Optimal Design of Experiments on Riemannian Manifolds

Hang Li\(^a\) and Enrique Del Castillo\(^a,b\)

\(^a\)Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA; \(^b\)Department of Statistics, Pennsylvania State University, University Park, PA

ABSTRACT
The theory of optimal design of experiments has been traditionally developed on an Euclidean space. In this article, new theoretical results and an algorithm for finding the optimal design of an experiment located on a Riemannian manifold are provided. It is shown that analogously to the results in Euclidean spaces, D-optimal and G-optimal designs are equivalent on manifolds, and we provide a lower bound for the maximum prediction variance of the response evaluated over the manifold. In addition, a converging algorithm that finds the optimal experimental design on manifold data is proposed. Numerical experiments demonstrate the importance of considering the manifold structure in a designed experiment when present, and the superiority of the proposed algorithm. Supplementary materials for this article are available online.

1. Introduction

Supervised learning models typically need to be trained on large amounts of labeled instances to perform well. While many modern systems can easily produce a large number of unlabeled instances at low cost, the labeling process can be very difficult, expensive or time-consuming. For example, audio data require experienced linguists to spend much longer time than the audio itself to precisely annotate the speech utterances. Given a learning model, nonidentical labeled instances contain different amounts of information and contribute to the learning process in different ways. Therefore, an interesting and practical question arises: how to choose the most informative instances to label so that one can improve the learning rate of the model and reduce the labeling cost at the same time?

In statistics, the problem of selecting which instances to label is closely related to classical Design of Experiments (DOE, see, e.g., Wu and Hamada 2009). Traditional DOE was developed for physical experiments in agricultural applications, where the goal was to explore the relationship between several input covariates and one output response under limited experimental resources. An important question widely studied within DOE is to define optimality criteria for experimental design and algorithms to obtain such designs, that is, Optimal Design of Experiments (ODOE). The goal of ODOE is to find experimental designs that are optimal with respect to some statistical criteria. In the classical theory of ODOE, a linear statistical model is usually assumed and the criteria are typically related either to the model parameter estimates or to the model predictions (see Kiefer and Wolfowitz 1960; Fedorov 1972; Pukelsheim 2006; Fedorov and Leonov 2013b). In this literature, the number of covariates or “factors” of interest in an experiment is relatively small and the experimental region is usually assumed to be Euclidean. However, in some modern learning tasks, such as image recognition and text categorization, the dimension of text or image data is often much higher than the dimension of covariates in a traditional agricultural or industrial experiment.

In order to perform a statistical learning task under these conditions, a manifold hypothesis is made, which assumes that although the training data of interest are available in a high-dimensional ambient space, there exists a lower-dimensional manifold where the data are located. This is in contrast to traditional linear dimensionality reduction based on principal components, where a linear subspace is assumed with the hope the data is concentrated in it. Starting with the work by Roweis and Saul (2000) and Tenenbaum, de Silva, and Langford (2000), a wide body of literature has shown how high-dimensional data, such as text or image data, frequently lie on a lower dimensional manifold and is usually sparse in its high-dimensional ambient space (see, e.g., Cheng and Wu 2013; Lin et al. 2017; Zhu et al. 2018; Yao and Zhang 2020). In many experimental design situations, it is infeasible to obtain and label enough training instances to fill up the high-dimensional ambient space. However, one could select points from a lower dimensional manifold space if the data points were much more dense on this low-dimensional space. Unfortunately, traditional DOE methods fail to take into account these complex characteristics of modern high-dimensional data (Li, Del Castillo, and Runger 2020).

Figure 1(a) displays an easy to visualize example, where the covariate data points are available in a three-dimensional Euclidean space but lie on the two-dimensional surface of a Torus embedded in the \( \mathbb{R}^3 \) ambient space. Two different experimental designs on this dataset are provided in Figure 1(b) and (c). The motivating question is therefore: how to find the
optimal design that improves the learning performance the most, while incorporating the manifold structure where data lie into account?

The goal of this article is to develop theory and an algorithm for constructing optimal experimental designs on high-dimensional manifold data, which minimize the number of experimental runs and at the same time acquire as much useful information about the response as possible. We assume training data are located on a lower dimensional Riemannian manifold, loosely defined as a curved space which when seen over a sufficiently small neighborhood resembles Euclidean (flat) space.

Although some previous authors have implemented ODOE criteria as Active Learning strategies for high-dimensional data (He 2010; Chen et al. 2010; Alaeddini et al. 2019), as far as we know, no existing work has provided theoretical guarantees or justification for such experimental designs on Riemannian manifolds.

Our contributions are summarized as follows: (a) we prove a new Equivalence Theorem for continuous optimal designs on manifold data, which shows how a D-optimal and a G-optimal design are equivalent on Riemannian manifolds; (b) we provide a new lower bound for the maximum prediction variance over the manifold and show how the lower bound can be achieved by a D/G optimal design; (c) we propose a new algorithm, ODOEM (Optimal Design of Experiments on Manifolds), for finding a continuous D/G optimal design on a Riemannian manifold, and prove that it is guaranteed to converge to the global D/G optimal design, and finally, (d) we illustrate the superior performance of our ODOEM algorithm on both of synthetic manifold datasets and a real-world image dataset.

The rest of this article is organized as follows. In Section 2, we briefly review the traditional ODOE problem on Euclidean space, and then introduce the manifold regularization model of Belkin, Niyogi, and Sindhwani (2006) on which our results are based, explaining the ODOE problem on manifolds. Section 3 provides the theoretical justification behind our ODOEM algorithm, where a new equivalence theorem is given for Riemannian manifolds. Section 4 gives the proposed ODOEM algorithm and provides a convergence analysis. Finally, Section 5 presents several numerical experiments conducted to demonstrate the effectiveness of the proposed algorithm for finding optimal designs on manifold data. We conclude the article with a summary and some possible further research directions in Section 6.

2. Optimal Design of Experiments on Manifolds

2.1. Traditional ODOE on Euclidean Space

Consider the following model

\[ y = f(x, \beta) + \epsilon = \beta^\top g(x) + \epsilon, \]

where \( g: \mathbb{R}^d \to \mathbb{R}^p \) is some nonlinear function that maps from the input space \( x \in \mathbb{R}^d \) to the feature space \( \mathbb{R}^p \), \( \beta \in \mathbb{R}^p \) is a column vector of unknown parameters, and \( \epsilon \) is assumed to have a \( N(0, \sigma^2) \) distribution. Given a sample of \( n \) design points \( \{x_i\}_{i=1}^n \), if the corresponding response values \( \{y_i\}_{i=1}^n \) are available, the well-known ordinary least squares estimates of the \( \beta \) parameters are given by

\[ \hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n (y_i - \beta^\top g(x_i))^2 \right\} = (X^\top X)^{-1}X^\top Y \]

where \( X \) is a \( n \times p \) design matrix with ith row defined as \( g(x_i)^\top \), and \( Y \) is a \( n \times 1 \) response vector. As a result, the corresponding fitted function is \( \hat{f}(x) = \hat{\beta}^\top g(x) \).

Classical work on ODOE was developed by Kiefer and Wolfowitz (1960) and summarized by Fedorov (1972) (see also Pukelsheim 2006; Fedorov and Leonov 2013b). Examples of optimality criteria for the regression model (1) are the D-optimality criterion which minimizes the determinant of the covariance matrix of the parameter estimates \( \text{var}(\hat{\beta}) = \sigma^2(X^\top X)^{-1} \), and the G-optimality criterion which minimizes the maximum prediction variance \( \max_{x \in X} \{\text{var}(\hat{y}_i)\} \) where \( X \) is the experimental region. These and similar criteria have been called “alphabetic optimality” design criteria by Box and Draper (2007).

While there have been recent attempts at applying alphabetic optimality criteria to manifold learning models (He 2010; Chen et al. 2010; Alaeddini et al. 2019), no theoretical justification exists, as far as we know, to these methods, and no guarantees can be given for their success other than empirical experimentation. A new theory for optimal experimental design is therefore needed that explicitly considers high-dimensional manifold data, justify existing methods if possible, and that provides a principled way to develop new algorithms. Before we discuss the design of experiments on manifolds, first we need to introduce a manifold learning model by Belkin, Niyogi, and Sindhwani (2006) that will be used in the sequel.
2.2. Manifold Regularization Model

In the standard paradigm of machine learning, learning is understood as a process that uses the training data \( \{x_i\}_{i=1}^n \) to construct a function \( f : \mathcal{X} \to \mathbb{R} \) that maps a data instance \( x \) to a label variable \( y \). Let \( P \) be the joint distribution that generates labeled data \( \{ (x_i, y_i) \}_{i=1}^n \subset \mathcal{X} \times \mathbb{R} \) and \( P_X \) be the marginal distribution that generates unlabeled data \( \{ x_i \}_{i=1}^n \subset \mathcal{X} \subset \mathbb{R}^d \). In order to extend the learning of functions to general Riemannian manifolds, Belkin, Niyogi, and Sindhwani (2006) assume that the conditional distribution \( P(y|x) \) varies smoothly as \( x \) moves along a manifold that supports \( P_X \). In other words, if two data points \( x_1, x_2 \in \mathcal{X} \) are close as measured by an intrinsic (or geodesic) distance on this manifold, then the two probabilities of the labels, \( P(y|x_1) \) and \( P(y|x_2) \), will be similar. These authors developed a semi-supervised learning framework that involves solving the following double regularized objective function:

\[
\hat{f} = \arg\min_{f \in \mathcal{H}_K} \left\{ \sum_{i=1}^l V(x_i, y_i, f) + \lambda_A \|f\|^2_{\mathcal{H}_K} + \lambda_1 \|f\|^2 \right\} \tag{3}
\]

where \( V \) is a given loss function (such as squared loss \( (y_i - f(x_i))^2 \)), \( \mathcal{H}_K \) is a Reproducing Kernel Hilbert Space (RKHS, Aronszajn 1950) with associated Mercer kernel \( K \), \( \|f\|^2_{\mathcal{H}_K} \) is a penalty term with the norm of \( \mathcal{H}_K \) that imposes smoothness conditions in the ambient space (Wahba 1990), and \( \|f\|^2 \) is a penalty term for non-smoothness along geodesics on the intrinsic manifold structure of \( P_X \). Moreover, \( \lambda_A \) and \( \lambda_1 \) are two regularization parameters that control the amount of penalization in the ambient space and in the intrinsic manifold that supports \( P_X \), respectively. Recent work on non-Euclidean data methods are related to (3), for instance, the spatial regression model proposed by Ettinger, Sangalli, and Perotto (2016) can be seen as the manifold regularization model (3) without the ambient space regularization. While there are also numerous nonparametric regression models on manifolds (see, e.g., Cheng and Wu 2013; Marzio, Panzera, and Taylor 2014; Lin et al. 2017), we focus on this article on the manifold regularization model from Belkin, Niyogi, and Sindhwani (2006) since it provides a nice representer theorem, an advantage that will be clear in what follows.

Intuitively, the choice of \( \|f\|^2 \) should be a roughness penalty of \( f \) corresponding to the probability distribution \( P_X \). When the support of \( P_X \) is a compact manifold \( \mathcal{M} \), one choice for \( \|f\|^2 \) is

\[
\int_{x \in \mathcal{M}} \|\nabla_M f\|^2 dP_X(x) = \int_{x \in \mathcal{M}} f \Delta_M f dP_X(x),
\]

where \( \nabla_M \) is the gradient of \( f \) along the manifold \( \mathcal{M} \) and \( \Delta_M \) is the Laplace-Beltrami operator. Clearly, minimizing this term will penalize \( f \) along \( \mathcal{M} \). However, in most real-world applications \( P_X \) is not known, and therefore empirical estimates of the marginal distribution must be used. Considerable research has been devoted to the case when \( P_X \) is supported on a compact manifold \( \mathcal{M} \subset \mathbb{R}^p \) and a finite set of points is used to infer the manifold structure (Roweis and Saul 2000; Tenenbaum, de Silva, and Langford 2000; Donoho and Grimes 2003; Coifman et al. 2005). Under these assumptions, it can be shown (see Belkin 2003; Lafon 2004) that problem (3) can be reduced to

\[
\hat{f} = \arg\min_{f \in \mathcal{H}_K} \left\{ \sum_{i=1}^l V(x_i, y_i, f) + \lambda_A \|f\|^2_{\mathcal{H}_K} + \lambda_1 \|f\|^2 + \lambda_2 \|L f\|^2 \right\} \tag{4}
\]

where \( f = [f(x_1), \ldots, f(x_n)]^\top \) and \( L \) is the Laplacian matrix associated with the data adjacency graph \( G \) that is constructed on all the labeled and the unlabeled data points \( \{x_i\}_{i=1}^n \). In particular, the graph Laplacian \( L \) approximates the Laplace-Beltrami operator acting on the continuous Riemannian manifold \( \mathcal{M} \) (see Belkin and Niyogi 2005; Coifman et al. 2005; Hein, Audibert, and von Luxburg 2005). The convergence of the graph Laplacian provides a theoretical justification to the common practice in manifold learning of using a graph and the corresponding geodesic distances as an approximate representation of the manifold \( \mathcal{M} \), providing a precise sense in which the graph approaches \( \mathcal{M} \) as the number of data points gets denser. This way, the term \( f^\top L f \) serves as an approximation for \( \|f\|^2 \), and enforces the penalization on the lack of smoothness of \( f \) as it varies between adjacent points in the graph \( G \).

The solution of the infinite dimensional problem (4) can be represented in terms of a finite sum over the labeled and unlabeled points:

\[
f(x) = \sum_{i=1}^n \alpha_i K(x_i, x) \tag{5}
\]

where \( K(\cdot, \cdot) \) is the Mercer kernel associated with the ambient space \( \mathcal{H}_K \). This constitutes a representer theorem for problem (4), similar to that in the theory of splines (Kimeldorf and Wahba 1970; Wahba 1990).

2.3. Regularized ODOE on Manifolds

Vuchkov (1977) provided the first discussion of a regularized method in the literature on ODOE (for Euclidean spaces), based on the ridge regression estimator:

\[
\hat{\beta}_{\text{ridge}} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^l (y_i - \beta^\top g(x_i))^2 + \lambda_{\text{ridge}} \|\beta\|^2 \right\}. \tag{6}
\]

Vuchkov’s motivation was to use the ridge estimator to solve the singular or ill-conditioned problems that exist in the sequential application of a D-optimal design algorithm when the number of design points is smaller than the number of parameters to estimate. The ridge solution (6) can be seen as a particular case of the more general learning problem (4) where \( V \) is a squared-loss function, the RKHS \( \mathcal{H}_K \) is equipped with a \( L^2 \)-norm and the manifold regularization parameter \( \lambda_1 \) is zero.

To discuss the optimal experimental design for the general manifold regularization model (4), we first introduce some notation. Without loss of generality, assume a sequential experimental design problem, starting with no labeled data at the beginning of the sequence. Let \( \{z_i\}_{i=1}^k \subset \{x_i\}_{i=1}^n \) be the set of points that has been labeled at the \( k \)th iteration, and \( y = (y_1, \ldots, y_k)^\top \) be the corresponding vector of responses or labels. Given a square loss function, the manifold regularization model
(4) becomes the Laplacian Regularized Least Squares (LapRLS) problem (Belkin, Niyogi, and Sindhwani 2006):

$$\hat{f} = \arg\min_{f \in \mathcal{H}_K} \left\{ \sum_{i=1}^{N} (y_i - f(z_i))^2 + \lambda_A \|f\|_{\mathcal{H}_K}^2 + \lambda_1 f^T L f \right\}. \quad (7)$$

Substituting the representer theorem solution (5) into (7), we get a convex differentiable objective function with respect to $\alpha$:

$$\hat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^n} \left\{ (y - K_{XX}^T \alpha)^T (y - K_{XX}^T \alpha) + \lambda_A \alpha^T K \alpha + \lambda_1 \alpha^T KLK \alpha \right\}, \quad (8)$$

where $K_{XX}$ and $K$ are the Gram matrices defined by

$$K_{XX} = \begin{bmatrix} K(x_1, z_1) & \ldots & K(x_1, z_k) \\ \vdots & \ddots & \vdots \\ K(x_n, z_1) & \ldots & K(x_n, z_k) \end{bmatrix}_{n \times k},$$

$$K = \begin{bmatrix} K(x_1, x_1) & \ldots & K(x_1, x_n) \\ \vdots & \ddots & \vdots \\ K(x_n, x_1) & \ldots & K(x_n, x_n) \end{bmatrix}_{n \times n}$$

and $K$ is the kernel embedded in the RKHS $\mathcal{H}_K$. Taking the derivative of (8) with respect to $\alpha$ and making it equal to 0, yields the following expression:

$$\hat{\alpha} = (K_{XX} K_{XX}^T + \lambda_A K + \lambda_1 KLK)^{-1} K_{XX} y. \quad (9)$$

Without loss of generality, consider a linear model and a linear kernel for $\mathcal{H}_K$, the model parameters $\beta$ can be estimated by

$$\hat{\beta} = X^T \hat{\alpha} = X^T (Z_k^T Z_k)^{-1} + \lambda_A XX^T + \lambda_1 XX^T LXX^T)^{-1} X Z_k^T y \quad (10)$$

where

$$Z_k = \begin{bmatrix} g(z_1)^T \\ \vdots \\ g(z_k)^T \end{bmatrix}, \quad X = \begin{bmatrix} g(x_1)^T \\ \vdots \\ g(x_n)^T \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix}. \quad (11)$$

Here a linear model is linear in the parameters $\beta$, but the function $g(x)$ is in general nonlinear in $x$. The assumption of a model linear in the parameters is necessary for the theory of optimal design of experiments. However, for classical optimal designs on Euclidean space, a nonlinear function $f(x, \beta)$ can be linearized using a Taylor expansion so that one finds a locally optimal design assuming a nominal parameter vector $\beta_0$ is known, or integrating the vector of prior parameters similar to Bayesian Designs (Fedorov and Leonov 2013; Pronzato and Pazman 2013).

By some simple linear algebra (a formal proof is provided in the supplementary materials), the estimated parameters $\hat{\beta}$ can be simplified to

$$\hat{\beta} = (Z_k^T Z_k + \lambda_A I_p + \lambda_1 X^T L X)^{-1} Z_k^T y. \quad (12)$$

Similarly to the theory of ODOE on Euclidean space, the regularized estimator resembles a Bayesian linear estimator, with the difference being that the regularization comes from the manifold penalization of high-dimensional data instead of some a priori covariance estimate (Pukelsheim 2006).

He (2010) demonstrated that the covariance matrix of (12) can be approximated as

$$\text{cov}(\hat{\beta}) \approx \sigma^2 (Z_k^T Z_k + \lambda_A I_p + \lambda_1 X^T L X)^{-1}, \quad (13)$$

since the regularization parameters $\lambda_A$ and $\lambda_1$ are usually set to be numerically small. The determinant of covariance matrix (13) is the statistical criterion we need to minimize to obtain a D-optimal design for manifold data. Before we discuss the optimal design algorithm, we will first provide its main theoretical justification, which just as the ODOE theory for Euclidean spaces, models a design as a probability measure over the manifold.

### 3. Equivalence Theorem on Manifolds

In Euclidean space, the theory of ODOE indicates an equivalence between the D-optimality criteria and the G-optimality criteria, which minimizes the maximum prediction variance, as stated by the celebrated Kiefer-Wolfowitz (KW) theorem (Kiefer and Wolfowitz 1960; Kiefer 1974). In analogy with the KW theorem, in this section we aim to develop a new equivalence result for optimal experimental design based on the manifold regularization model (4), which can then be used to justify algorithms for designing an optimal experiment on a Riemannian manifold, in an analogous way the Federov-Wynn ODOE algorithms for Euclidean space were motivated based on the continuous KW theory (Wynn 1970; Fedorov 1972).

Assume there is an infinite number of points $x$ that are uniformly distributed on a Riemannian manifold $\mathcal{M}$. Let $\epsilon$ be a continuous design on $\mathcal{M}$, a probability measure where $\epsilon(z)$, $z \in \mathcal{M}$ models the proportion of experimental effort allocated at point $z$. For any continuous design $\epsilon$, based on the Carathéodory Theorem, it is known (see Fedorov 1972) that it can be represented over a finite set of $n_0$ points of support, that is,

$$\epsilon = \left\{ \frac{z_1, z_2, \ldots, z_{n_0}}{q_1, q_2, \ldots, q_{n_0}} \right\}, \quad \text{where} \sum_{i=1}^{n_0} q_i = 1. \quad (14)$$

For any $\epsilon$, the corresponding information matrix of LapRLS model is defined as

$$M_{\text{Lap}}(\epsilon) = \int_{z \in \mathcal{X}} \xi(z) g(z) g(z)^T dz + \lambda_A I_p + \lambda_1 \int_{x \in \mathcal{M}} g(x) \Delta_{\mathcal{M}} g(x)^T d\mu, \quad (15)$$

where $\xi$ is a probability measure of design $\epsilon$ on the experimental region $\mathcal{X} \subseteq \mathcal{M} \subset \mathbb{R}^p$ and $\mu$ is the uniform measure on $\mathcal{M}$. For simplicity of notation, define

$$C = \lambda_A I_p + \lambda_1 \int_{x \in \mathcal{M}} g(x) \Delta_{\mathcal{M}} g(x)^T d\mu. \quad (16)$$

Then (15) can be written as

$$M_{\text{Lap}}(\epsilon) = \int_{z \in \mathcal{X}} \xi(z) g(z) g(z)^T dz + C. \quad (17)$$
Based on the parameters estimates (12), for a given continuous design \( \epsilon \), the prediction variance at a test point \( z \) is
\[
d(z, \epsilon) = \text{var}\left( \hat{\beta}^\top g(z) \right) = g(z)^\top \text{cov}(\hat{\beta}) g(z) = \sigma^2 g(z)^\top M^{-1}_\text{Lap}(\epsilon) g(z). \tag{18}
\]

As it can be seen, under the LapRLS model one can obtain a D-optimal design by maximizing the determinant of \( M_{\text{Lap}}(\epsilon) \) and a G-optimal design by minimizing \( \max_{z \in \mathcal{X}} d(z, \epsilon) \). Similarly to the optimal design of experiments in Euclidean space, we prove next an equivalence theorem on Riemannian manifolds that shows how the D and G optimality criteria lead to the same optimal design. Before the equivalence theorem is discussed, we need to prove some auxiliary results. The proofs of these propositions are provided in the supplementary materials.

**Proposition 1.** Let \( \epsilon_1 \) and \( \epsilon_2 \) be two designs with the corresponding information matrices \( M_{\text{Lap}}(\epsilon_1) \) and \( M_{\text{Lap}}(\epsilon_2) \). Then
\[
M_{\text{Lap}}(\epsilon_3) = (1 - \alpha)M_{\text{Lap}}(\epsilon_1) + \alpha M_{\text{Lap}}(\epsilon_2), \tag{19}
\]
where \( M_{\text{Lap}}(\epsilon_3) \) is the information matrix of the design \( \epsilon_3 = (1 - \alpha)\epsilon_1 + \alpha \epsilon_2 \), for \( 0 < \alpha < 1 \).

**Proposition 2.** Let \( \epsilon_1 \) and \( \epsilon_2 \) be two designs with the corresponding information matrices \( M_{\text{Lap}}(\epsilon_1) \) and \( M_{\text{Lap}}(\epsilon_2) \). Then
\[
\frac{d \log |M_{\text{Lap}}(\epsilon_3)|}{d \alpha} = \text{Tr}\left[M^{-1}_{\text{Lap}}(\epsilon_3)\left[M_{\text{Lap}}(\epsilon_2) - M_{\text{Lap}}(\epsilon_1)\right]\right], \tag{21}
\]
where \( M_{\text{Lap}}(\epsilon_3) \) is the information matrix of the design \( \epsilon_3 = (1 - \alpha)\epsilon_1 + \alpha \epsilon_2 \), for \( 0 < \alpha < 1 \).

**Proposition 3.** For any continuous design \( \epsilon \),
\[
1. \int_{z \in \mathcal{X}} d(z, \epsilon) \xi(z) dz = p - \text{Tr}\left[M^{-1}_{\text{Lap}}(\epsilon) C\right], \tag{23}
\]
\[
2. \max_{z \in \mathcal{X}} d(z, \epsilon) \geq p - \text{Tr}\left[M^{-1}_{\text{Lap}}(\epsilon) C\right]. \tag{24}
\]

**Proposition 4.** The function \( \log |M_{\text{Lap}}(\epsilon)| \) is a strictly concave function.

Based on Propositions 1–4, we can now prove the equivalence theorem for the LapRLS model. In summary, the following theorem demonstrates that the D-optimal design and G-optimal design are equivalent on the Riemannian manifold \( \mathcal{M} \). It also provides the theoretical value of maximum prediction variance of the LapRLS model when the D/G optimal design is achieved.

**Theorem 1 (Equivalence Theorem on Manifolds).** The following statements are equivalent:
1. the design \( \epsilon^* \) maximizes \( \det(M_{\text{Lap}}(\epsilon)) \);
2. the design \( \epsilon^* \) minimizes \( \max_{z \in \mathcal{X}} d(z, \epsilon) \);
3. \( \max_{z \in \mathcal{X}} d(z, \epsilon^*) = p - \text{Tr}\left[M^{-1}_{\text{Lap}}(\epsilon^*) C\right] \).

**Proof.** The proof of this theorem is provided in the supplementary materials.

**Theorem 2 (Convergence Theorem).** The iterative procedure in Algorithm 1 converges to the D-optimal design \( \epsilon^* \),
\[
\lim_{k \to \infty} |M_{\text{Lap}}(\epsilon_k)| = |M_{\text{Lap}}(\epsilon^*)|. \tag{32}
\]

**Proof.** Let the design \( \epsilon_0 \) not be D-optimal. Based on Proposition 6, we have
\[
|M_{\text{Lap}}(\epsilon_0)| < |M_{\text{Lap}}(\epsilon_1)| < \cdots < |M_{\text{Lap}}(\epsilon_k)| < \cdots \leq |M_{\text{Lap}}(\epsilon^*)|. \tag{33}
\]
Thus, the sequence \((\epsilon_k)\) converges. The proof proceeds by contradiction. Assume 
\[ |M_Lap(\epsilon_k)| < |M_Lap(\epsilon^*)| \] for all \(k\). Next we need to show 
\[ |M_Lap(\hat{\epsilon})| = |M_Lap(\epsilon^*)| . \] 
The proof proceeds by contradiction. Assume 
\[ |M_Lap(\hat{\epsilon})| < |M_Lap(\epsilon^*)| \]. 
By the convergence of the sequence \(|M_Lap(\epsilon_0)|, |M_Lap(\epsilon_1)|, \ldots, |M_Lap(\epsilon_k)|\), we know that, for any \(\eta > 0\), there \(\exists k_0 \in \mathbb{N}\) s.t. 
\[ |M_Lap(\epsilon_{k+1})| - |M_Lap(\epsilon_k)| < \eta \] for all \(k > k_0\). 
Based on Proposition 5, we have 
\[ (1 - \alpha_k)p(1 + \frac{\alpha_k}{1 - \alpha_k}d(z_{k+1}, \hat{\epsilon}_k) + \frac{\alpha_k}{1 - \alpha_k} \text{Tr}(M_Lap^{-1}(\epsilon_k)C)) \]
\[ |M_Lap(\hat{\epsilon})| - |M_Lap(\epsilon_k)| < \eta . \]
Then, 
\[ (1 - \alpha_k)p(1 + \frac{\alpha_k}{1 - \alpha_k}[d(z_{k+1}, \hat{\epsilon}_k) + \text{Tr}(M_Lap^{-1}(\epsilon_k)C)]) \]
\[ < 1 + \eta|M_Lap(\epsilon_k)|^{-1} . \] 
Defining \(\tau_k = d(z_{k+1}, \hat{\epsilon}_k) - [p - \text{Tr}(M_Lap^{-1}(\epsilon_k)C)]\), we can rewrite \((37)\) as 
\[ (1 - \alpha_k)p(1 + \frac{\alpha_k}{1 - \alpha_k}[\tau_k + p]) < 1 + \eta|M_Lap(\epsilon_k)|^{-1} . \] 
Next, define a function \(T(\tau_k, \alpha_k)\) as 
\[ T(\tau_k, \alpha_k) = (1 - \alpha_k)p(1 + \frac{\alpha_k}{1 - \alpha_k}[\tau_k + p]) \] 
such that 
\[ \frac{\partial T}{\partial \tau_k} = (1 - \alpha_k)p \alpha_k \frac{1}{1 - \alpha_k} . \] 
Clearly, \(\frac{\partial T}{\partial \tau_k} > 0\) for \(0 < \alpha_k < 1\). Thus, for a given \(0 < \alpha_k < 1\), \(T(\tau_k, \alpha_k)\) is a monotonic increasing function with respect to \(\tau_k\). 
On the other hand, 
\[ \frac{\partial T}{\partial \alpha_k} = -p(1 - \alpha_k)^{p-1}(1 + \frac{\alpha_k}{1 - \alpha_k}[\tau_k + p]) \]
\[ + (1 - \alpha_k)^p \frac{\tau_k}{(1 - \alpha_k)^2} . \]
Let \(\frac{\partial T}{\partial \alpha_k} \geq 0\), we have 
\[ (1 - \alpha_k)^{p-2}[\tau_k + p] \geq p(1 - \alpha_k)^{p-1}(1 + \frac{\alpha_k}{1 - \alpha_k}[\tau_k + p]) \]
\[ \alpha_k \leq \frac{\tau_k}{p(\tau_k - 1)} . \]
Thus, for \(0 < \alpha_k \leq \frac{\tau_k}{p(\tau_k - 1)}\) and \(\tau_k > 0\), \(T(\tau_k, \alpha_k)\) is a monotone increasing function. In particular, plugging in the expression for \(\tau_k\), we get 
\[ \frac{\tau_k}{p(\tau_k - 1)} = \frac{d(z_{k+1}, \hat{\epsilon}_k) - [p - \text{Tr}(M_Lap^{-1}(\epsilon_k)C)]}{p(d(z_{k+1}, \hat{\epsilon}_k) - [1 - \text{Tr}(M_Lap^{-1}(\epsilon_k)C)])} . \] 
Notice that \(0 < \alpha_k \leq \frac{\tau_k}{p(\tau_k - 1)}\) is the same choice of \(\alpha_k\) in the proposed Algorithm 1.

From the assumption \((35)\), Proposition 3 and Theorem 1, it follows that \(\tau_k \geq 0\). This guarantees the existence of \(\alpha_k\) such that \(0 < \alpha_k \leq \frac{\tau_k}{p(\tau_k - 1)}\). Thus, for any \(\tau_k > 0\) and \(0 < \alpha_k \leq \frac{\tau_k}{p(\tau_k - 1)}\), we have \(T(\tau_k, \alpha_k) > 1\). Note that \(\eta\) is an arbitrary positive number in Equation \((38)\), which implies \(\tau_k\) need to be an infinitely small positive number to satisfy Equation \((38)\), that is, given \(\forall \zeta > 0\), there \(\exists \tilde{k}(\xi) \in \mathbb{N}\) s.t. 
\[ \tau_k = d(z_{k+1}, \hat{\epsilon}_k) - [p - \text{Tr}(M_Lap^{-1}(\epsilon_k)C)] < \zeta \] for \(k > \tilde{k}(\xi)\). 
However, based on the assumption \((35)\) and Theorem 1, we have that 
\[ d(z_{k+1}, \hat{\epsilon}_k) - [p - \text{Tr}(M_Lap^{-1}(\epsilon_k)C)] \geq \delta_k > 0 \] for \(\forall k\). 
Choosing \(\zeta < \delta_k\), we have a contradiction, and therefore, the convergence theorem is proved. 

When the experimental design space is continuous and the intrinsic manifold structure is known, the information matrix \((15)\) can be analytically derived and Algorithm 1 can be implemented to find the optimal D/G design \(\epsilon^*\) on the manifold \(M\). As it was shown before, Algorithm 1 is a converging algorithm.
Algorithm 2 ODOEM with discrete candidate points

Inputs: \( \lambda_A, \lambda_I, \) kernel function \( K, \) discrete candidate points \( \{x_i\}_{i=1}^n \) within the experimental region, and initial labeled set \( \mathcal{L}_k = \{(z_i, y_i)\}_{i=1}^k \).
Define the unlabeled set as \( \mathcal{U}_k = \{x_i\}_{i=1}^n - \{z_i\}_{i=1}^k \).
Compute the initial information matrix from the labeled set:
\[
M_{\text{Lap}}(\mathcal{L}_k) = K_{XX}K_{XX}^\top + \lambda_A K + \lambda_I KLK
\]
while more data need to be labeled do
1. Find \( z_{k+1} \) s.t.
\[
z_{k+1} = \arg\max_{u \in \mathcal{U}_k} K_{Xu}M_{\text{Lap}}^{-1}(\mathcal{L}_k)K_{Xu}
\]
where \( K_{Xu} = [K(x_1, u), K(x_2, u), \ldots, K(x_n, u)]^\top \).
2. Update the labeled set and unlabeled set
\[
\mathcal{L}_{k+1} = \mathcal{L}_k \cup \{(z_{k+1}, y_{k+1})\}
\]
\[
\mathcal{U}_{k+1} = \mathcal{U}_k - z_{k+1}
\]
3. Compute the information matrix \( M_{\text{Lap}}(\mathcal{L}_{k+1}) \), set \( k = k + 1 \) and repeat Steps 1–3.
end while
Output: Optimal design on manifold \( \mathcal{M} \).

The resulting sequence of determinants in Algorithm 2, \( |M_{\text{Lap}}(\epsilon_k)| \), is still monotonic increasing, since
\[
|M_{\text{Lap}}(\epsilon_k) + g(z_{k+1})g(z_{k+1})^\top| = |M_{\text{Lap}}(\epsilon_k)[1 + g(z_{k+1})^\top M_{\text{Lap}}^{-1}(\epsilon_k)g(z_{k+1})]| > |M_{\text{Lap}}(\epsilon_k)|
\]
where \( z_{k+1} = \arg\max_{z \in \mathcal{X} \setminus Z_k} d(z, \epsilon_k) = \arg\max_{z \in \mathcal{X} \setminus Z_k} g(z)^\top M_{\text{Lap}}^{-1}(\epsilon_k)g(z) \).

5. Numerical Results
To illustrate the empirical performance of Algorithm 2 in practice, we consider its application to both synthetic datasets and real-world high-dimensional image datasets. The synthetic datasets are low-dimensional manifold examples that permit straightforward visualization of the resulting designs on manifolds and are shown first.

5.1. Synthetic Manifold Datasets
We generate four different two-dimensional manifold datasets: data on a Torus, on a Möbius Strip, on a figure “8” immersion of a Klein bottle and on a classic Klein bottle (Gray, Abbena, and Salamon 2006). Each of the first three datasets contains 400 instances and the last dataset contains 1600 instances. For all four datasets, we plot these two-dimensional manifolds in a three-dimensional Euclidean space, as shown in Figures 2 and 3. The colors on these manifolds represent the corresponding response values \( \{y_i\}_{i=1}^n \) or their estimates \( \{\hat{y}_i\}_{i=1}^n \) based on different experimental designs. The true response values \( \{y_i\}_{i=1}^n \) are defined by
\[
y = \sin(u) + \sin^2(u) + \cos^2(v)
\]
where \( u \in [0, 2\pi) \) and \( v \in [0, 2\pi] \). The red numbers on the manifolds represent the sequence of labeled instances by different design algorithms.

The regularization parameters \( \lambda_A \) and \( \lambda_I \) in model (7) are usually selected by cross-validation. However, ODOEM is a sequential design algorithm and the order in which instances (points on the manifold) are labeled is important. The cross-validation idea, which randomly divides the labeled instances into a training set and a validation set, is impractical in a sequential design framework. Thus, in these numerical experiments, we set \( \lambda_A = 0.01 \) for numerical stability and generate a decreasing sequence of \( \lambda_I = -\ln(k/n) \), where \( k \) is the number of labeled instance at the \( k \)th iteration and \( n \) is the total number of instances. The reason for such a decreasing sequence of \( \lambda_I \) comes from the penalized loss function (7) and the performance evaluation criterion MSE = \( \sum_{i=1}^n (y_i - \hat{f}(z_i))^2 \). For the manifold regularization model, the estimated learning function \( \hat{f} \) is obtained by minimizing the objective function (7). At early iterations, there are only few labeled instances, and \( \hat{f} \) would benefit more from penalizing the learning function along the manifold structure (second regularization term). As the number of labeled instances increases, larger \( \lambda_I \) might not lead to smaller MSE. For example, we consider the extreme scenario when all the instances have been labeled, that is, \( k = n \). If one desires to achieve a smaller MSE = \( \sum_{i=1}^n (y_i - \hat{f}(z_i))^2 \), it is better to...
Figure 2. Torus and Möbius Strip examples. Top row: when $\{x_i\}_{i=1}^n$ lie on a Torus. Bottom row: when $\{x_i\}_{i=1}^n$ lie on a Möbius Strip. (a) The colors represent the true response values defined on the Torus and the Möbius Strip. (b) 100 labeled instances (red numbers) and fitted response values (surface colors) by a kernel regression and a standard D-optimal design, (which does not consider the manifold structure of the data). (c) 100 labeled instances (red numbers) and fitted response values (surface colors) by our ODOEM algorithm. As it can be seen by comparing the coloring of the surfaces, for both examples the fitted functions in (c) better approximate the true functions in (a) than the fitted function in (b). Similar cases with noise are provided in the supplementary materials, where similar results are observed.

Figure 3. Klein bottle examples. Top row: when $\{x_i\}_{i=1}^n$ lie on a figure “8” immersion of Klein bottle. Bottom row: when $\{x_i\}_{i=1}^n$ lie on a bottle shape Klein bottle. (a) The colors represent the true response values defined on the Klein bottles. (b) 100 labeled instances (red numbers) and fitted response values (surface colors) by a kernel regression with a standard D-optimal Design for Euclidean spaces (which does not consider the manifold structure of the data). (c) 100 labeled instances (red numbers) and fitted response values (surface colors) by our ODOEM algorithm. Once again, for both examples the fitted functions in (c) better approximate the true functions in (a) than the fitted function in (b) as can be seen comparing the coloring of the surfaces. Similar cases with noise are provided in the supplementary materials, where similar results are observed.
estimate \( \hat{f} \) by

\[
\hat{f} = \arg\min_{f \in \mathcal{H}_K} \sum_{i=1}^{N} (y_i - f(z_i))^2,
\]

instead of using (7). Therefore, setting \( \lambda_I = -\ln(k/n) \), as \( k \) increases \( \lambda_I \) decreases such that \( \lambda_I = 0 \) when all the instances have been labeled, and this choice provides good empirical performance as seen in the numerical experiments shown below.

We compare the ODOEM algorithm with a classical D-optimal design algorithm on a kernel regression model, which does not consider the manifold structure. In both of these algorithms, we use a nonlinear Radial Basis Function (RBF) kernel with range parameter set to 0.01.

For some applications, the data may not strictly lie on a given manifold due to noise. In order to explore the robustness of the ODOEM algorithm to noise, we also let the four synthetic datasets fluctuate around their manifolds by adding noise to \( \{x_i\}_{i=1}^{n} \). In other words, for each of the four manifolds, we investigate both the case when the data \( \{x_i\}_{i=1}^{n} \) lie exactly on the given manifold and the case when \( \{x_i\}_{i=1}^{n} \) are not exactly on the manifold. The figures of the noisy cases are provided in the supplementary materials.

Based on these results, the following comments can be made:
(a) in the Torus, Möbius Strip, and Figure “8” Immersion examples, the instances selected by the classical D-optimal design tend to be clustered in certain regions, while the instances selected by ODOEM are widely spread over these manifolds. Although the fitted values from using the classical D-optimal design is close to the true values in some very small regions on manifolds, it is clear that ODOEM provides better overall fitting performance. (b) In the Klein bottle example, the classical D-optimal design selects relatively dispersive instances, but the
function is still poorly fitted on the bottle. It is illustrated that the kernel regression is not able to capture the manifold structure and incorporate it into the learning process. (c) ODOEM is adaptive to various manifold structures. It picks instances all over the manifolds and provides stable and superior fitting performance. In summary, on all four synthetic manifold datasets, ODOEM performs much better than kernel regression D-optimal Design in terms of instance selection and function fitting, under both of the noise-free cases and the noisy cases.

5.2. Columbia Object Image Library

To demonstrate an application to a more realistic manifold learning problem, we tested the ODOEM algorithm on the Columbia Object Image Library (COIL-20). COIL-20 is a database of grey-scale images of 20 different objects and these images were taken at pose intervals of 5 degrees for each object. There are two versions of this database. In this article, we choose the processed database that contains 1440 normalized images each made of $32 \times 32$ pixels.

In this set of experiments, the input data $\{x_i\}_{i=1}^n$ are the object images and the response values $\{y_i\}_{i=1}^n$ are the corresponding pose angles of these images with respect to the observer. Given an object image, our goal is to estimate the angle of this object in the image. Among 20 different objects, we choose images of four different objects as illustration: a “Rubber Duck”, a “Cannon,” a “Toy Car,” and a “Piggy Bank.” For each object, we apply the ODOEM algorithm (Algorithm 2) to decide which instances to label and then train the LapRLS model (7) to predict the angles of the images using the labeled and unlabeled instances. Comparisons were made with the following alternative algorithms: Kernel regression model with a classical D-optimal Design; Kernel regression model with a random sampling scheme; Kernel regression model with a $L_2$-discrepancy uniform design (Fang, Li, and Sudjianto 2006); Kernel regression model with a minimax uniform design (Fang, Li, and Sudjianto 2006); Kernel regression model with a maximin uniform design (Fang, Li, and Sudjianto 2006); SVM model with MAED (Manifold Adaptive Experimental Design, Cai and He 2012); SVM model with TED (Transductive Experimental Design, Yu et al. 2008).

Similarly to the synthetic manifold experiments, we used a RBF kernel with range parameter fixed at 0.01 for both kernel regression and SVM. In addition, we set $\lambda_A = 0.01$ and $\lambda_I = -\ln(k/n)$ in ODOEM and kernel regressions. The results are shown in Figures 4 and 5. A sensitivity analysis of these hyperparameters is provided in the supplementary materials and further demonstrates the robustness of the ODOEM algorithm.

Figure 4, in particular, illustrates the first four images selected by classical D-optimal design and ODOEM for training the models, and Figure 5 demonstrates the fitting performance of the different algorithms in terms of mean square error (MSE). The results for the Toy Car and Piggy Bank show similar behavior and are provided in the supplementary materials.

Based on the results obtained, the following comments can be made: (a) compared to the classical D-optimal design, there is a greater dispersion (in terms of angles) within the first four images selected by ODOEM, which improves the learning curve in Figure 5; (b) for some uniform design criteria, the corresponding optimization is not convex. Since the images are labeled sequentially, there is no guarantee that the global optimum can be achieved. This explains why some uniform designs do not work very well in these experiments. (c) since the MAED algorithm (Cai and He 2012) also benefits from incorporating the manifold structure into the design process, it leads to better fitting performance than most algorithms compared, except ODOEM. (d) ODOEM outperforms all the other algorithms in this experiment.

6. Conclusions

In this article, we have developed a theoretical framework for the optimal design of experiments on Riemannian manifolds. Similarly to the Euclidean case, we have shown that D-optimal designs and G-optimal designs are equivalent when the regressors lie on a manifold. Moreover, we have provided a new lower
bound for the maximum prediction variance, which includes the intrinsic manifold geometrical information, demonstrating that this lower bound is achieved at the D/G optimal design. In our proofs, the hyperparameters are treated as constants, a common assumption in most similar proofs in machine learning theory. We also proposed a converging algorithm for finding an optimal experimental design on a manifold. Finally, we compared our proposed algorithm with other popular designs and models proposed for both manifold and Euclidean optimal design of experiments on several synthetic datasets and real-world image problems, and demonstrated the overall best performance of our ODOEM algorithm.

Further research could aim to develop a systematic procedure for choosing the regularization parameters $\lambda_A$ and $\lambda_I$. While we provided empirical justification for our choices of $\lambda_A$ and $\lambda_I$, a model selection criterion with theoretical guarantees is desirable, given that cross-validation is not an option in an experimental design situation. Related work has been discussed by Li, Del Castillo, and Runger (2020), who maximize the likelihood function to choose the values of $\lambda_A$ and $\lambda_I$ in a Gaussian Process model. Another area for further work is to consider other optimality criteria. As it is well-known, there are optimality criteria other than D/G optimality in the literature on ODOE for Euclidean spaces which could be explored likewise in the manifold case. Also, for very large scale problems with many millions of discrete candidate points, evaluating each point with the corresponding design criteria is computationally prohibitive, and some modifications to our algorithm can be investigated, for example, one could apply first clustering techniques to the covariate data and then evaluate a representative point from each cluster.

**Supplementary Materials**

1. Proofs.
2. Further performance evidence of ODOEM for synthetic manifold datasets.
3. Further performance on image datasets.
4. Sensitivity analysis of the regularization and range hyperparameters used by ODOEM.
5. Matlab codes and datasets (zipped file).

**Funding**

The authors gratefully acknowledge NSF grant CMMI 1537987.

**ORCID**

Enrique Del Castillo http://orcid.org/0000-0002-6148-5692

**References**

Alaedini, A., Craft, E., Meka, R., and Martinez, S. (2019), "Sequential Laplacian Regularized v-optimal Design of Experiments for Response Surface Modeling of Expensive Tests: An Application in Wind Tunnel Testing," *IIE Transactions*, 51, 559–576. [876]

Aronszajn, N. (1950), “Theory of Reproducing Kernels,” *Transactions of the American Mathematical Society*, 68, 337–404. [877]

Belkin, M. (2003), “Problems of Learning on Manifolds,” Ph.D. thesis, The University of Chicago. [877,881]

Belkin, M., and Niyogi, P. (2005), “Towards a Theoretical Foundation for Laplacian-based Manifold Methods,” in *Proceedings of Conference on Learning Theory*. [877]

Belkin, M., Niyogi, P., and Sindhwani, V. (2006), “Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples,” *Journal of Machine Learning Research*, 7, 2399–2434. [876,877,878]

Box, G., and Draper, N. (2007), *Response Surfaces, Mixtures, and Ridge Analyses*. Wiley Series in Probability and Statistics, Hoboken, NJ: Wiley. [876]

Cai, D., and He, X. (2012), “Manifold Adaptive Experimental Design for Text Categorization,” *IEEE Transactions on Knowledge and Data Engineering*, 24, 707–719. [884]

Chen, C., Chen, Z., Bu, J., Wang, C., Zhang, L., and Zhang, C. (2010), “G-optimal Design with Laplacian Regularization,” *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 1, 413–418. [876]

Cheng, M., and Wu, H. (2013), “Local Linear Regression on Manifolds and its Geometric Interpretation,” *Journal of the American Statistical Association*, 108, 1421–1434. [875,877]

Coifman, R., Lafon, S., Lee, A., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. (2005), “Geometric Diffusions as a Tool for Harmonic Analysis and Structure Definition of Data: Diffusion Maps,” *Proceedings of the National Academy of Sciences*, 102, 7426–7431. [877,881]

Donoho, D., and Grimes, C. (2003), “Hessian Eigenmaps: Locally Linear Embedding Techniques for High Dimensional Data,” *Proceedings of the National Academy of Sciences*, 100, 5591–5596. [877]

Etingber, S., Sangalli, L. M., and Perotto, S. (2016), “Spatial Regression Models over Two-Dimensional Manifolds,” *Biometrika*, 103, 71–88. [877]

Fang, K., Li, R., and Sudjianto, A. (2006), *Design and Modeling for Computer Experiments*. Computer Science and Data Analysis Series, Boca Raton, FL: Chapman & Hall/CRC. [884]

Fedorov, V. V. (1972), *Theory of Optimal Experiments*. New York: Academic Press. [875,876,878]

Fedorov, V. V., and Hackle, P. (1997), *Model-Oriented Design of Experiments*, New York: Springer. [881]

Fedorov, V. V., and Leonov, S. (2013a), *Optimal Design for Nonlinear Response Models*, Boca Raton, FL: CRC Press. [878]

Fedorov, V. V. (2013b), *Optimal Design for Nonlinear Response Models*, Boca Raton, FL: CRC Press. [875,876]

Gray, A., Abbena, E., and Salamon, S. (2006), *Modern Differential Geometry of Curves and Surfaces with Mathematica, Third Edition (Studies in Advanced Mathematics)*, Boca Raton, FL: Chapman & Hall/CRC. [881]

He, X. (2010), “Laplacian Regularized d-optimal Design for Active Learning and its Application to Image Retrieval,” *IEEE Transactions on Image Processing*, 19, 254–263. [876,878]

Hein, M., Audibert, J. Y., and von Luxburg, U. (2005), “From Graphs to Manifolds-Weak and Strong Pointwise Consistency of Graph Laplacians,” in *Proceedings of the 18th Conference on Learning Theory*. [877]

Kiefer, J. (1974), “General Equivalence Theory for Optimum Designs (Approximate Theory),” *The Annals of Statistics*, 2, 849–879. [878]

Kiefer, J., and Wolfowitz, J. (1960), “The Equivalence of Two Extremum Problems,” *Canadian Journal of Mathematics*, 12, 363–366. [875,876,878]

Kimeldorf, G. S., and Wahba, G. (1970), “A Correspondence between Bayesian Estimation on Stochastic Processes and Smoothing by Splines,” *Annals of Mathematical Statistics*, 41, 495–502. [877]

Lafon, S. (2004), “Diffusion Maps and Geometric Harmonics,” Ph. D. thesis, Yale University. [877,881]

Li, H., Del Castillo, E., and Runger, G. (2020), “On Active Learning Methods for Manifold Data,” *TEST*, 29, 1–33. [875,885]

Lin, L., Thomas, B. S., Zhu, H., and Dunson, D. B. (2017), “Extrinsic Local Regression on Manifold-Valued Data,” *Journal of the American Statistical Association*, 112, 1261–1273. [875,877]

Marzio, M. D., Panzera, A., and Taylor, C. C. (2014), “Nonparametric Regression for Spherical Data,” *Journal of the American Statistical Association*, 109, 748–763. [877]
Pronzato, L., and Pazman, A. (2013), Design of Experiments in Nonlinear Models: Asymptotic Normality, Optimality Criteria and Small-sample Properties, New York: Springer. [878]
Pukelsheim, F. (2006), Optimal Design of Experiments, Philadelphia, PA: Society for Industrial and Applied Mathematics. [875,876,878]
Roweis, S. T., and Saul, L. K. (2000), "Nonlinear Dimensionality Reduction by Locally Linear Embedding," Science, 290, 2323–2326. [875,877]
Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000), "A Global Geometric Framework for Nonlinear Dimensionality Reduction," Science, 290, 2319–2323. [875,877]
Vuchkov, I. (1977), "A Ridge-Type Procedure for Design of Experiments," Biometrika, 64, 147–150. [877]
Wahba, G. (1990), Spline Models for Observational Data, Philadelphia, PA: Society for Industrial and Applied Mathematics. [877]
Wu, C.-F. J., and Hamada, M. S. (2009), Experiments: Planning, Analysis, and Parameter Design Optimization, (2nd ed.), New York: Wiley. [875]
Wynn, H. P. (1970), "The Sequential Generation of d-optimum Experimental Designs," The Annals of Mathematical Statistics, 41, 1655–1664. [878]
Yao, Z., and Zhang, Z. (2020), "Principal Boundary on Riemannian Manifolds," Journal of the American Statistical Association, 115, 1435–1448. [875]
Yu, K., Zhu, S., Xu, W., and Gong, Y. (2008), "Non-greedy Active Learning for Text Categorization using Convex Transductive Experimental Design," in Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore, pp. 635–642. [884]
Zhu, B., Liu, J. Z., Cauley, S. F., Rosen, B. R., and Rosen, M. S. (2018), "Image Reconstruction by Domain-Transform Manifold Learning," Nature, 555, 487–492. [875]