D_L^{-1/2} \); note that \( M \) is \( q \times d \) of rank \( q \leq d \). Then the above display equals

\[
\sup_{v: \|v\| = 1} v^T M^T M v = \lambda_{\max}(M^T M).
\]

The matrix \( M^T M \) is non-negative definite, in general, with \( d - q \) eigenvalues equal to zero. But the \( q \) positive eigenvalues of \( M^T M \) coincide with those of the \( q \times q \) positive definite matrix \( MM^T \). And since

\[
MM^T = D_L^{-1/2} L^{-1/2} E^T D^T = D_L^{-1} D^T,
\]

it follows that \( \lambda_{\max}(M^T M) = \lambda_{\max}(MM^T) = \lambda_{\max}(D_L^{-1} D^T) \).

S2.3 Proof of Proposition 3

Let \( \tilde{h}(\eta, \eta') = H^2(g_\eta, g_{\eta'}) \), and recall that \( \eta = g(\theta) \), a function of \( \theta \). Since \( g \) is smooth, a small change in \( \theta \) corresponds to a small change in \( \eta \). In particular,

\[
g(\theta + \varepsilon u) - g(\theta) = \varepsilon \dot{g}(\theta)^T u + o(\varepsilon), \quad \varepsilon \to 0.
\]

Call the above difference \( \delta \), so that \( h(\theta; \theta + \varepsilon u) = \tilde{h}(\eta, \eta + \delta) \) for small \( \varepsilon \). Then

\[
\frac{h(\theta; \theta + \varepsilon u)}{\varepsilon^\alpha} = \frac{\tilde{h}(\eta, \eta + \delta)}{\varepsilon^\alpha} = \frac{\tilde{h}(\eta, \eta + \delta)}{\delta^\alpha} \frac{|\dot{g}(\theta)^T u|^\alpha}{\varepsilon^\alpha} + o(1).
\]

Now let \( \varepsilon \to 0 \) and, therefore, \( \delta \to 0 \), and recall the definition of \( \dot{J}(\eta) \). This proves that \( J(\theta; u) = |\dot{g}(\theta)^T u|^\alpha \dot{J}(g(\theta)) \) as was to be shown.

S2.4 Proof of Proposition 4

Theorem 2 in the paper says that the optimal design must be symmetric. So it suffices to show that the two-point symmetric design with points on the boundary, \( \{(0.5, -A), (0.5, A)\} \), has information at least as big as the maximum among the symmetric designs. That is, we intend to show that

\[
\mathcal{J}_{\{(0.5, -A), (0.5, A)\}}(\theta) \geq \max_{\xi \in \Xi_{\text{sym}}} \mathcal{J}_\xi(\theta).
\]  

Denote \( \mathcal{J}_{\{(0.5, -A), (0.5, A)\}}(\theta; u) \) as the Hellinger information of \( \theta \) in the direction of \( u \) based on design \( \{(0.5, -A), (0.5, A)\} \). For non-regular linear model, \( g(\theta; x) = \theta_0 + \theta_1 x \), the
expressions for $J_{(0, 5, \pm A)}(\theta; u)$ and $J_{\xi^\dagger}(\theta; u)$ for any symmetric design, $\xi^\dagger \in \Xi_{\text{sym}}$, are listed below:

$$J_{(0, 5, \pm A)}(\theta; u) = 0.5(|u_1 + u_2 A|^\alpha + |u_1 - u_2 A|^\alpha),$$

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

Assuming that,

$$J_{(0, 5, \pm A)}(\theta; u) \geq J_{\xi^\dagger}(\theta; u), \quad \text{for all unit vectors } u \text{ and all } \xi^\dagger \in \Xi_{\text{sym}},$$

let $\hat{u} = \arg \min_{u:||u||=1} J_{(0, 5, \pm A)}(\theta; u)$, then following from (20), for any $\xi^\dagger$,

$$\min_{u:||u||=1} J_{(0, 5, \pm A)}(\theta; u) = J_{\xi^\dagger}(\theta; \hat{u}) \geq J_{\xi^\dagger}(\theta; \tilde{u}).$$

Consequently, $\min_{u} J_{(0, 5, \pm A)}(\theta; u) \geq \max_{\xi^\dagger} \min_{u} J_{\xi^\dagger}(\theta; u)$, which implies (19).

To complete the proof, we only need to establish (20). Towards this, $J_{(0, 5, \pm A)}(\theta; u) - J_{\xi^\dagger}(\theta; u)$

$$= 0.5(|u_1 + u_2 A|^\alpha + |u_1 - u_2 A|^\alpha) - \sum_{i=1}^m w_i (|u_1 + u_2 x_i|^\alpha + |u_1 - u_2 x_i|^\alpha)$$

$$= \sum_{i=1}^m w_i (|u_1 + u_2 A|^\alpha + |u_1 - u_2 A|^\alpha - |u_1 + u_2 x_i|^\alpha - |u_1 - u_2 x_i|^\alpha)$$

To evaluate the above expression, first see that one can rewrite any unit vector, except $u = (1, 0)$, in the following format:

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

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

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

then,

$$J_{(0, 5, \pm A)}(\theta; u) - J_{\xi^\dagger}(\theta; u)$$

$$= (1 + B^2)^{-0.5 \alpha} \sum_{i=1}^m w_i (|B + A|^\alpha + |B - A|^\alpha - (|B + x_i|^\alpha - |B - x_i|^\alpha))$$

Based on the expression above, to see if $J_{(0, 5, \pm A)}(\theta; u) - J_{\xi^\dagger}(\theta; u)$ is non-negative or not for all $u, \xi^\dagger$ boils down to checking the sign of $\sum_{i=1}^m w_i (|B + A|^\alpha + |B - A|^\alpha - (|B + x_i|^\alpha + |B - x_i|^\alpha))$ for any $0 \leq x_1, \ldots, x_m \leq A, B \in [0, \infty)$.

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 (three cases in total) defined by relationships between $x_i, A, B$ in location: $x_i \leq A \leq B$, $B \leq x_i \leq A$ and $x_i \leq B \leq A$.

---

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

28
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 $B > 0$,

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

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

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

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 it’s first derivative is always positive,

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

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

\[ (B + A)^\alpha + (A - B)^\alpha - (B + x_i)^\alpha - (x_i - B)^\alpha > 0, i = 1, \ldots, m, \]

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

\[
|B+A|^\alpha + |B-A|^\alpha - (|B+x_i|^\alpha + |B-x_i|^\alpha) \\
=(B + A)^\alpha + (A - B)^\alpha - (B + x_i)^\alpha - (B - x_i)^\alpha \\
=(B + A)^\alpha - (B + x_i)^\alpha + (A - B)^\alpha - (B - x_i)^\alpha. \tag{22}
\]

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 (22) 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 \[ B + x_i = W, \text{ then } B + A = x_i + d_i + B + d_m = W + d_i + d_m. \]

29
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 \( w > 0, d_i > 0 \),
\[
(W + d_i + d_m)^\alpha - (W)^\alpha - (d_i + d_m)^\alpha > 0 \quad \text{and} \quad (d_i + d_m)^\alpha - d_i^\alpha - d_m^\alpha > 0.
\]

Therefore, when \( 0 \leq A - B < B - x_i \)
\[
(B + A)^\alpha - (B + x_i)^\alpha + (A - B)^\alpha - (B - x_i)^\alpha
\]
\[
= (W + d_i + d_m)^\alpha - (W)^\alpha + (d_m)^\alpha - (d_i)^\alpha
\]
\[
> (d_i + d_m)^\alpha + (d_m)^\alpha - (d_i)^\alpha
\]
\[
> d_i^\alpha + d_m^\alpha + (d_m)^\alpha - (d_i)^\alpha
\]
\[
\geq 0.
\]

In summary of all three cases, no matter where \( B \) is in relation to \( x_i \) and \( A \),
\[
|B + A|^\alpha + |B - A|^\alpha - (|B + x_i|^\alpha + |B - x_i|^\alpha) \geq 0 \quad \text{for all } i=1,\ldots,m.
\]

**S2.5. Proof of Proposition 5**

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

Given any symmetric design
\[
\xi = \{(w_1, -x_1), \ldots, (w_m, -x_m), (w_1, x_1), \ldots, (w_m, x_m)\}
\]
and direction vector \( u \), for the non-regular quadratic regression model, the Hellinger information of \( \xi \) in the direction of \( u = (u_1, u_2, u_3) \) has the expression of \( J_{\xi\dagger}(\theta; u) = \sum_{i=1}^{m} w_i(|u_1 + u_2x_i + u_3x_i^2|^\alpha + |u_1 + u_2(-x_i) + u_3x_i^2|^\alpha) \). For simplicity, denote \( f_u(x) = u_1 + u_2x + u_3x^2 \); then, when \( \alpha = 1 \), the above becomes
\[
J_{\xi\dagger}(\theta; u) = \sum_{i=1}^{m} w_i(|f_u(x_i)| + |f_u(-x_i)|).
\]

Let’s assume that there exists \( r_i \in [0, 1] \) such that, for all \( x_i \in [-A, A] \), the following relation is true:
\[
2r_i|f_u(0)| + (1 - r_i)|f_u(A)| + (1 - r_i)|f_u(-A)| > |f_u(x_i)| + |f_u(-x_i)|. \quad (23)
\]

Then, given \( w_i \), after multiplying \( w_i \) on both sides of the inequality (23), we have
\[
2w_ir_i|f(0)| + w_i(1 - r_i)|f(A)| + w_i(1 - r_i)|f(-A)| \geq w_i(|f(x_i)| + |f(-x_i)|). \quad (24)
\]

30
Let \( w = \sum_{i=1}^{m} (1 - r_i)w_i \), based on 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 (24) as 
\[
\xi_w = \{(w, -A), (1 - 2w, 0), (w, A)\}, 0 \leq w \leq 0.5
\]

Hellinger information based on design \( \xi_w \) 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 (24), for any \( u \), for any symmetric design \( \xi^1 \), there is a \( w \) such that 
\[
J_{\xi_w}(\theta; u) \geq J_{\xi^1}(\theta; u).
\]

Now, via Theorem 2 and the exact same argument that established Proposition 4, the conclusion of this proposition holds.

The only step we need in order to complete the proof is to show (23) is true. Notice that \(|f_u(x)| = |f_{-u}(x)|\), i.e. \( |u_1 + u_2x + u_3x^2| = |-u_1 - u_2x - u_3x^2| \). Thus, for every given \( \bar{u} \) with \( \bar{u}_3 < 0 \), there is a \( \hat{u} = -\bar{u} \) such that \(|f_{\hat{u}}(x)| = |f_u(x)|\), and \( f_u(x) \) is convex. Thus, for simplicity, the following only shows (23) 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, (23) can be shown to be true. Here we only consider cases in which \( u \) is such that its \( u_3 \neq 0 \), as 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 \([-B, B]\) for some \( B > 0 \) and \( x_i \in [0, B] \), and there is a \( r_i \in (0, 1) \), such that \( x_i = r_i0 + (1 - r_i)B \), and 
\[
 r_if(0) + (1 - r_i)f(B) > f(x_i) \text{ and } r_if(0) + (1 - r_i)f(-B) > f(-x_i), \]

then
\[
2r_i|f(0)| + (1 - r_i)|f(B)| + (1 - r_i)|f(-B)| > |f(x_i)| + |f(-x_i)|. \tag{25}
\]

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 is at most one root for \( f(x) \), i.e. \( f(x) \geq 0 \) for all \( x \in R \).

The following goes through these cases and shows that (23) is true for each of them. Notice that cases 1 and 2 are equivalent, and cases 5 and 6 are equivalent. So we shall focus on cases 1, 3, 4, 5, and 7.
In case 1 both roots are above \( x \); 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 \([-A, A]\), so by the argument of convexity in [25], [23] is true.

- 1.2) The left root \( x_L \) is below \( A \), i.e. \( x_L < A \). Here, \( f(-x_i), f(x_i), f(-A) > 0 \), which implies that

\[
|f(-x_i)| + |f(x_i)| = 2u_1 + 2u_3x^2, \quad \text{and} \quad f(-A) = u_1 - u_2A + u_3A^2. \tag{26}
\]

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

\[
|f(A)| = -u_1 - u_2A - u_3A^2 > 0, \quad \text{and} \quad -u_2A > u_1 + u_3A^2. \quad \text{Then with (26),}
\]

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

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

\[
|f(A)| = u_1 + u_2A + u_3A^2 > 0. \quad \text{Then with (26),}
\]

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

Then for 1.2) one can find a ratio \( r_A \) such that \( r_A(|f(A)| + f(-A)) > f(x_i) + f(-x_i) \), letting \( r_i = 1 - r_A \), then [23] 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 \([-x_i, x_i]\), so \( f(x) \) 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 \( \alpha = 1 \), [23] 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 \( |f(x)| = -f(x) = -u_1 - u_2x - u_3x^2 \) is concave over \([-x_i, x_i]\). Thus, \( |f(0)| > \frac{1}{2}|f(x_i)| + \frac{1}{2}|f(-x_i)| \), and consequently, [23] holds.

Case 5: \(-x_i \leq x_L \leq x_i \leq x_R\).

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

\[
|f(-A)| = u_1 - u_2A + u_3A^2, \quad |f(-x_i)| = u_1 - u_2x_i + u_3x_i^2. \tag{27}
\]

Based on the set up of case 5, and the possible relations of \( A \) 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. $f(x_i) < 0$ and $f(A) < 0$, so
\[ |f(A)| = -f(A) = -u_1 - u_2 A - u_3 A^2, \text{ and } |f(x_i)| = -f(x_i) = -u_1 - u_2 x_i - u_3 x_i^2. \]
Therefore, with the fact that $-u_2 > 0$, $A \geq x_i$, and \[\text{(27)}, \text{ we have} \]
\[
|f(-A)| + |f(A)| - |f(-x_i)| - |f(x_i)| \\
= -2u_2 A + 2u_2 x_i \\
= -2u_2 (A - x_i) \\
\geq 0.
\]

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 \[\text{(27)}\]
\[
|f(-A)| + |f(A)| - |f(-x_i)| - |f(x_i)| \\
= 2u_1 + 2u_3 A^2 + 2u_2 x_i \\
= 2(u_1 + u_2 A + u_3 A^2) - u_2 (A - x_i) \\
= 2|f(A)| - u_2 (A - x_i) \\
\geq 0.
\]

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 $r_A(|f(A)| + f(-A)) > f(x_i) + f(-x_i)$, letting $r_i = 1 - r_A$, then \[\text{(23)}\] is true, as
\[2r_i|f(0)| + (1 - r_i)|f(A)| + (1 - r_i)|f(-A)| > |f(x_i)| + |f(-x_i)|.\]

- Case 7: There is at most one root, which means, $f(x) \geq 0$ for all $x \in [-A, A]$. Thus, by the argument in \[\text{(25)}, \text{ implies (23)}.\]

In summary of these all seven cases, \[\text{(23)}\] holds.

**S2.6. Proof of Lemma 1**

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:
\[
m_\psi(\theta) - m_\psi(\vartheta) = \int \left[ T(y) - \frac{1}{2} \{m_\psi(\theta) + m_\psi(\vartheta)\} \right] [p_\theta(y) - p_\vartheta(y)] \mu(dy).
\]
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:

$$||m_\psi(\theta) - m_\psi(\vartheta)|| = \left\| \int v_{\theta, \vartheta}(p_\theta - p_\vartheta) \, dy \right\|$$

$$\leq \int \|v_{\theta, \vartheta}\| |p_\theta^{1/2} + p_\vartheta^{1/2}| |p_\theta^{1/2} - p_\vartheta^{1/2}| \, dy.$$  

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

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

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

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

An analogous bound holds for $\int \|v_{\theta, \vartheta}\|^2 p_\vartheta \, dy$, yielding the expression

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

Rearranging terms gives the bound

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

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

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

- 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.