ON OPTIMAL DESIGNS FOR NONREGULAR MODELS

BY YI LIN[*], RYAN MARTIN[†] AND MIN YANG[*]

University of Illinois at Chicago[*] *and North Carolina State University*[†]

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 nonregular 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 nonregular optimal design problems. Hellinger optimal designs are derived for several nonregular regression problems, with numerical results empirically demonstrating the efficiency of these designs compared to alternatives.

1. Introduction. Optimal experimental design is a classical problem with substantial recent developments. For example, Biedermann, Dette and Zhu (2006), Dette et al. (2008), Feller et al. (2017) and Schorning et al. (2017) studied optimal designs for dose-response models; Dette, Pepelyshev and Zhigljavsky (2016) and Dette, Konstantinou and Zhigljavsky (2017) investigated optimal designs for correlated observations; Dror and Steinberg (2006) and Gotwalt, Jones and Steinberg (2009) studied robustness issues in optimal designs; López-Fidalgo, Tommasi and Trandafir (2007), Waterhouse et al. (2008), Biedermann, Dette and Hoffmann (2009), Dette and Titoff (2009), and Dette et al. (2018) studied optimal discrimination designs; Biedermann, Dette and Woods (2011) studied optimal design for additive partially nonlinear models; Yu (2011), Yang, Biedermann and Tang (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), Hirose and Lai (1997), Cousineau (2009)). In particular, Smith (1994) describes a class of nonregular linear regression models,

$$y = x^\top \theta + \varepsilon,$$

where the error ε is nonnegative, which implies a nonregular model for y, given x, since its distribution has θ-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 nonregular problems.

Toward formulating a design problem in a nonregular 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 noninformative priors in the Bayesian context, thereby backtracking the path taken by Lindley (1956) and Bernardo (1979) from information in an experiment to noninformative priors. In particular, Shemyakin (2014) proposes an alternative to Fisher information and generalizes the noninformative 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 multiparameter 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_θ and P_ϑ, 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 nonregular 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 θ, but not differentiable, so important features of the local approximation will generally depend on both the magnitude and the direction of the departure of ϑ from θ. 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 nonregular 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 nonregular 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 nonregular 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 nonregular models in Section 4. Some specific nonregular 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 the Appendix; the remaining details are given in the Supplementary Material (Lin, Martin and Yang (2019)).

2. Review of information in regular models.

The proposed model assumes that the $\mathbb{Y}$-valued observations $Y_1, \ldots, Y_n$ are independent, and the marginal distribution of Y_i is $P_{i,\theta}$, where θ 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 θ 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 σ-finite measure μ. When the index i is not important, and there is no risk of confusion, we will drop the index and write simply p_θ for the density function with respect to μ.

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 θ-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, ℓ) entry is given by

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

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 $\dot{\ell}_\theta$, typically the gradient of $\log p_\theta$, taking values in $\mathbb{R}^d$, such that

$$\int \left(p_{\theta+\varepsilon}^{1/2} - p_\theta^{1/2} - \frac{1}{2} \varepsilon^\top \dot{\ell}_\theta p_\theta^{1/2} \right)^2 d\mu = o(\|\varepsilon\|^2), \qquad \varepsilon \to 0,$$

where $\| \cdot \|$ denotes the ℓ_2-norm. Then the Fisher information matrix exists and is given by the formula $I(\theta) = \int \dot{\ell}_\theta \dot{\ell}_\theta^\top p_\theta \, d\mu$. If we let $H(P_\theta, P_\vartheta)$, denote the Hellinger distance and define h as

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

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 nonregular 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 nonregular cases and, unfortunately, these differ from their regular counterparts in several fundamental ways. First, the Fisher information is not well defined in nonregular 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), page 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 nonregular 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 nonregular problems.

3. Information in nonregular 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 θ, the limit $J(\theta) = \lim_{\varepsilon \to 0} |\varepsilon|^{-\alpha} h(\theta; \theta + \varepsilon)$ exists, is finite and nonzero. If such an α exists, then it must be unique; but there are cases where existence fails, for example, when θ is not identifiable, so that $h(\theta, \theta + \varepsilon) \equiv 0$ for all sufficiently small ε. The case $\alpha = 2$ corresponds to differentiable in quadratic mean and, hence, "regular," while $\alpha \in (0, 2)$ corresponds to "nonregular." Differentiability of $\vartheta \mapsto H(P_\theta, P_\vartheta)$ or lack thereof determines a model's regularity, so the largest value α can take is 2; otherwise, the limit is infinite. From the above limit, there is a local approximation,

$$(2) \qquad h(\theta; \vartheta) = J(\theta)|\theta - \vartheta|^\alpha + o(|\theta - \vartheta|^\alpha).$$

This resembles the local Hölder condition considered in Ibragimov and Has'minskiĭ (1981), Section I.6. We call α 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 $P_\theta = \mathsf{Unif}(0, \theta)$, $\theta > 0$, then $J(\theta) = \theta^{-1}$.
- If $P_\theta = \mathsf{Unif}(\theta^{-1}, \theta)$, $\theta > 1$, then $J(\theta) = (\theta^2 + 1)\{\theta(\theta^2 - 1)\}^{-1}$.
- If $P_\theta = \mathsf{Unif}(\theta; \theta^2)$, $\theta > 1$, then $J(\theta) = (2\theta + 1)\{\theta(\theta - 1)\}^{-1}$.

A class of nonregular 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

$$(3) \qquad p_0(y) = \beta c y^{\beta-1} \qquad \text{as } y \to 0,$$

where $\beta \geq 1$ and $c = c(\beta) \in (0, \infty)$. For example, the gamma and Weibull families, with shape parameter β and scale σ, have $c = \{\beta \sigma^\beta \Gamma(\beta)\}^{-1}$ and $c = \sigma^{-\beta}$, respectively. The next result identifies the regularity index α and the Hellinger information $J(\theta)$ for this class of location parameter problems, with $p_\theta(y) = p_0(y - \theta)$. It shows that α need not be an integer and the Hellinger information, like Fisher's, is constant in location models.

PROPOSITION 1. *Let p_0 satisfy* (3) *with $\beta \in [1, 2)$. If, for some $\Delta > 0$,*

$$(4) \qquad \int_\Delta^\infty \left(\frac{d}{dy} \log p_0(y)\right)^2 p_0(y)\, dy < \infty,$$

then $\alpha = \beta$ and $J(\theta) \equiv c\{1 + \beta\, r(\beta)\}$, where c is as in (3) *and*

$$(5) \qquad r(\beta) = \int_0^\infty \{(w + 1)^{(\beta-1)/2} - w^{(\beta-1)/2}\}^2\, dw.$$

PROOF. See Section S2.1 in the Supplementary Material. □

Ibragimov and Has'minskiĭ (1981), Theorem VI.1.1, show that $h(\theta; \theta + \varepsilon) = O(|\varepsilon|^\alpha)$ as $\varepsilon \to 0$, but they do not identify the constant c. Similar results have appeared elsewhere in the literature on nonregular 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, nonregular multiparameter case, where Θ is an open subset of $\mathbb{R}^d$, defining Hellinger information requires some additional effort. In particular, nonregularity 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 ϑ approaches θ, 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 Θ 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 ∞:

$$\text{(6)} \qquad \lim_{\varepsilon \to 0} \frac{h(\theta; \theta + \varepsilon u)}{|\varepsilon|^{\alpha}} = J(\theta; u).$$

Then, the following local approximation holds:

$$\text{(7)} \qquad h(\theta; \theta + \varepsilon u) = J(\theta; u)|\varepsilon|^{\alpha} + o(|\varepsilon|^{\alpha}), \qquad \varepsilon \to 0.$$

This defines the *index of regularity* α and the *Hellinger information* $J(\theta; u)$ at θ 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, that is, $u = \pm 1$, 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 α 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 $P_\theta = \mathsf{Unif}(\theta_1, \theta_1 + \theta_2)$, where $\theta_1 \in \mathbb{R}$ and $\theta_2 > 0$. In this form, θ_1 and θ_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 S1 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 nonregular case (see Remark 2), when the model is regular, that is, 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 θ and on the direction u of departure from θ.

PROPOSITION 2. *For a regular model, with $\alpha = 2$, if $I(\theta)$ denotes the $d \times d$ Fisher information matrix, then $J(\theta; u) = \frac{1}{4} u^\top I(\theta) u$.*

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 θ is an unknown parameter taking values in an open subset Θ of $\mathbb{R}^d$ for some $d \geq 1$. Let P_θ^n 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 θ, in the direction of u, based on the sample of size n, as

$$(8) \qquad \mathcal{J}_n(\theta; u) = \sum_{i=1}^n J_i(\theta; u),$$

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 1 below will establish a suitable connection between $\mathcal{J}_n(\theta; u)$ and the quality of an estimator, and this will provide the necessary foundation for defining optimal designs for nonregular 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

$$(9) \qquad R_\psi(T_n, \theta) = E_\theta^n \| T_n - \psi(\theta) \|^2,$$

the q-vector version of mean square error, where expectation, E_θ^n, is with respect to P_θ^n. This covers the case where $\psi(\theta) = \theta$ and $q = d$, so that interest is in the full parameter θ, and the case where $\psi(\theta)$ is a single component of θ 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 $\mathcal{J}_n(\theta; u)$ 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 $\varepsilon_{n,u} = \{3\mathcal{J}_n(\theta; u)\}^{-1/\alpha}$, and*

$$(10) \qquad \lim_{n \to \infty} \inf_u [n^{-1} \mathcal{J}_n(\theta; u)] > 0,$$

then, for all large n,

$$(11) \qquad \inf_{T_n} \sup_{\vartheta \in B_n(\theta)} R_\psi(T_n, \vartheta) \gtrsim \left[\inf_u \{ \| D_\psi(\theta) u \|^{-\alpha} \mathcal{J}_n(\theta; u) \} \right]^{-2/\alpha},$$

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 "$\gtrsim$" 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 θ is that a lucky choice of $T_n \equiv \psi(\theta)$ has excellent performance at θ, but poor performance at a nearby ϑ. The theorem basically says that, if one looks at a locally uniform measure of risk, which prevents "cheating" toward or luck at a particular θ, 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 $\mathcal{J}_n(\theta; u) = n J_1(\theta; u)$, then it follows that the lower bound is of order $n^{-2/\alpha}$, which agrees with the known minimax rate for estimators in nonregular models (Ibragimov and Has'minskiĭ (1981), Section 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 θ, 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 $\mathcal{I}_n(\theta)$ is the positive definite $d \times d$ Fisher information matrix, then the lower bound in (11) is proportional to*

$$\lambda_{\max}\{D_\psi(\theta)\mathcal{I}_n(\theta)^{-1}D_\psi(\theta)^\top\},$$

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

Y. LIN, R. MARTIN AND M. YANG

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

$$(12) \qquad \lambda_{\max}\{\mathcal{I}_n(\theta)^{-1}\} = \lambda_{\min}^{-1}\{\mathcal{I}_n(\theta)\}.$$

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 $\mathcal{J}_n(\theta; u)$ 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 nonregular model as formulated above, with index of regularity $\alpha \in (0, 2]$, we set the direction-free Hellinger information at θ, for interest parameter $\psi(\theta)$, as

$$(13) \qquad \mathcal{J}_n^{\psi}(\theta) = \inf_u\{\|D_{\psi}(\theta)u\|^{-\alpha} \mathcal{J}_n(\theta; u)\}.$$

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

$$(14) \qquad \mathcal{J}_n(\theta) = \inf_u \mathcal{J}_n(\theta; u).$$

Moreover, in the regular case with $\alpha = 2$, it follows from Corollary 1 and, in particular, (12), that $\mathcal{J}_n(\theta)$ 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 α to depend on u, so each component of θ, treated individually, must have the same index of regularity. To see this, consider an exponential distribution with location and rate parameters θ_1 and θ_2, respectively. If θ_1 was fixed and only θ_2 was unknown, then it is a regular problem and the above definition would hold with $\alpha = 2$. Similarly, if θ_2 was fixed and only θ_1 was unknown, then the definition holds with $\alpha = 1$ according to Proposition 1. However, if both θ_1 and θ_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 θ, 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 multiparameter models. Shemyakin defines a "Hellinger information matrix" for nonregular 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 θ, so if such a matrix could describe the local behavior, then it ought to be nonnegative definite. However, Shemyakin ((2014), page 931) admits that a general nonnegative definiteness result has not been established for his information matrix. Without a nonnegative definiteness result for his Hellinger information matrix, lower bounds like those in, for example, 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 $\mathcal{J}_n(\theta; u) = \sum_{i=1}^{n} J_i(\theta; u)$, 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_θ^n and P_ϑ^n, assuming independence. A straightforward calculation reveals

$$H^2(P_\theta^n, P_\vartheta^n) = 2 - 2 \prod_{i=1}^{n} \int \{p_{i,\theta}(y_i) p_{i,\vartheta}(y_i)\}^{1/2} \, dy_i$$

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

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

$$H^2(P_\theta^n, P_\vartheta^n) \approx \sum_{i=1}^{n} H^2(P_{i,\theta}, P_{i,\vartheta}).$$

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 $\mathsf{Unif}(0, \theta)$ 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 nonregular models are

consistent with our lower bound, but, to our knowledge, more precise results concerning the asymptotic behavior of estimators in nonregular 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 nonregular 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 nonregular 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 nonregular. This section defines what we mean by an optimal design for nonregular 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_{η_i}, for $i = 1, \ldots, n$. That is, each Y_i has its own parameter η_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 η_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 η_i to a common θ, we obtain the setup from previous sections, that is, $Y_i \sim p_{i,\theta}$, 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_η be a density function depending on a scalar parameter η, 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_θ, for $\theta \in \Theta \subseteq \mathbb{R}^d$, as $q_{g(\theta)}$ where $g : \Theta \to \mathbb{R}$ is a smooth function with nonvanishing gradient $\dot{g}$. Then p_θ also has index of regularity α, and the corresponding Hellinger information at θ, in the direction of u, is*

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

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, that is, $Y_i \sim p_{i,\theta} = q_{g_i(\theta)}$, the Hellinger information at θ, in direction of u, is

$$\mathcal{J}_n(\theta; u) = \sum_{i=1}^{n} |\dot{g}_i(\theta)^\top u|^\alpha \tilde{J}(g_i(\theta)).$$

For the special case where $g_i(\theta) = g(x_i, \theta)$ for covariates x_i, it is clear that $\mathcal{J}_n(\theta; u)$ depends on $x_1, \ldots, x_n$. For example, if $Y_1, \ldots, Y_n$ are independent, with $Y_i \sim g(x_i, \theta) + \mathsf{Gamma}(\beta, 1)$, where $g(x, \theta) = \theta_0 + \sum_{k=1}^{p} \theta_k x^{k+1}$, then it follows from Propositions 1 and 3 that

$$\mathcal{J}_n(\theta; u) = \frac{1 + \beta r(\beta)}{\beta \Gamma(\beta)} \sum_{i=1}^{n} \left| \sum_{k=0}^{p} x_i^k u_{k+1} \right|^\beta.$$

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 ξ 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 ξ 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 $\mathcal{J}_\xi(\theta; u)$ to indicate that the Hellinger information in the direction u depends on the design ξ through the specific covariate values. For example, given design $\xi = \{(w_i, x_i) : i = 1, \ldots, m\}$, then $\mathcal{J}_\xi(\theta; u) = \sum_{i=1}^{m} w_i J_i(\theta; u)$, 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 ξ is defined as

$$\mathcal{J}_\xi(\theta) = \inf_u \mathcal{J}_\xi(\theta; u).$$

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

DEFINITION 2. Under the nonregular model setup presented above, the optimal design $\xi^\star$ is one which maximizes the Hellinger information, that is,

$$\xi^\star = \arg\max_{\xi} \mathcal{J}_\xi(\theta).$$

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 nonregular 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, nonregularity will cause sampling distributions of estimators to be nonellipsoidal, 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 nonregular cases. There are some similarities to the regular case, in particular, the Hellinger information is nonnegative and additive like Fisher information. Also, the map $\xi \mapsto \mathcal{J}_\xi(\theta)$ is concave for fixed θ, that is, for any two designs ξ and ξ' and any $w \in [0, 1]$,

$$(15) \qquad \mathcal{J}_{w\xi+(1-w)\xi'}(\theta) \geq w\mathcal{J}_\xi(\theta) + (1-w)\mathcal{J}_{\xi'}(\theta),$$

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 nonregular polynomial models.* Motivated by the setup in Smith (1994), we consider a nonregular model of the form

$$(16) \qquad y_i = g(x_i, \theta) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

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 ε_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 (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 ξ given by

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

that is, 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*

$$\max_{\xi \in \Xi_{\text{sym}}} \mathcal{J}_\xi(\theta) = \max_{\xi \in \Xi} \mathcal{J}_\xi(\theta).$$

PROOF.   See Appendix A.2.   □

The next section applies this general result to identify optimal designs in some special cases of the nonregular 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 nonregular case as well, which would be an interesting theoretical topic to pursue in future work.

5. Optimal designs for some nonregular 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 nonnegative 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 nonregular case as well.

PROPOSITION 4.   *The optimal design $\xi^\star$, according to Definition 2, for the nonregular 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 ε_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 $\xi^\star$, 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 Ξ_{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 α. Based on Proposition 5 and the definition of Hellinger information, the optimal weight can be obtained by solving the optimization problem

$$(17) \qquad \pi_A(\alpha) = \arg \max_{\pi \in [0, 1]} f(\pi),$$

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 1 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 α or A increases. Our optimal designs for nonregular 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

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

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
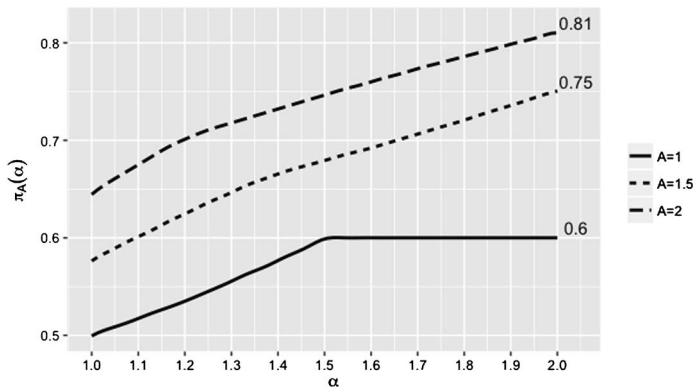

FIG. 1.   *Optimal weight $\pi_A(\alpha)$ as a function of α for several A values.*

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 nonregular model "regular-optimal." For the nonregular 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 nonregular 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 nonnegative 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 nonregular 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 θ_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), page 174) argues, it is generally superior to maximum likelihood in nonregular 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 nonregular 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 α and θ, 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
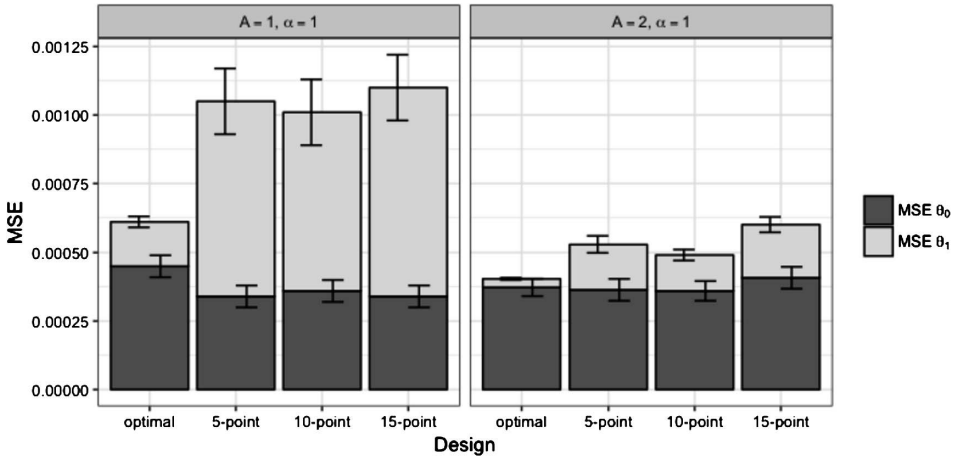
(a) Linear model, $\alpha = 1$
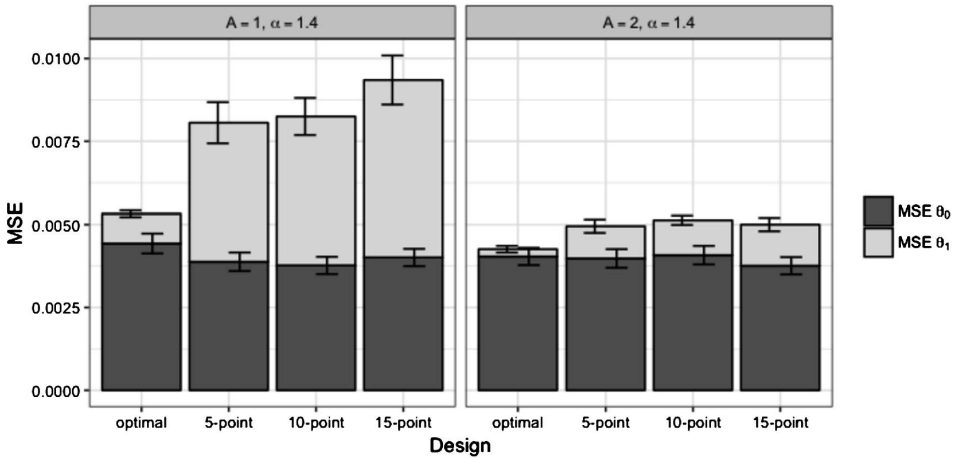
(b) Linear model, $\alpha = 1.4$

FIG. 2.   *Comparison for the nonregular linear model, based on $n = 120$ and $\theta = (6, 0.5)$.*

better in the estimation of the slope, θ_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 $\alpha > 1$.

6. Conclusion.   This paper aims to establish a framework for optimal design in the context of nonregular models where the Fisher information matrix does not
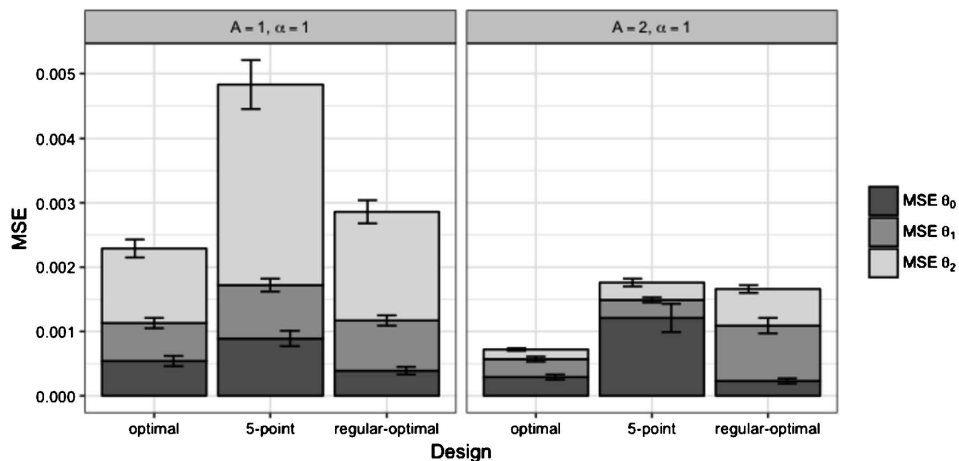
(a) Left panel: $\pi_1(1) = 0.5$ and the regular-optimal design is $\{(-1, 0.2), (0, 0.6), (1, 0.2)\}$; Right panel: $\pi_2(1) = 0.75$ and the regular-optimal design is $\{(-2, 0.095), (0, 0.81), (2, 0.095)\}$
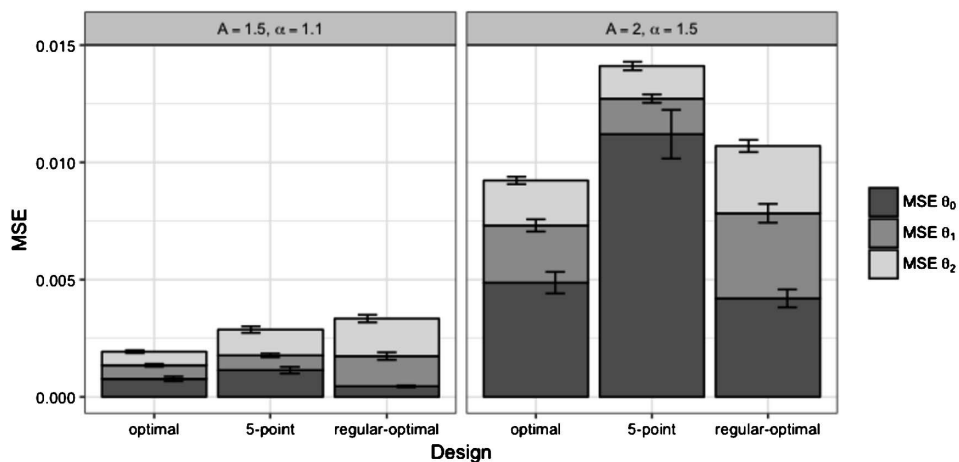
(b) Left panel: $\pi_{1.5}(1.1) = 0.6$ and the regular-optimal design is $\{(-1.5, 0.125), (0, 0.75), (1.5, 0.125)\}$; Right panel: $\pi_2(1.5) = 0.75$ and the regular-optimal design is : $\{(-2, 0.095), (0, 0.81), (2, 0.095)\}$

FIG. 3. *Comparison for nonregular quadratic model, based on $n = 120$ and $\theta = (2, 4, 0.8)$.*

exist. Toward this goal, we defined an alternative measure of information, based on a local approximation of the squared Hellinger distance between models, suitable for nonregular 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 nonregular problems, that is, 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 nonregular 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 nonregular 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 nonregular 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 nonregular models and corresponding optimal design problems, there is still much more to be done.

APPENDIX: 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 Has'minskiĭ (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 $Y \in \mathbb{Y}$, consider a model P_θ, with μ-density p_θ, indexed by a parameter $\theta \in \Theta \subseteq \mathbb{R}^d$. Let $\psi = \psi(\theta)$ be the interest parameter, where $\psi : \mathbb{R}^d \to \mathbb{R}^q$. For an estimator $T = T(Y)$ of ψ, the risk function $R_\psi(T, \theta)$ for the estimator T satisfies*

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

For the proof of Theorem 1, start with the squared Hellinger distance between joint distributions P_θ^n and P_ϑ^n, given by

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

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

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

it follows that

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

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

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

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

$$h^n(\theta; \theta + \varepsilon_{n,u}u) \leq \frac{2}{3} + o(1), \qquad n \to \infty,$$

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

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

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

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

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

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

and, hence,

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

Plugging in the definition of $\varepsilon_{n,u}$ establishes that

$$R_\psi(T_n, \theta + \varepsilon_{n,u}u) + R_\psi(T_n, \theta) \gtrsim \|D_\psi(\theta)u\|^2 \mathcal{J}_n(\theta; u)^{-2/\alpha}.$$

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

$$L(u; x) = \mathcal{J}_\xi(\theta; u) = \sum_{m=1}^{M} w_m \left| \sum_{k=0}^{p} x_m^k u_{k+1} \right|^\alpha.$$

The L function does not depend on θ because it is based on the information in a location parameter problem, but it does depend implicitly on the w component of the design ξ. From the trivial identity,

$$a x_m^k = a(-1)^k (-x_m)^k \qquad \text{any } a \in \mathbb{R}, \text{ any } m, \text{ and any } k,$$

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 ξ'—the one that replaces the original x_m in ξ with $-x_m$, but keeps the same weights—satisfies $\mathcal{J}_\xi(\theta) = \mathcal{J}_{\xi'}(\theta)$. Define the mixture design $\xi^\dagger = \frac{1}{2}\xi + \frac{1}{2}\xi'$, which is symmetric by construction, and by concavity (15) satisfies

$$\mathcal{J}_{\xi^\dagger}(\theta) = \min_u \left\{ \frac{1}{2}\mathcal{J}_\xi(\theta; u) + \frac{1}{2}\mathcal{J}_{\xi'}(\theta; u) \right\}$$

$$\geq \frac{1}{2}\min_u \mathcal{J}_\xi(\theta; u) + \frac{1}{2}\min_u \mathcal{J}_{\xi'}(\theta; u).$$

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

SUPPLEMENTARY MATERIAL

Further technical details and proofs (DOI: 10.1214/18-AOS1780SUPP; .pdf). This file contains details concerning the two-parameter uniform example on page 3341 along with the remaining proofs.

REFERENCES

BERNARDO, J.-M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B* **41** 113–147. MR0547240

BIEDERMANN, S., DETTE, H. and HOFFMANN, P. (2009). Constrained optimal discrimination designs for Fourier regression models. *Ann. Inst. Statist. Math.* **61** 143–157. MR2481031

BIEDERMANN, S., DETTE, H. and WOODS, D. C. (2011). Optimal design for additive partially nonlinear models. *Biometrika* **98** 449–458. MR2806440

BIEDERMANN, S., DETTE, H. and ZHU, W. (2006). Optimal designs for dose-response models with restricted design spaces. *J. Amer. Statist. Assoc.* **101** 747–759. MR2256185

CHERNOZHUKOV, V. and HONG, H. (2004). Likelihood estimation and inference in a class of nonregular econometric models. *Econometrica* **72** 1445–1480. MR2077489

COUSINEAU, D. (2009). Fitting the three-parameter Weibull distribution: Review and evaluation of existing and new methods. *IEEE Trans. Dielectr. Electr. Insul.* **16** 281–288.

DETTE, H., KONSTANTINOU, M. and ZHIGLJAVSKY, A. (2017). A new approach to optimal designs for correlated observations. *Ann. Statist.* **45** 1579–1608. MR3670189

DETTE, H. and MELAS, V. B. (2011). A note on the de la Garza phenomenon for locally optimal designs. *Ann. Statist.* **39** 1266–1281. MR2816354

DETTE, H., PEPELYSHEV, A. and ZHIGLJAVSKY, A. (2016). Optimal designs in regression with correlated errors. *Ann. Statist.* **44** 113–152. MR3449764

DETTE, H. and SCHORNING, K. (2013). Complete classes of designs for nonlinear regression models and principal representations of moment spaces. *Ann. Statist.* **41** 1260–1267. MR3113810

DETTE, H. and TITOFF, S. (2009). Optimal discrimination designs. *Ann. Statist.* **37** 2056–2082. MR2533479

DETTE, H., BRETZ, F., PEPELYSHEV, A. and PINHEIRO, J. (2008). Optimal designs for dose-finding studies. *J. Amer. Statist. Assoc.* **103** 1225–1237. MR2462895

DETTE, H., GUCHENKO, R., MELAS, V. B. and WONG, W. K. (2018). Optimal discrimination designs for semiparametric models. *Biometrika* **105** 185–197. MR3768873

DE LA GARZA, A. (1954). Spacing of information in polynomial regression. *Ann. Math. Stat.* **25** 123–130. MR0060777

DROR, H. A. and STEINBERG, D. M. (2006). Robust experimental design for multivariate generalized linear models. *Technometrics* **48** 520–529. MR2328621

FELLER, C., SCHORNING, K., DETTE, H., BERMANN, G. and BORNKAMP, B. (2017). Optimal designs for dose response curves with common parameters. *Ann. Statist.* **45** 2102–2132. MR3718163

GOTWALT, C. M., JONES, B. A. and STEINBERG, D. M. (2009). Fast computation of designs robust to parameter uncertainty for nonlinear settings. *Technometrics* **51** 88–95. MR2664866

HARMAN, R. and BENKOVÁ, E. (2017). Barycentric algorithm for computing *D*-optimal size- and cost-constrained designs of experiments. *Metrika* **80** 201–225. MR3597582

HIROSE, H. and LAI, T. L. (1997). Inference from grouped data in three-parameter Weibull models with applications to breakdown-voltage experiments. *Technometrics* **39** 199–210. MR1452348

IBRAGIMOV, I. A. and HAS'MINSKIĬ, R. Z. (1981). *Statistical Estimation. Asymptotic Theory. Applications of Mathematics* **16**. Springer, New York–Berlin. Translated from the Russian by Samuel Kotz. MR0620321

KIEFER, J. (1974). General equivalence theory for optimum designs (approximate theory). *Ann. Statist.* **2** 849–879. MR0356386

KOENKER, R. (2013). quantreg: Quantile regression. R package version 5.05. Available at http://CRAN.R-project.org/package=quantreg.

KOENKER, R. and HALLOCK, K. (2001). Quantile regression: An introduction. *J. Econ. Perspect.* **15** 43–56.

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR1639875

LIN, Y., MARTIN, R. and YANG, M. (2019). Supplement to "On optimal designs for nonregular models." DOI:10.1214/18-AOS1780SUPP.

LINDLEY, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Stat.* **27** 986–1005. MR0083936

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** 231–242. MR2325274

POLLARD, D. (1997). Another look at differentiability in quadratic mean. In *Festschrift for Lucien Le Cam* 305–314. Springer, New York. MR1462953

POLLARD, D. (2005). Asymptopia. In *Hellinger Differentiability*, Chapter 6, 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** 2198–2224. MR3396983

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** 1003–1010. MR3737318

SHEMYAKIN, A. E. (1991). Multidimensional integral inequalities of Rao–Cramér type for parametric families with singularities. *Sibirsk. Mat. Zh.* **32** 204–215, 230. MR1142080

SHEMYAKIN, A. E. (1992). On information inequalities in parametric estimation theory. *Teor. Veroyatn. Primen.* **37** 121–123. MR1211210

SHEMYAKIN, A. (2014). Hellinger distance and non-informative priors. *Bayesian Anal.* **9** 923–938. MR3293962

SMITH, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72** 67–90. MR0790201

SMITH, R. L. (1994). Nonregular regression. *Biometrika* **81** 173–183. MR1279665

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

VAN DER VAART, A. (2002). The statistical work of Lucien Le Cam. *Ann. Statist.* **30** 631–682. Dedicated to the memory of Lucien Le Cam. MR1922537

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** 132–144. MR2369620

WOODROOFE, M. (1974). Maximum likelihood estimation of translation parameter of truncated distribution. II. *Ann. Statist.* **2** 474–488. Collection of articles dedicated to Jerzy Neyman on his 80th birthday. MR0356343

YANG, M. (2010). On the de la Garza phenomenon. *Ann. Statist.* **38** 2499–2524. MR2676896

YANG, M., BIEDERMANN, S. and TANG, E. (2013). On optimal designs for nonlinear models: A general and efficient algorithm. *J. Amer. Statist. Assoc.* **108** 1411–1420. MR3174717

YANG, M. and STUFKEN, J. (2009). Support points of locally optimal designs for nonlinear models with two parameters. *Ann. Statist.* **37** 518–541. MR2488361

YANG, M. and STUFKEN, J. (2012). Identifying locally optimal designs for nonlinear models: A simple extension with profound consequences. *Ann. Statist.* **40** 1665–1681. MR3015039

YU, Y. (2011). D-optimal designs via a cocktail algorithm. *Stat. Comput.* **21** 475–481. MR2826686

Y. LIN
M. YANG
DEPARTMENT OF MATHEMATICS, STATISTICS,
    AND COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT CHICAGO
851 S. MORGAN ST.
CHICAGO, ILLINOIS 60607
USA
E-MAIL: yilin2718@gmail.com
        myang2@uic.edu

R. MARTIN
DEPARTMENT OF STATISTICS
NORTH CAROLINA STATE UNIVERSITY
2311 STINSON DR.
RALEIGH, NORTH CAROLINA 27695
USA
E-MAIL: rgmarti3@ncsu.edu