Journal of Machine Learning Research 7 (2006) 519–549 Submitted 6/05; Revised 11/05; Published 3/06

Learning Coordinate Covariances via Gradients

Sayan Mukherjee SAYAN@STAT.DUKE.EDU
Institute of Statistics and Decision Sciences
Institute for Genome Sciences and Policy
Department of Computer Science
Duke University
Durham, NC 27708, USA

Ding-Xuan Zhou MAZHOU@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong, China

Editor: John Shawe-Taylor

Abstract

We introduce an algorithm that learns gradients from samples in the supervised learning framework. An error analysis is given for the convergence of the gradient estimated by the algorithm to the true gradient. The utility of the algorithm for the problem of variable selection as well as determining variable covariance is illustrated on simulated data as well as two gene expression data sets. For square loss we provide a very efficient implementation with respect to both memory and time.

Keywords: Tikhnonov regularization, variable selection, reproducing kernel Hilbert space, generalization bounds

1. Introduction

The advent of data sets with many variables or coordinates in the biological and physical sciences has driven the use of a variety of machine learning approaches based on Tikhonov regularization or global shrinkage such as support vector machines (SVMs) (Vapnik, 1998) and regularized least square classification (Poggio and Girosi, 1990). These algorithms have been very successful in both classification and regression problems. However, in a number of applications the classical questions from statistical linear modelling of which variables are most relevant to the prediction and how the coordinates vary with respect to each other have been revived. In the context of high dimensional data with few examples, the "large p, small n" paradigm (West, 2003), this leads to foundational questions in constructing and interpreting statistical models. Since statistical models based on shrinkage or regularization (Vapnik, 1998; West, 2003) have had success in the framework of both classification and regression, we formulate the problem of learning coordinate covariation and relevance in this framework.

We first describe the Tikhonov regularization method for classification and regression in order to define notation and basic concepts. We then introduce an algorithm that learns gradients of a function. We also motivate the algorithm and give an intuition of how the gradient can be used to learn coordinate covariation and relevance.

1.1 Classification and Regression

Classification and regression problems can be addressed in the framework of learning or estimating functions from a hypothesis space given sample values. An efficient learning method is the Tikhonov regularization scheme. Let X be a compact metric space and the hypothesis space, $\mathcal{H}$, be a set of functions $X \to Y \subset \mathbb{R}$. If we assign a penalty functional $\Omega : \mathcal{H} \to \mathbb{R}_+$ on $\mathcal{H}$ and choose a loss function $V : \mathbb{R}^2 \to \mathbb{R}_+$, the Tikhonov regularization scheme in $\mathcal{H}$ associated with (V, Ω) is defined for a sample $\mathbf{z} = \left\{(x_i, y_i)\right\}_{i=1}^m \in (X \times Y)^m$ and $\lambda > 0$ as

$$f_{\mathbf{z}}^V = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{m} \sum_{i=1}^m V(y_i, f(x_i)) + \lambda \Omega(f) \right\}. \tag{1}$$

The efficiency of learning algorithms of type (1) in machine learning can be seen when $\mathcal{H}$ takes the special choice of a reproducing kernel Hilbert space generated by a Mercer kernel.

Let $K : X \times X \to \mathbb{R}$ be continuous, symmetric and positive semidefinite, i.e., for any finite set of distinct points $\{x_1, \cdots, x_m\} \subset X$, the matrix $(K(x_i, x_j))_{i,j=1}^m$ is positive semidefinite. Such a function is called a *Mercer kernel*.

The *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}_K$ associated with the Mercer kernel K is defined (see Aronszajn (1950)) to be the completion of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ satisfying $\langle K_x, K_y \rangle_K = K(x, y)$. The reproducing property of $\mathcal{H}_K$ is

$$\langle K_x, f \rangle_K = f(x), \qquad \forall x \in X, f \in \mathcal{H}_K. \tag{2}$$

If $\mathcal{H} = \mathcal{H}_K$ and $\Omega(f) = \|f\|_K^2$ in (1), the reproducing property (2) tells us that

$$f_{\mathbf{z}}^V = \sum_{i=1}^m c_i K_{x_i}$$

and the coefficients $\{c_i\}_{i=1}^m$ can be found by solving an optimization problem in $\mathbb{R}^m$.

Assume that ρ is a probability distribution on $Z := X \times Y$ and $\mathbf{z} = \left\{(x_i, y_i)\right\}_{i=1}^m \in Z^m$ is a random sample independently drawn according to ρ.

When the loss function is the least-square loss $V(y, t) = (y - t)^2$, the algorithm (1) is least-square regression and the objective is to learn the regression function

$$f_\rho(x) = \int_Y y d\rho(y|x), \qquad x \in X \tag{3}$$

from the random sample $\mathbf{z}$. Here $\rho(\cdot|x)$ is the conditional distribution of ρ at x. Denote ρ_X as the marginal distribution of ρ on X and $L_{\rho_X}^2$ as the L^2 space with the metric $\|f\|_\rho := (\int_X |f(x)|^2 d\rho_X)^{1/2}$. There has been a vast literature (e.g. (Evgeniou et al., 2000; Zhang, 2003; Vito et al., 2005; Smale and Zhou, 2006b)) in learning theory showing for this least-square regression algorithm the convergence of $f_{\mathbf{z}}^V$ to f_ρ in the metric $\|\cdot\|_\rho$ under the assumption that f_ρ lies in the closure of $\mathcal{H}_K$ and $\lambda = \lambda(m) \to 0$ as $m \to \infty$.

For the (binary) classification purpose, we take $Y = \{1, -1\}$. A real valued function $f : X \to \mathbb{R}$ induces a classifier $\mathrm{sgn}(f) : X \to Y$. In this case, one uses a (convex) loss function $\phi : \mathbb{R} \to \mathbb{R}_+$ to measure the empirical error $\phi(t)$, $t = yf(x)$, when $\mathrm{sgn}(f(x))$ is applied to predict $y \in Y$. Examples of such a convex loss function ϕ include the logistic loss

$$\phi(t) = \log\left(1 + e^{-t}\right), \qquad t \in \mathbb{R} \tag{4}$$

and the hinge loss $\phi(t) = \max\{0, 1-t\}$. For $V(y, f(x)) = \phi(t)$ in (1) extensive investigation in learning theory (e.g. (Cortes and Vapnik, 1995; Evgeniou et al., 2000; Schoelkopf and Smola, 2001; Vapnik, 1998; Wu and Zhou, 2005)) has shown that $\operatorname{sgn}(f_{\mathbf{z}}^V)$ converges to the Bayes rule $\operatorname{sgn}(f_\rho)$ with respect to the misclassification error:

$$\mathcal{R}(\operatorname{sgn}(f)) = \operatorname{Prob}\{\operatorname{sgn}(f(x)) \neq y\}.$$

1.2 Learning the Gradient

In this paper we are interested in learning the gradient of f_ρ from the function sample values. Let $X \subset \mathbb{R}^n$. Denote $x = (x^1, x^2, \ldots, x^n)^T \in \mathbb{R}^n$. The gradient of f_ρ is the vector of functions (if the partial derivatives exist)

$$\nabla f_\rho = \left(\frac{\partial f_\rho}{\partial x^1}, \ldots, \frac{\partial f_\rho}{\partial x^n} \right)^T. \tag{5}$$

The relevance of learning the gradient with respect to the problems of variable selection and estimating coordinate covariation is that the gradient provides the following information:
(a) variable selection: the norm of a partial derivative $\| \frac{\partial f_\rho}{\partial x^i} \|$ indicates the relevance of this variable, since a small norm implies a small change in the function f_ρ with respect to the i-th coordinate,
(b) coordinate covariation: the inner product between partial derivatives $\left\langle \frac{\partial f_\rho}{\partial x^i}, \frac{\partial f_\rho}{\partial x^j} \right\rangle$ indicates the covariance of the i-th and j-th coordinates with respect to variation in f_ρ.

We now motivate the derivation of our gradient learning algorithm. Taylor expanding a function $g(u)$ around the point x gives us

$$g(u) \approx g(x) + \int_{\Delta x \in \Gamma_x} \langle \nabla g, \Delta x \rangle,$$

where the inner product and a neighborhood Γ_x of x are determined according to what is natural for different settings. For example, in the manifold setting we know the marginal ρ_X is concentrated on a manifold $\mathcal{M}$ and it is natural to formulate the following expansion

$$g(u) \approx g(x) + \int_{\Delta x \in \mathcal{M}_x} \langle \nabla_{\mathcal{M}} g, \Delta x \rangle,$$

where $\Delta x \in \mathcal{M}_x$ are points on the manifold around x with respect to the intrinsic distance on the manifold and the inner product is L^2 over the manifold (Belkin and Niyogi, 2004). In the graph setting we are given a sparse sample on the manifold which can be thought of as vertices of a graph and the distance between the points is the weight matrix of the graph. A natural formulation in this setting is to set Γ_x to be vertices connected to x and the inner product as the weight matrix. Minimizing the empirical error (with regularization) between $g(u)$ and its expansion $g(x) + \int_{\Delta x \in \Gamma_x} \langle \nabla g, \Delta x \rangle \approx g(x) + \nabla g(x) \cdot (u - x)$ for $u \approx x$ results in various learning algorithms.

For regression the algorithm to learn gradients will use least-square loss to minimize the error of the Taylor expansion at sample points. To learn vectors of functions we use the hypothesis space $\mathcal{H}_K^n$ which is an n-fold of $\mathcal{H}_K$: each $\vec{f} \in \mathcal{H}_K^n$ can be written as a column vector of functions $\vec{f} = (f_1, f_2, \ldots, f_n)^T$ with $f_\ell \in \mathcal{H}_K$. Define $\langle \vec{f}, \vec{h} \rangle_K = \sum_{\ell=1}^n \langle f_\ell, h_\ell \rangle_K$. Then $\| \vec{f} \|_K^2 = \sum_{\ell=1}^n \| f_\ell \|_K^2$. The empirical error on sample points $x = x_i, u = x_j$ will be measured by the square loss

$$\big(g(u) - g(x) - \nabla g(x) \cdot (u - x) \big)^2 = \big(y_i - y_j + \vec{f}(x_i) \cdot (x_j - x_i) \big)^2.$$

The restriction $u \approx x$ will be enforced by weights: $w_{i,j} = w_{i,j}^{(s)} > 0$ corresponding to (x_i, x_j) with the requirement that $w_{i,j}^{(s)} \to 0$ as $|x_i - x_j|/s \to \infty$. For $x = (x^1, x^2, \ldots, x^n)^T \in \mathbb{R}^n$, we denote $|x| = \left(\sum_{j=1}^n (x^j)^2\right)^{1/2}$.

One possible choice of weights is given by a Gaussian with variance s. Let $w = w_s$ be the function on $\mathbb{R}^n$ given by $w(x) = \frac{1}{s^{n+2}} e^{-\frac{|x|^2}{2s^2}}$. Then this choice of weights is

$$w_{i,j} = w_{i,j}^{(s)} = \frac{1}{s^{n+2}} e^{-\frac{|x_i - x_j|^2}{2s^2}} = w(x_i - x_j), \qquad i, j = 1, \ldots, m. \tag{6}$$

For regression we define the algorithm by the following optimization problem with weights being arbitrary positive numbers $w_{i,j} = w_{i,j}^{(s)}$ which depend on an index $s > 0$.

Definition 1 *The least-square type learning scheme is defined for the sample $\mathbf{z} \in Z^m$ as*

$$\vec{f}_{\mathbf{z},\lambda} := \arg \min_{\vec{f} \in \mathcal{H}_K^n} \left\{ \frac{1}{m^2} \sum_{i,j=1}^m w_{i,j}^{(s)} \left(y_i - y_j + \vec{f}(x_i) \cdot (x_j - x_i) \right)^2 + \lambda \|\vec{f}\|_K^2 \right\}, \tag{7}$$

where λ, s are two positive constants called the regularization parameters.

A similar algorithm can be defined for classification with a convex loss function $\phi(\cdot)$ like the hinge or logistic loss.

Definition 2 *The regularization scheme for classification is defined for the sample $\mathbf{z} \in Z^m$ as*

$$\vec{f}_{\mathbf{z},\lambda} = \arg \min_{\vec{f} \in \mathcal{H}_K^n} \left\{ \frac{1}{m^2} \sum_{i,j=1}^m w_{i,j}^{(s)} \phi\left(y_i \left(y_j + \vec{f}(x_i) \cdot (x_i - x_j) \right) \right) + \lambda \|\vec{f}\|_K^2 \right\}. \tag{8}$$

Remark 3 *Some algorithms for computing numerical derivative by means of partition were introduced in Wahba and Wendelberger (1980). They work well in low dimensional spaces. In high dimensional spaces, partition is difficult. Our method can be regarded as an algorithm for numerical derivatives in high dimensional spaces.*

At first thought, a natural approach to computing partial derivatives would be to estimate the regression function and then compute partial derivatives. The problem with this approach is that the partial derivatives are no longer in the RKHS of the regression function. This leaves us with the problem of not having a norm or computable metric to work with. The advantage of our method is the derived functions are already approximations of the partial derivatives and they have RKHS inner products which are computed in the estimation process. The inner products reflect the nature of the measure, which is often on a low dimensional manifold embedded in a large dimensional space.

The hypothesis space $\mathcal{H}_K^n$ in the optimization problem (7) may be replaced by some other space of vector-valued functions (Micchelli and Pontil, 2005) in order to learn the gradients.

Remark 4 *Estimation of coordinate covariation is not possible in standard regression models that allow for variable selection such as: recursive feature elimination (RFE) (Guyon et al., 2002), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), and basis pursuits denoising (Chen et al., 1999).*

1.3 Overview

In Sections 2 and 3, we shall derive linear systems for solving the optimization problem (7). In particular, when $m << n$, an efficient algorithm will be provided.

The regularization parameters in (7) depend on m: $\lambda = \lambda(m)$, $s = s(m)$ and generally $\lambda(m), s(m) \to 0$ as m becomes large. In Section 4, we show for a Gaussian weight function (6) how a particular choice of the two regularization parameters leads to rates of convergence of our estimate of the gradient to the true gradient, $\vec{f}_{\mathbf{z},\lambda}$ to ∇f_ρ.

The utility of the algorithm is demonstrated in Section 5 in applications to simulated data as well as gene expression data. We close with a brief discussion in Section 6.

2. Representer Theorem

The optimization problem defining the least-square algorithm (7) can be solved as a linear system of equations. Denote $\mathbb{R}^{p\times q}$ as the space of $p \times q$ matrices, I_n the $n \times n$ identity matrix, and $\mathrm{diag}\{B_1, B_2, \cdots, B_m\}$ the $m \times m$ block diagonal matrix with each $B_i \in \mathbb{R}^{n\times n}$. To save space, we express an mn column vector with blocks $\{c_i \in \mathbb{R}^n\}$ by the following abuse of notion $c = (c_1, c_2, \ldots, c_m)^T$.

The following theorem is an analog of the standard representer theorem (Schoelkopf and Smola, 2001) that states the minimizer of the optimization problem defined by (7) has the form

$$\vec{f}_{\mathbf{z},\lambda} = \sum_{i=1}^{m} c_{i,\mathbf{z}} K_{x_i} \tag{9}$$

with $c_{\mathbf{z}} = (c_{1,\mathbf{z}}, \ldots, c_{m,\mathbf{z}})^T \in \mathbb{R}^{mn}$.

Theorem 5 *For $i = 1, \ldots, m$, let B_i*

$$B_i = \sum_{j=1}^{m} w_{i,j}(x_j - x_i)(x_j - x_i)^T \in \mathbb{R}^{n\times n}, \quad Y_i = \sum_{j=1}^{m} w_{i,j}(y_j - y_i)(x_j - x_i) \in \mathbb{R}^n. \tag{10}$$

Then $\vec{f}_{\mathbf{z},\lambda} = \sum_{i=1}^{m} c_{i,\mathbf{z}} K_{x_i}$ with $c_{\mathbf{z}} = (c_{1,\mathbf{z}}, \ldots, c_{m,\mathbf{z}})^T \in \mathbb{R}^{mn}$ satisfying the linear system

$$\left\{ m^2 \lambda I_{mn} + diag\{B_1, B_2, \cdots, B_m\} \left[K(x_i, x_j) I_n\right]_{i,j=1}^{m} \right\} c = (Y_1, Y_2, \ldots, Y_m)^T. \tag{11}$$

Proof By projecting onto the span of $\{K_{x_i}\}_{i=1}^m$ the reproducing property (2) ensures that $\vec{f}_{\mathbf{z},\lambda} = \sum_{i=1}^{m} c_{i,\mathbf{z}} K_{x_i}$, with $c_{i,\mathbf{z}} \in \mathbb{R}^n$ for each i. Note that $x \cdot v = \sum_{i=1}^{n} x^i v^i = x^T v$ for $x, v \in \mathbb{R}^n$. To find $\{c_{i,\mathbf{z}}\}$, we consider $\vec{f} = \sum_{i=1}^{m} c_i K_{x_i} \in \mathcal{H}_K^n$ with $c_i \in \mathbb{R}^n$. Then

$$\vec{f}(x_i) \cdot (x_j - x_i) = \sum_{p=1}^{m} K(x_p, x_i) c_p \cdot (x_j - x_i) = \sum_{p=1}^{m} K(x_p, x_i)(x_j - x_i)^T c_p$$

and

$$\|\vec{f}\|_K^2 = \sum_{i,j=1}^{m} K(x_i, x_j) c_i \cdot c_j.$$

Define the **empirical error** $\mathcal{E}_{\mathbf{z}}$ as

$$\mathcal{E}_{\mathbf{z}}(\vec{f}) = \frac{1}{m^2} \sum_{i,j=1}^{m} w_{i,j}^{(s)} \left(y_i - y_j + \vec{f}(x_i) \cdot (x_j - x_i) \right)^2.$$

It is a function of mn variables $\{c_q^k : 1 \leq q \leq m, 1 \leq k \leq n\}$ where the q-th coefficient $c_q \in \mathbb{R}^n$ of $\vec{f}$ is expressed as $(c_q^k)_{k=1}^n = (c_q^1, \ldots, c_q^n)^T$. For $q \in \{1, \ldots, m\}$, $k \in \{1, \ldots, n\}$,

$$\begin{aligned}\frac{\partial}{\partial c_q^k}\left\{\mathcal{E}_{\mathbf{z}}(\vec{f}) + \lambda\|\vec{f}\|_K^2\right\} &= 2\lambda\sum_{i=1}^m K(x_q, x_i)c_i^k \\ &+ \frac{2}{m^2}\sum_{i,j=1}^m w_{i,j}\left(y_i - y_j + \sum_{p=1}^m K(x_p, x_i)(x_j - x_i)^T c_p\right)K(x_q, x_i)(x_j^k - x_i^k).\end{aligned}$$

Notice from (2) that for $g, h \in \operatorname{span}\{K_{x_i}\}_{i=1}^m$, $g(x_i) - h(x_i) = 0$ for $i = 1, \ldots, m$ implies that $g - h$ is orthogonal to each K_{x_i}, and hence $g - h = 0$. Then we know that $\vec{f}_{\mathbf{z},\lambda} = \sum_{i=1}^m \tilde{c}_{i,\mathbf{z}} K_{x_i}$ where $\tilde{c}_{\mathbf{z}} = \{\tilde{c}_{i,\mathbf{z}}\}_{i=1}^m$ is the solution to the linear system

$$\lambda c_i + \frac{1}{m^2}\sum_{j=1}^m w_{i,j}\left(y_i - y_j + \sum_{p=1}^m K(x_p, x_i)(x_j - x_i)^T c_p\right)(x_j - x_i) = 0,\ i = 1, \ldots, m.$$

Since $(x_j - x_i)^T c_p$ is a scalar, $[(x_j - x_i)^T c_p](x_j - x_i) = (x_j - x_i)(x_j - x_i)^T c_p$. So the above system can be expressed as

$$B_i \sum_{p=1}^m K(x_i, x_p)c_p + m^2\lambda c_i = Y_i,\ i = 1, \ldots, m. \tag{12}$$

This is exactly the system in (11). ■

Remark 6 *One might consider solving the optimization problem (7) by finding each component of $\vec{f}_{\mathbf{z},\lambda}$ separately. However, to find $(\vec{f}_{\mathbf{z},\lambda})_\ell$ by minimizing over $f \in \mathcal{H}_K$, one needs to replace $y_i - y_j$ by $y_i - y_j + \sum_{k \neq \ell}(\vec{f}_{\mathbf{z},\lambda})_k(x_i)(x_j^k - x_i^k)$. So the optimization problems for components of $\vec{f}_{\mathbf{z},\lambda}$ are not completely separable. It would be interesting to have a separable method for (7).*

3. Reducing the Matrix Size

In some applications of variable selection, the number n of variables is much larger than the sample size m. In such a situation, the system (11) for implementing the learning algorithm (7) is not satisfactory, since the size of the linear system (11) is $(mn) \times (mn)$.

Observe that each term in the summation defining B_i in (10) is a rank one matrix. Hence the rank of the $n \times n$ matrix B_i is at most m for each i. This raises the expectation of reducing the matrix size in the linear system (11). In this section, we show how to reduce this size to $(\mathcal{S}m) \times (\mathcal{S}m)$ with $\mathcal{S} \leq m - 1$. Moreover, an approximation algorithm will be introduced which is often implemented with $\mathcal{S} << m$.

We use the well known approach of singular value decomposition. It may be applied to the coefficient matrix of (11) to reduce the matrix size. Here we prefer to apply the approach to a matrix involving the data only, leaving us flexibility for the weights $w_{i,j}$.

Consider the matrix involving the data $\mathbf{x}$ given by

$$M_{\mathbf{x}} = [x_1 - x_m, x_2 - x_m, \ldots, x_{m-1} - x_m, x_m - x_m] \in \mathbb{R}^{n \times m}. \tag{13}$$

Assume the rank of $M_{\mathbf{x}}$ is d. Then $d \leq \min\{m-1, n\}$ since the last column of the matrix is zero. The theory of singular value decomposition tells us that there exists an $n \times n$ orthogonal matrix

$V = [V_1, V_2, \ldots, V_n]$ and a $m \times m$ orthogonal matrix $U = [U_1, U_2, \ldots, U_m]$ such that

$$M_{\mathbf{x}} = V \Sigma U^T = [V_1 \; V_2 \; \cdots \; V_n] \begin{bmatrix} \mathrm{diag}\{\sigma_1, \sigma_2, \cdots, \sigma_d\} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \\ \vdots \\ U_m^T \end{bmatrix}. \tag{14}$$

Here $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d > \sigma_{d+1} = \ldots = \sigma_{\min\{m,n\}} = 0$ are the singular values of $M_{\mathbf{x}}$. The matrix Σ is $n \times m$ and has entries zero except that $(\Sigma)_{i,i} = \sigma_i$ for $i = 1, \ldots, d$. From expression (14), we see that

$$M_{\mathbf{x}} = \sum_{\ell=1}^{d} \sigma_\ell V_\ell U_\ell^T.$$

Note that $U_\ell^T = [U_\ell^1, \ldots, U_\ell^m]$. The j-th column of $M_{\mathbf{x}}$ equals $x_j - x_m = \sum_{\ell=1}^{d} \sigma_\ell V_\ell U_\ell^j$ and

$$x_j - x_i = \sum_{\ell=1}^{d} \sigma_\ell \left(U_\ell^j - U_\ell^i \right) V_\ell. \tag{15}$$

It follows that $Y_i = \sum_{j=1}^{m} w_{i,j} (y_j - y_i) \sum_{\ell=1}^{d} \sigma_\ell \left(U_\ell^j - U_\ell^i \right) V_\ell$ and

$$B_i = \sum_{j=1}^{m} w_{i,j} \sum_{\ell=1}^{d} \sum_{p=1}^{d} \sigma_\ell \sigma_p \left(U_\ell^j - U_\ell^i \right) \left(U_p^j - U_p^i \right) V_\ell V_p^T. \tag{16}$$

Now we can reduce the matrix size by solving an approximation to the linear system derived from the singular values. A strong correlation among the vectors $\{x_i\}$ would result in a large number of small singular values. If we ignore the small singular values $\sigma_{\mathcal{S}+1}, \ldots, \sigma_d$, the error is proportional to $\sigma_{\mathcal{S}+1}$. This follows from the idea of low-rank approximations in singular value decomposition. The following theorem quantifies the above statement.

Theorem 7 *Assume $|y| \leq M$ almost surely. Denote $\kappa = \sup_{x \in X} \sqrt{K(x,x)}$. Let $1 \leq \mathcal{S} \leq d$. Set*

$$\mathcal{B}_i = \sum_{j=1}^{m} w_{i,j} \left[\sigma_\ell \sigma_p \left(U_\ell^j - U_\ell^i \right) \left(U_p^j - U_p^i \right) \right]_{\ell,p=1}^{\mathcal{S}} \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}, \qquad i = 1, \ldots, m \tag{17}$$

and

$$\mathcal{Y}_i = \sum_{j=1}^{m} w_{i,j} (y_j - y_i) \left[\sigma_\ell \left(U_\ell^j - U_\ell^i \right) \right]_{\ell=1}^{\mathcal{S}} \in \mathbb{R}^{\mathcal{S}}, \qquad i = 1, \ldots, m. \tag{18}$$

Solve the linear system

$$\left\{ m^2 \lambda I_{m\mathcal{S}} + diag\{\mathcal{B}_1, \cdots, \mathcal{B}_m\} \left[K(x_i, x_j) I_{\mathcal{S}} \right]_{i,j=1}^{m} \right\} \hat{b} = (\mathcal{Y}_1, \ldots, \mathcal{Y}_m)^T. \tag{19}$$

The solution $\hat{b}_{\mathbf{z}} = (\hat{b}_{1,\mathbf{z}}, \ldots, \hat{b}_{m,\mathbf{z}})^T \in \mathbb{R}^{m\mathcal{S}}$ gives an approximation $\vec{f}_{\mathbf{z},\lambda,\mathcal{S}} = \sum_{i=1}^{m} b_{i,\mathbf{z}} K_{x_i}$ with $b_{i,\mathbf{z}} = \sum_{\ell=1}^{\mathcal{S}} \hat{b}_{i,\mathbf{z}}^\ell V_\ell$. The error between $b_{\mathbf{z}} = (b_{i,\mathbf{z}})_{i=1}^m$ and $c_{\mathbf{z}}$ can be bounded as

$$\left| b_{\mathbf{z}} - c_{\mathbf{z}} \right|_{\ell^2(\mathbb{R}^{mn})} \leq \frac{4M\sigma_{\mathcal{S}+1}}{m^2 \lambda} \left\{ \sqrt{d\Delta_m} + \frac{2\kappa^2 \sigma_1^2}{m\lambda} \sqrt{\mathcal{S}} \Delta_m \right\}, \tag{20}$$

where $\Delta_m := \max_{1 \leq i \leq m} \left(\sum_{j=1}^{m} w_{i,j} \right)^2 + \sum_{i,j=1}^{m} \left(w_{i,j} \right)^2$.

See Appendix B for the proof.

If we set $\mathcal{S} = d$, we can solve for $\vec{f}_{\mathbf{z},\lambda}$ exactly with a linear system of reduced size. This is stated in the following corollary.

Corollary 8 *Let* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d > 0$ *be all the positive singular values of* $M_{\mathbf{x}}$ *and* U, V *be the orthogonal matrices in (14). Then* $\vec{f}_{\mathbf{z},\lambda} = \sum_{i=1}^{m} \left\{ \sum_{\ell=1}^{d} \widetilde{c}_{i,\mathbf{z}}^{\ell} V_\ell \right\} K_{x_i}$ *where* $\widetilde{c}_{\mathbf{z}}$ *satisfies (19) with* $\mathcal{B}_i$ *and* $\mathcal{Y}_i$ *given by (17) and (18) with* $\mathcal{S} = d$*, respectively.*

Theorem 7 provides a theoretical foundation for the following approximation algorithm with reduced matrix size that accurately approximates $\vec{f}_{\mathbf{z},\lambda}$ by $\vec{f}_{\mathbf{z},\lambda,\mathcal{S}}$.

3.1 Reduced Matrix Size Algorithm

The following is an outline of the reduced matrix algorithm. See appendix C for Matlab® code implementing this algorithm.

Algorithm 1: Approximation algorithm with reduced matrix size to approximate $\vec{f}_{\mathbf{z},\lambda}$

```
input : inputs $(x_i)_{i=1}^m$, labels $(y_i)_{i=1}^m$, kernel $K$, weights $(w_{i,j})$, eigenvalue threshold $\varepsilon > 0$
return: coefficients $(b_{i,\mathbf{z}})_{i=1}^m$

$M_{\mathbf{x}} = [x_1 - x_m, x_2 - x_m, \ldots, x_{m-1} - x_m, x_m - x_m] \in \mathbb{R}^{n \times m}$;
Given $M_{\mathbf{x}}$ compute the singular value decomposition (14) with orthogonal matrices $U, V$ and singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\mathcal{S}} > \varepsilon$;
$t_j = \left(\sigma_1 U_1^j, \ldots, \sigma_{\mathcal{S}} U_{\mathcal{S}}^j\right)^T \in \mathbb{R}^{\mathcal{S}}$ for $1 \leq j \leq m$;
$\mathcal{B}_i = \sum_{j=1}^m w_{i,j}(t_j - t_i)(t_j - t_i)^T$ for $1 \leq i \leq m$;
$\mathcal{Y}_i = \sum_{j=1}^m w_{i,j}(y_j - y_i)(t_j - t_i)$ for $1 \leq i \leq m$;
$$[\widetilde{K}] = \begin{bmatrix} \mathcal{B}_1 K(x_1,x_1) & \mathcal{B}_1 K(x_1,x_2) & \cdots & \mathcal{B}_1 K(x_1,x_m) \\ \mathcal{B}_2 K(x_2,x_1) & \mathcal{B}_2 K(x_2,x_2) & \cdots & \mathcal{B}_2 K(x_2,x_m) \\ \vdots & \ddots & \cdots & \vdots \\ \mathcal{B}_m K(x_m,x_1) & \mathcal{B}_m K(x_m,x_2) & \cdots & \mathcal{B}_m K(x_m,x_m) \end{bmatrix}, \qquad \vec{\mathcal{Y}} = \begin{bmatrix} \mathcal{Y}_1 \\ \mathcal{Y}_2 \\ \vdots \\ \mathcal{Y}_m \end{bmatrix}$$
$\hat{b}_{\mathbf{z}} = (\hat{b}_{1,\mathbf{z}}, \ldots, \hat{b}_{m,\mathbf{z}})^T \in \mathbb{R}^{m\mathcal{S}}$ where
$$\left\{ m^2 \lambda I_{m\mathcal{S}} + [\widetilde{K}] \right\} \hat{b}_{\mathbf{z}} = \vec{\mathcal{Y}} \qquad (21)$$
$b_{i,\mathbf{z}} = \sum_{\ell=1}^{\mathcal{S}} \hat{b}_{i,\mathbf{z}}^{\ell} V_\ell$ and $\vec{f}_{\mathbf{z},\lambda,\mathcal{S}} = \sum_{i=1}^m b_{i,\mathbf{z}} K_{x_i}$ is an approximation of $\vec{f}_{\mathbf{z},\lambda}$;
return $(b_{i,\mathbf{z}})_{i=1}^m$
```

4. Error Analysis

In what follows we use Gaussian weights (equation (6)) and estimate error bounds. We show that $\vec{f}_{\mathbf{z},\lambda} \to \nabla f_\rho$ as $m \to \infty$ for suitable choices of the regularization parameters going to zero, $\lambda = \lambda(m) \to 0, s = s(m) \to 0$. Since we are learning gradients, some regularity conditions on both the marginal distribution and the density are required. The following case illustrates the idea (this case corresponds to the realizable setting in the PAC learning paradigm and will be a corollary of the error analysis that follows).

Proposition 9 *Assume $|y| \leq M$ almost surely. Suppose that for some $0 < \tau \leq 2/3, c_\rho > 0$, the marginal distribution ρ_X satisfies*

$$\rho_X\left(\{x \in X : \inf_{u \in \mathbb{R}^n \setminus X} |u - x| \leq s\}\right) \leq c_\rho^2 s^{4\tau}, \qquad \forall s > 0, \tag{22}$$

and the density $p(x)$ of $d\rho_X(x)$ exists and satisfies

$$\sup_{x \in X} p(x) \leq c_\rho, \quad |p(x) - p(v)| \leq c_\rho |v - x|^\tau, \qquad \forall v, x \in X. \tag{23}$$

Choose $\lambda = \lambda(m) = m^{-\frac{\tau}{n+2+3\tau}}$ and $s = s(m) = (\kappa c_\rho)^{\frac{2}{\tau}} m^{-\frac{1}{n+2+3\tau}}$. If $\nabla f_\rho \in \mathcal{H}_K^n$ and the kernel K is C^3, then there is a constant $C_{\rho,K}$ such that for any $0 < \delta < 1$ and $m \geq 1$, with confidence $1 - \delta$, we have

$$\| \vec{f}_{\mathbf{z},\lambda} - \nabla f_\rho \|_\rho \leq C_{\rho,K} \log\left(\frac{2}{\delta}\right)\left(\frac{1}{m}\right)^{-\frac{\tau}{2(n+2+3\tau)}}. \tag{24}$$

The condition (23) means the density of the marginal distribution is Hölder τ. The condition (22) is about the behavior of ρ_X near the boundary of X. They are natural assumptions for learning gradients of the regression function. When the boundary is piecewise smooth, (23) implies (22).

The idea behind the proof for the convergence of the gradient consists of simultaneously controlling a sample or estimation error term and a regularization or approximation error term. The first term, the sample error, is bounded using a concentration inequality since it is a function of the sample, **z**. The second term, the regularization error, does not depend on the sample and we use functional analysis to bound this quantity.

4.1 Sample Error

First we estimate the sample error by means of the sampling operator introduced in Smale and Zhou (2004, 2006b,a).

Definition 10 *The* **sampling operator** $S_\mathbf{x} : \mathcal{H}_K^n \to \mathbb{R}^{mn}$ *associated with a discrete subset $\mathbf{x} = \{x_i\}_{i=1}^m$ of X is defined by*

$$S_\mathbf{x}(\vec{f}) = \left(\vec{f}(x_i)\right)_{i=1}^m = \left(\vec{f}(x_1), \ldots, \vec{f}(x_m)\right)^T.$$

The adjoint of the sampling operator, $S_\mathbf{x}^T : \mathbb{R}^{mn} \to \mathcal{H}_K^n$, is given by

$$S_\mathbf{x}^T c = \sum_{i=1}^m c_i K_{x_i}, \qquad c = (c_i)_{i=1}^m = (c_1, \ldots, c_m)^T \in \mathbb{R}^{mn}.$$

Denote $D_{\mathbf{x}} = \mathrm{diag}\{B_1, B_2, \cdots, B_m\}$ and $\vec{Y} = (Y_1, Y_2, \ldots, Y_m)^T$.

Consider equation (12) satisfied by $c_{\mathbf{z}}$. The quantity $\sum_{p=1}^m K(x_i, x_p) c_{p,\mathbf{z}}$ equals $\vec{f}_{\mathbf{z},\lambda}(x_i)$. Then (12) implies $\left(S_{\mathbf{x}}^T D_{\mathbf{x}} S_{\mathbf{x}} + m^2 \lambda I\right) \vec{f}_{\mathbf{z},\lambda} = S_{\mathbf{x}}^T \vec{Y}$. Therefore,

$$\vec{f}_{\mathbf{z},\lambda} = \left(\frac{1}{m^2} S_{\mathbf{x}}^T D_{\mathbf{x}} S_{\mathbf{x}} + \lambda I\right)^{-1} \frac{1}{m^2} S_{\mathbf{x}}^T \vec{Y}. \tag{25}$$

We introduce an s-generalization error corresponding to the empirical error as follows.

Definition 11 *The s-**generalization error** $\mathcal{E} = \mathcal{E}_s$ is defined for vectors of functions as*

$$\mathcal{E}(\vec{f}) = \int_Z \int_Z w(x-u)\left(y - v + \vec{f}(x)\cdot(u-x)\right)^2 d\rho(x,y) d\rho(u,v).$$

If we denote $\sigma_s^2 = \int_Z \int_Z w(x-u)(y - f_\rho(x))^2 d\rho(x,y) d\rho(u,v)$, then

$$\mathcal{E}(\vec{f}) = 2\sigma_s^2 + \int_X \int_X w(x-u)\left[f_\rho(x) - f_\rho(u) + \vec{f}(x)\cdot(u-x)\right]^2 d\rho_X(x) d\rho_X(u). \tag{26}$$

A data independent limit of $\vec{f}_{\mathbf{z},\lambda}$ is

$$\vec{f}_\lambda = \arg\min_{\vec{f} \in \mathcal{H}_K^n} \left\{\mathcal{E}(\vec{f}) + \lambda \|\vec{f}\|_K^2\right\}. \tag{27}$$

Taking the functional derivatives, we know from (26) that $\vec{f}_\lambda$ can be expressed in terms of the following integral operator on the space $\left(L^2_{\rho_X}\right)^n$ with norm $\|\vec{f}\|_\rho = \left(\|f_\ell\|_\rho^2\right)^{1/2}$.

Proposition 12 *Let $L_{K,s} : \left(L^2_{\rho_X}\right)^n \to \left(L^2_{\rho_X}\right)^n$ be the integral operator defined by*

$$L_{K,s}\vec{f} = \int_X \int_X w(x-u)(u-x)K_x(u-x)^T \vec{f}(x) d\rho_X(x) d\rho_X(u). \tag{28}$$

It is a positive operator on $\left(L^2_{\rho_X}\right)^n$ and

$$\vec{f}_\lambda = \left(L_{K,s} + \lambda I\right)^{-1} \vec{f}_{\rho,s}. \tag{29}$$

where

$$\vec{f}_{\rho,s} := \int_X \int_X w(x-u)(u-x)K_x\left(f_\rho(u) - f_\rho(x)\right) d\rho_X(x) d\rho_X(u). \tag{30}$$

The operator $L_{K,s}$ has its range in $\mathcal{H}_K^n$. It can also be regarded as a positive operator on $\mathcal{H}_K^n$. We shall use the same notion for the operators on these two different domains.

To bound the sample error $\vec{f}_{\mathbf{z},\lambda} - \vec{f}_\lambda$, we shall introduce a McDiarmid-Bernstein type probability inequality for vector-valued random variables.

Proposition 13 *Let $\mathbf{z} = \{z_i\}_{i=1}^m$ be i.i.d. draws from a probability distribution* ρ *on Z, $(H, \|\cdot\|)$ be a Hilbert space, and $F : Z^m \to H$ be measurable. If there is $\widetilde{M} \geq 0$ such that $\|F(\mathbf{z}) - E_{z_i}(F(\mathbf{z}))\| \leq \widetilde{M}$ for each $1 \leq i \leq m$ and almost every $\mathbf{z} \in Z^m$, then for every $\varepsilon > 0$,*

$$Prob_{\mathbf{z}\in Z^m}\left\{\|F(\mathbf{z}) - E_{\mathbf{z}}(F(\mathbf{z}))\| \geq \varepsilon\right\} \leq 2\exp\left\{-\frac{\varepsilon^2}{2(\widetilde{M}\varepsilon + \sigma^2)}\right\}, \tag{31}$$

where $\sigma^2 := \sum_{i=1}^m \sup_{\mathbf{z}\setminus\{z_i\}\in Z^{m-1}} E_{z_i}\{\|F(\mathbf{z}) - E_{z_i}(F(\mathbf{z}))\|^2\}$. For any $0 < \delta < 1$, with confidence $1-\delta$, there holds

$$\|F(\mathbf{z}) - E_{\mathbf{z}}(F(\mathbf{z}))\| \leq 2\log\frac{2}{\delta}\{\widetilde{M} + \sqrt{\sigma^2}\}.$$

Proof Apply Theorem 3.3 of Pinelis (1994) (see Appendix A) to the finite sequence $\{f_j = E_{z_m,\ldots,z_j}F - E_{z_m,\ldots,z_1}F : j = 1, 2, \ldots, m+1\}$. So $f_1 = 0$ and $f_{m+1} = F(\mathbf{z}) - E_{\mathbf{z}}(F(\mathbf{z}))$. Note that $\|f_j - f_{j-1}\| \leq \widetilde{M}$ almost surely and $\sum_{j=2}^{m+1} E_{z_{j-1}}\|f_j - f_{j-1}\|^2 \leq \sigma^2$. The conditions of Theorem 3.3 of Pinelis (1994) hold with $B^2 = \sigma^2, \Gamma = \widetilde{M}$. As pointed out in a correction of Pinelis (1999), the probability should be $2\exp\{-\frac{r^2}{r\Gamma + B^2 + B\sqrt{B^2 + 2r\Gamma}}\}$, which is bounded by $2\exp\{-\frac{r^2}{2(r\Gamma + B^2)}\}$. Inequality (31) follows from the theorem.

Choose ε such that $\frac{\varepsilon^2}{2\widetilde{M}\varepsilon + 2\sigma^2} = \log\frac{2}{\delta}$. That is, ε satisfies

$$\varepsilon^2 = 2\widetilde{M}\log\frac{2}{\delta}\varepsilon + 2\sigma^2\log\frac{2}{\delta}.$$

Therefore, with confidence at least $1-\delta$, we have

$$\|F(\mathbf{z}) - E_{\mathbf{z}}(F(\mathbf{z}))\| \leq \varepsilon \leq 2\widetilde{M}\log\frac{2}{\delta} + \sqrt{2\sigma^2\log\frac{2}{\delta}} \leq 2\log\frac{2}{\delta}\{\widetilde{M} + \sqrt{\sigma^2}\}.$$

This proves the proposition. ■

Now we can give the main result on the sample error $\|\vec{f}_{\mathbf{z},\lambda} - \vec{f}_\lambda\|_K$. Denote the diameter of X as $\mathrm{Diam}(X) = \max_{x,t\in X}|x - t|$ and the moments of the Gaussian as

$$J_p := \int_{\mathbb{R}^n} e^{-\frac{|x|^2}{2}}|x|^p dx, \qquad p \geq 0.$$

In the following sample error estimates, the bounds are valid for any m, λ, and s, though they yield reasonable learning rates only for suitable choices of $\lambda = \lambda(m)$ and $s = s(m)$ which we state in Proposition 9.

Theorem 14 *Assume $|y| \leq M$ almost surely.*

1. *For any $0 < \delta < 1$, with confidence $1-\delta$, we have*

$$\|\vec{f}_{\mathbf{z},\lambda} - \vec{f}_\lambda\|_K \leq \frac{16\kappa Diam(X)\log(2/\delta)}{\sqrt{m}\lambda s^{n+2}}\left\{2M + \kappa Diam(X)\|\vec{f}_\lambda\|_K\right\} + \frac{1}{m}\|\vec{f}_\lambda\|_K. \tag{32}$$

2. *If the density $p(x)$ of $d\rho_X(x)$ exists and satisfies $\sup_{x\in X} p(x) \leq c_\rho$, then for any $0 < s \leq 1$, with confidence $1-\delta$, there holds*

$$\begin{aligned}\|\vec{f}_{\mathbf{z},\lambda} - \vec{f}_\lambda\|_K \quad \leq \quad & \frac{8\kappa\log(2/\delta)}{\sqrt{m}\lambda s^{1+n/2}}\left(\sqrt{c_\rho} + \frac{Diam(X)}{\sqrt{m}s^{1+n/2}}\right) \\ & \left(2M\sqrt{J_2} + \kappa(Diam(X) + \sqrt{J_4})\|\vec{f}_\lambda\|_K\right) + \frac{1}{m}\|\vec{f}_\lambda\|_K. \end{aligned} \tag{33}$$

Proof By (25), we have

$$\vec{f}_{\mathbf{z},\lambda} - \vec{f}_\lambda = \big(\frac{1}{m^2} S_{\mathbf{x}}^T D_{\mathbf{x}} S_{\mathbf{x}} + \lambda I\big)^{-1} \left\{ \frac{1}{m^2} S_{\mathbf{x}}^T \vec{Y} - \frac{1}{m^2} S_{\mathbf{x}}^T D_{\mathbf{x}} S_{\mathbf{x}} \vec{f}_\lambda - \lambda \vec{f}_\lambda \right\}.$$

Define a vector-valued function $F : Z^m \to \mathcal{H}_K^n$ by

$$F(\mathbf{z}) = \frac{1}{m^2} S_{\mathbf{x}}^T \vec{Y} - \frac{1}{m^2} S_{\mathbf{x}}^T D_{\mathbf{x}} S_{\mathbf{x}} \vec{f}_\lambda.$$

That is,

$$F(\mathbf{z}) = \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m w_{i,j}(x_j - x_i) K_{x_i}(y_j - y_i) - \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m w_{i,j}(x_j - x_i) K_{x_i}(x_j - x_i)^T \vec{f}_\lambda(x_i).$$

By independence, the expected value of $F(\mathbf{z})$ equals

$$\begin{aligned} & \frac{1}{m^2} \sum_{i=1}^m E_{\mathbf{z}_i} \sum_{j \neq i} E_{\mathbf{z}_j} \left\{ w_{i,j}(x_j - x_i) K_{x_i} \left[(y_j - y_i) - (x_j - x_i)^T \vec{f}_\lambda(x_i) \right] \right\} \\ = \quad & \frac{m-1}{m^2} \sum_{i=1}^m E_{\mathbf{z}_i} \left\{ \int_X w(x_i - u)(u - x_i) K_{x_i} \left[(f_\rho(u) - y_i) - (u - x_i)^T \vec{f}_\lambda(x_i) \right] d\rho_X(u) \right\}. \end{aligned}$$

It follows that

$$E_{\mathbf{z}}(F(\mathbf{z})) = \frac{m-1}{m} \vec{f}_{\rho,s} - \frac{m-1}{m} L_{K,s} \vec{f}_\lambda.$$

By (29), $L_{K,s} \vec{f}_\lambda + \lambda \vec{f}_\lambda = \vec{f}_{\rho,s}$. Hence $\lambda \vec{f}_\lambda = \vec{f}_{\rho,s} - L_{K,s} \vec{f}_\lambda = \frac{m}{m-1} E_{\mathbf{z}}(F(\mathbf{z}))$. Therefore,

$$\|\vec{f}_{\mathbf{z},\lambda} - \vec{f}_\lambda\|_K \leq \frac{1}{\lambda} \|F(\mathbf{z}) - \frac{m}{m-1} E_{\mathbf{z}}(F(\mathbf{z}))\|_K \leq \frac{1}{\lambda} \|F(\mathbf{z}) - E_{\mathbf{z}}(F(\mathbf{z}))\|_K + \frac{1}{m} \|\vec{f}_\lambda\|_K.$$

The reproducing property (2) together with the upper bound κ implies

$$\|f\|_\rho \leq \|f\|_\infty \leq \kappa \|f\|_K, \qquad \forall f \in \mathcal{H}_K. \tag{34}$$

Then we apply Proposition 13 to the function $F(\mathbf{z})$ to get our error bound.

Let $i \in \{1, \ldots, m\}$. We know that $F(\mathbf{z}) - E_{z_i}(F(\mathbf{z}))$ equals

$$\begin{aligned} & \frac{1}{m^2} \sum_{j \neq i} w(x_i - x_j)(x_j - x_i) \Big\{ K_{x_i} \left[y_j - y_i - (x_j - x_i)^T \vec{f}_\lambda(x_i) \right] \\ & \qquad + K_{x_j} \left[y_j - y_i - (x_j - x_i)^T \vec{f}_\lambda(x_j) \right] \Big\} \\ & - \frac{1}{m^2} \sum_{j \neq i} \int_X w(x - x_j)(x_j - x) \Big\{ K_x \left[y_j - f_\rho(x) - (x_j - x)^T \vec{f}_\lambda(x) \right] \\ & \qquad + K_{x_j} \left[y_j - f_\rho(x) - (x_j - x)^T \vec{f}_\lambda(x_j) \right] \Big\} d\rho_X(x). \end{aligned}$$

Using (34) for $\vec{f}_\lambda$ and $|x - x_j| \leq \mathrm{Diam}(X)$ for any $x \in X$, we see that

$$\|F(\mathbf{z}) - E_{z_i}(F(\mathbf{z}))\|_K \leq \widetilde{M} = \frac{4\kappa \mathrm{Diam}(X)}{m s^{n+2}} \left\{ 2M + \kappa \mathrm{Diam}(X) \|\vec{f}_\lambda\|_K \right\}.$$

1. We first prove (32). Apply the trivial bound $\sigma^2 \le m\widetilde{M}^2$. Then Proposition 13 tells us that for any $0<\delta<1$, with confidence $1-\delta$, there holds

$$\|F(\mathbf{z})-E_{\mathbf{z}}(F(\mathbf{z})\|_K \le 2\log\frac{2}{\delta}\left\{\widetilde{M}+\sqrt{m}\widetilde{M}\right\} \le 4\log\frac{2}{\delta}\sqrt{m}\widetilde{M}.$$

This proves (32).

2. To prove (33) we need to improve on our estimate of the variance σ^2, we bound $\|F(\mathbf{z})-E_{z_i}(F(\mathbf{z}))\|_K$ by

$$\frac{1}{m^2}\sum_{j\ne i} w(x_i-x_j)|x_j-x_i|\left\{2\kappa 2M+2|x_j-x_i|\kappa^2\|\vec{f}_\lambda\|_K\right\}$$
$$+\frac{1}{m^2}\sum_{j\ne i}\int_X w(x-x_j)|x_j-x|\left\{2\kappa 2M+2|x_j-x|\kappa^2\|\vec{f}_\lambda\|_K\right\}d\rho_X(x).$$

It follows that $\left(E_{z_i}(\|F(\mathbf{z})-E_{z_i}(F(\mathbf{z}))\|_K^2)\right)^{1/2}$ is bounded by

$$\frac{2}{m^2}\sum_{j\ne i}\left\{\int_X \left(w(x-x_j)\right)^2|x_j-x|^2\left\{4\kappa M+2|x_j-x|\kappa^2\|\vec{f}_\lambda\|_K\right\}^2 d\rho_X(x)\right\}^{1/2}$$
$$\le \frac{2}{m^2}\sum_{j\ne i}\left\{\int_X s^{-2(n+2)}e^{-\frac{|x-x_j|^2}{s^2}}|x_j-x|^2\left\{4\kappa M\right\}^2 c_\rho dx\right\}^{1/2}$$
$$+\frac{2}{m^2}\sum_{j\ne i}\left\{\int_X s^{-2(n+2)}e^{-\frac{|x-x_j|^2}{s^2}}|x_j-x|^4(2\kappa^2)^2\|\vec{f}_\lambda\|_K^2 c_\rho dx\right\}^{1/2}.$$

Here we have used the assumption $d\rho_X(x)=p(x)dx$ with $p(x)\le c_\rho$. Bounding the above integrals by those on the whole space $\mathbb{R}^n$, we see from the definition of the moments M_r that

$$E_{z_i}(\|F(\mathbf{z})-E_{z_i}(F(\mathbf{z}))\|_K^2) \le \frac{2(m-1)}{m^2}\left\{4\kappa M\sqrt{c_\rho}\sqrt{\frac{J_2}{s^{n+2}2^{1+n/2}}}+2\kappa^2\|\vec{f}_\lambda\|_K\sqrt{c_\rho}\sqrt{\frac{J_4}{s^n 2^{2+n/2}}}\right\}.$$

It follows then that for $s\le 1$

$$\sigma^2 \le \frac{16c_\rho\kappa^2}{ms^{n+2}}\left\{2^{1/4}M\sqrt{J_2}+\kappa\|\vec{f}_\lambda\|_K\sqrt{J_4}\right\}^2.$$

The second statement (inequality (33)) follows from Proposition 13. ■

4.2 Regularization Error

In this subsection, we shall bound the regularization error $\|\vec{f}_\lambda-\nabla f_\rho\|$ by a functional analysis approach. To illustrate the idea, we state the result for a special case when $\nabla f_\rho\in\mathcal{H}_K^n$. It is a corollary of Theorem 17 and Theorem 19 with $r=1/2$.

Proposition 15 *Assume (22) and (23). Denote $V_\rho=\int_X(p(x))^2dx>0$. Suppose that $\nabla f_\rho\in\mathcal{H}_K^n$ and for some $c_\rho'>0$,*

$$|f_\rho(u)-f_\rho(x)-\nabla f_\rho(x)\cdot(u-x)|\le c_\rho'|u-x|^2,\ \forall\, u,x\in X. \tag{35}$$

Then for any $\lambda > 0$ *and* $0 < s \leq \min\big\{\big\{2\kappa^2 c_\rho\big(M_{2+\tau}+J_4+c_\rho J_2\big)\big\}^{1/\tau}\lambda^{1/\tau}, 1\big\}$, *there holds*

$$\|\vec{f}_\lambda - \nabla f_\rho\|_\rho \leq \big(\kappa^2 c'_\rho J_3\big)\frac{s}{\lambda} + \big\{2\big(V_\rho n(2\pi)^{n/2}\big)^{-1/2}\|\nabla f_\rho\|_K\big\}\sqrt{\lambda}.$$

To estimate the regularization error, we need to consider the convergence of $L_{k,s}$ as $s \to 0$.

Lemma 16 *Assume that for some* $0 < \tau < 1$, *conditions (22) and (23) hold. Then* $V_\rho \leq c_\rho$ *and for any* $0 < s \leq 1$ *we have*

$$\|L_{K,s} - V_\rho n(2\pi)^{n/2} L_K\|_{\mathcal{H}_K^n \to \mathcal{H}_K^n} \leq s^\tau \kappa^2 c_\rho\big(M_{2+\tau}+J_4+c_\rho J_2\big), \tag{36}$$

where L_K *is a positive operator on* $\mathcal{H}_K^n$ *defined by*

$$L_K \vec{f} = \int_X K_x \vec{f}(x)\frac{p(x)}{V_\rho} d\rho_X(x). \tag{37}$$

The operator L_K *is also a positive operator on* $(L^2_{\rho_X})^n$ *satisfying*

$$\|L_{K,s} - V_\rho n(2\pi)^{n/2} L_K\|_{(L^2_{\rho_X})^n \to (L^2_{\rho_X})^n} \leq s^\tau \kappa^2 c_\rho\big(M_{2+\tau}+J_4+c_\rho J_2\big), \qquad \forall\, 0 < s \leq 1. \tag{38}$$

Proof Let $\vec{f} \in (L^2_{\rho_X})^2$. Denote

$$\vec{g} = \int_X \left\{\int_X w(x-u)(u-x)K_x(u-x)^T du\right\} p(x)\vec{f}(x)\, d\rho_X(x).$$

Then by (23) and the Cauchy-Schwartz inequality we see that $\|L_{K,s}\vec{f} - \vec{g}\|_K$ is bounded by

$$\int_X \left\{\int_X \frac{1}{s^n} e^{-\frac{|u-x|^2}{2s^2}} \big|\frac{u-x}{s}\big|^2 \|K_x\|_K c_\rho |u-x|^\tau du\right\} \big|\vec{f}(x)\big|\, d\rho_X(x) \leq s^\tau \kappa c_\rho M_{2+\tau}\|\vec{f}\|_\rho.$$

Observe that $n(2\pi)^{n/2} = J_2$ and $\int_{\mathbb{R}^n} w(u-x)(u^i-x^i)(u^j-x^j)du = 0$ when $i \neq j$. Then $\int_{\mathbb{R}^n} \frac{1}{s^n} e^{-\frac{|u-x|^2}{2s^2}} \big(\frac{u-x}{s}\big)\big(\frac{u-x}{s}\big)^T du = J_2 I_n$. Hence

$$V_\rho n(2\pi)^{n/2} L_K \vec{f} = \int_X \left\{\int_{\mathbb{R}^n} \frac{1}{s^n} e^{-\frac{|u-x|^2}{2s^2}} \big(\frac{u-x}{s}\big)\big(\frac{u-x}{s}\big)^T du\right\} p(x) K_x \vec{f}(x) d\rho_X(x).$$

It follows that

$$\begin{aligned} \|\vec{g} - V_\rho n(2\pi)^{n/2} L_K \vec{f}\|_K &= \left\|\int_X \left\{\int_{\mathbb{R}^n \setminus X} \frac{1}{s^n} e^{-\frac{|u-x|^2}{2s^2}} \big(\frac{u-x}{s}\big)\big(\frac{u-x}{s}\big)^T du\right\} p(x) K_x \vec{f}(x) d\rho_X(x)\right\|_K \\ &\leq \int_X \left\{\int_{\mathbb{R}^n \setminus X} \frac{1}{s^n} e^{-\frac{|u-x|^2}{2s^2}} \big|\frac{u-x}{s}\big|^2 du\right\} \kappa |\vec{f}(x)| p(x) d\rho_X(x). \end{aligned}$$

Separate the domain X into $X_s := \{x \in X : \inf_{u \in \mathbb{R}^n \setminus X} \; |u-x| \leq \sqrt{s}\}$, consisting of those points whose distance to the boundary is at most $\sqrt{s}$, and its complement $X \setminus X_s$.

If $x \in X \setminus X_s$, any $u \in \mathbb{R}^n \setminus X$ satisfies $|u-x| \geq \sqrt{s}$ and thereby $1 \leq s\big|\frac{u-x}{s}\big|^2$. Hence

$$\int_{\mathbb{R}^n \setminus X} \frac{1}{s^n} e^{-\frac{|u-x|^2}{2s^2}} \big|\frac{u-x}{s}\big|^2 du \leq s \int_{\mathbb{R}^n \setminus X} \frac{1}{s^n} e^{-\frac{|u-x|^2}{2s^2}} \big|\frac{u-x}{s}\big|^4 du \leq s J_4.$$

It follows from (23) that

$$\int_{X\setminus X_s}\left\{\int_{\mathbb{R}^n\setminus X}\frac{1}{s^n}e^{-\frac{|u-x|^2}{2s^2}}|\frac{u-x}{s}|^2du\right\}\kappa|\vec{f}(x)|p(x)d\rho_X(x)\le s\kappa c_\rho M_4\int_{X\setminus X_s}|\vec{f}(x)|d\rho_X(x)$$

which is bounded by $s\kappa c_\rho M_4\|\vec{f}\|_\rho$.

For the subset X_s, we use the Cauchy-Schwartz inequality and (23) to obtain

$$\int_{X_s}\left\{\int_{\mathbb{R}^n\setminus X}\frac{1}{s^n}e^{-\frac{|u-x|^2}{2s^2}}|\frac{u-x}{s}|^2du\right\}\kappa|\vec{f}(x)|p(x)d\rho_X(x)\le\int_{X_s}\kappa c_\rho M_2|\vec{f}(x)|d\rho_X(x).$$

This is bounded by $\kappa c_\rho M_2\sqrt{\rho_X(X_s)}\|\vec{f}\|_\rho$. By (22), $\rho_X(X_s)\le c_\rho^2 s^{2\tau}$. Thus, for $0<s\le 1$,

$$\|\vec{g}-n(2\pi)^{n/2}L_K\vec{f}\|_K\le s^\tau\kappa c_\rho\big(M_4+c_\rho J_2\big)\|\vec{f}\|_\rho.$$

Combine the above two estimates. There holds for any $0<s\le 1$

$$\|L_{K,s}\vec{f}-V_\rho n(2\pi)^{n/2}L_K\vec{f}\|_K\le s^\tau\kappa c_\rho\big(M_{2+\tau}+J_4+c_\rho J_2\big)\|\vec{f}\|_\rho$$

which proves (38) by (34). When $\vec{f}\in\mathcal{H}_K^n$, we have from (34) again that

$$\|L_{K,s}\vec{f}-V_\rho n(2\pi)^{n/2}L_K\vec{f}\|_K\le s^\tau\kappa^2 c_\rho\big(M_{2+\tau}+J_4+c_\rho J_2\big)\|\vec{f}\|_K.$$

This verifies (36) and proves the lemma. ■

The measure $\frac{p(x)}{V_\rho}d\rho_X$ is probability one on X. So we know (see Cucker and Smale (2001)) that the operator L_K can be used to define the reproducing kernel Hilbert space: Let L_K^r be the r-th power of the positive operator L_K on $(L^2_{\rho_X})^n$ having range in $\mathcal{H}_K^n$. Then $\mathcal{H}_K^n$ is the range of $L_K^{1/2}$: $\|\vec{f}\|_\rho=\|L_K^{1/2}\vec{f}\|_K$ for any $\vec{f}\in(L^2_{\rho_X})^n$.

Theorem 17 *Under the assumption (35), we have*

$$\|\vec{f}_\lambda-\nabla f_\rho+\lambda\big(L_{K,s}+\lambda I\big)^{-1}\nabla f_\rho\|_K\le\frac{s}{\lambda}\kappa c'_\rho J_3.$$

Proof By (29), we find that

$$\vec{f}_\lambda-\nabla f_\rho+\lambda\big(L_{K,s}+\lambda I\big)^{-1}\nabla f_\rho=\big(L_{K,s}+\lambda I\big)^{-1}\left\{\vec{f}_{\rho,s}-L_{K,s}\nabla f_\rho\right\}.$$

Then

$$\|\vec{f}_\lambda-\nabla f_\rho+\lambda\big(L_{K,s}+\lambda I\big)^{-1}\nabla f_\rho\|_K\le\|\big(L_{K,s}+\lambda I\big)^{-1}\|_{\mathcal{H}_K^n\to\mathcal{H}_K^n}\|\vec{f}_{\rho,s}-L_{K,s}\nabla f_\rho\|_K$$

which is bounded by $\frac{1}{\lambda}\|\vec{f}_{\rho,s}-L_{K,s}\nabla f_\rho\|_K$. Using (35) on the integral

$$\vec{f}_{\rho,s}-L_{K,s}\nabla f_\rho=\int_X\int_X w(x-u)(u-x)K_x\left\{f_\rho(u)-f_\rho(x)-(u-x)^T\nabla f_\rho(x)\right\}d\rho_X(x)d\rho_X(u),$$

we know that

$$\|\vec{f}_{\rho,s}-L_{K,s}\nabla f_\rho\|_K\le\int_X\int_X w(x-u)|u-x|\|K_x\|_K c'_\rho|u-x|^2d\rho_X(x)d\rho_X(u)\le s\kappa c'_\rho J_3.$$

This proves the theorem. ■

Finally, we need to study $\lambda\big(L_{K,s}+\lambda I\big)^{-1}\nabla f_\rho$ in order to estimate the error $\|\vec{f}_\lambda-\nabla f_\rho\|$.

Lemma 18 *Assume (22) and (23). Denote* $c''_\rho = \big(2\kappa^2 c_\rho \big(M_{2+\tau}+J_4+c_\rho J_2\big)\big)^{1/\tau}$*. Then*

$$\|\big(L_{K,s}+\lambda I\big)^{-1}\vec{f}\| \le 2\|\big(V_\rho n(2\pi)^{n/2}L_K+\lambda I\big)^{-1}\vec{f}\|, \qquad \forall 0<s\le \min\big\{c''_\rho\lambda^{1/\tau},1\big\},$$

where $\vec{f}$ *is either in the space* $\mathcal{H}_K^n$ *or in* $(L^2_{\rho_X})^n$*, and* $\|\cdot\|$ *is the corresponding norm.*

Proof Write $\big(L_{K,s}+\lambda I\big)^{-1}\vec{f} = \big\{\big[V_\rho n(2\pi)^{n/2}L_K+\lambda I\big]-\big[n(2\pi)^{n/2}L_K-L_{K,s}\big]\big\}^{-1}\vec{f}$ as

$$\Big\{I-\big[V_\rho n(2\pi)^{n/2}L_K+\lambda I\big]^{-1}\big[V_\rho n(2\pi)^{n/2}L_K-L_{K,s}\big]\Big\}^{-1}\big[V_\rho n(2\pi)^{n/2}L_K+\lambda I\big]^{-1}\vec{f}.$$

This in connection with Lemma 16 implies

$$\|\big(L_{K,s}+\lambda I\big)^{-1}\vec{f}\| \le \Big\{1-\frac{1}{\lambda}s^\tau\kappa^2 c_\rho\big(M_{2+\tau}+J_4+c_\rho J_2\big)\Big\}^{-1}\|\big[V_\rho n(2\pi)^{n/2}L_K+\lambda I\big]^{-1}\vec{f}\|.$$

This verifies the lemma. ■

Lemma 18 yields the convergence of $\|\lambda\big(L_{K,s}+\lambda I\big)^{-1}\nabla f_\rho\|$. The convergence rates require some conditions on ∇f_ρ relative to the pair $(L^2_{\rho_X},\mathcal{H}_K)$. The assumption we shall use is $\|L_K^{-r}\nabla f_\rho\|_\rho < \infty$. It means that ∇f_ρ lies in the range of L_K^r. In particular, in the case $r=1/2$, the condition $\|L_K^{-1/2}\nabla f_\rho\|_\rho<\infty$ means $\nabla f_\rho \in \mathcal{H}_K^n$. For more examples about this condition, see Smale and Zhou (2006a).

Theorem 19 *Assume (22), (23), and (35). Let* $0<s\le\min\big\{c''_\rho\lambda^{1/\tau},1\big\}$*. If* $\|L_K^{-r}\nabla f_\rho\|_\rho<\infty$ *for some* $0<r\le 1$*, then*

$$\|\lambda\big(L_{K,s}+\lambda I\big)^{-1}\nabla f_\rho\|_\rho \le 2\lambda^r\big(V_\rho n(2\pi)^{n/2}\big)^{-r}\|L_K^{-r}\nabla f_\rho\|_\rho, \qquad \forall\lambda>0.$$

If moreover $r\ge 1/2$*, then we have for any* $\lambda>0$*,*

$$\|\lambda\big(L_{K,s}+\lambda I\big)^{-1}\nabla f_\rho\|_K \le 2\lambda^{r-1/2}\big(V_\rho n(2\pi)^{n/2}\big)^{-r}\|L_K^{-r}\nabla f_\rho\|_\rho.$$

In the general situation, we can see that $\|\lambda\big(L_{K,s}+\lambda I\big)^{-1}\nabla f_\rho\|_\rho\to 0$ as $\lambda\to 0$, provided that $\mathcal{H}_K$ is dense in $L^2_{\rho_X}$ (Smale and Zhou, 2003). This can be seen from the following convergence estimate.

Proposition 20 *Assume (22), (23), and (35). Then*

$$\|\lambda\big(L_{K,s}+\lambda I\big)^{-1}\nabla f_\rho\|_\rho \le 2\mathcal{K}\left(\nabla f_\rho, \frac{\sqrt{\lambda}}{V_\rho n(2\pi)^{n/2}}\right), \qquad \forall\ 0<s\le\min\big\{c''_\rho\lambda^{1/\tau},1\big\},$$

where $\mathcal{K}(\vec{f},t)$ *is the K-functional of the pair* $\big((L^2_{\rho_X})^n,\mathcal{H}_K^n\big)$ *defined as*

$$\mathcal{K}(\vec{f},t) = \inf_{\vec{g}\in\mathcal{H}_K^n}\Big\{\|\vec{f}-\vec{g}\|_\rho + t\|\vec{g}\|_K\Big\}, \qquad t>0. \tag{39}$$

The proof of Proposition 9 shows how our error analysis can be applied.

Proof of Proposition 9. Since the kernel K is C^3 and $\nabla f_\rho \in \mathcal{H}_K^n$, we know from Zhou (2003) that $\frac{\partial f_\rho}{\partial x^i}$ is C^1 for each i. It follows that f_ρ is C^2 and condition (35) is satisfied for some constant $c'_\rho > 0$.

Since $\lambda = \big(1/m\big)^\gamma$ with $\gamma = \frac{\tau}{n+2+3\tau}$ and $s = (\kappa c_\rho)^{2/\tau}\lambda^{1/\tau}$, we see from the fact $J_2 > 1$ that for $m \geq (\kappa c_\rho)^{2(n+2+3\tau)/\tau}$, the restriction $0 < s \leq \min\big\{\big\{2\kappa^2 c_\rho\big(M_{2+\tau} + J_4 + c_\rho J_2\big)\big\}^{1/\tau}\lambda^{1/\tau}, 1\big\}$ in Proposition 15 and Lemma 18 is satisfied. Then by Proposition 15, since $\frac{1}{\tau} - 1 \geq \frac{1}{2}$, we have for some constant $C_\rho > 0$ that

$$\|\vec{f}_\lambda - \nabla f_\rho\|_\rho \leq C_\rho\Big(\frac{s}{\lambda} + \sqrt{\lambda}\Big) \leq C_\rho(1 + (\kappa c_\rho)^{2/\tau})\Big(\frac{1}{m}\Big)^{\frac{\gamma}{2}}.$$

Applying Lemma 18, we know that

$$\|\lambda\big(L_{K,s} + \lambda I\big)^{-1}\nabla f_\rho\|_K \leq 2\|\lambda\big(V_\rho n(2\pi)^{n/2}L_K + \lambda I\big)^{-1}\nabla f_\rho\|_K \leq 2\|\nabla f_\rho\|_K.$$

This in connection with Theorem 17 implies that

$$\|\vec{f}_\lambda\|_K \leq \|\nabla f_\rho\|_K + 2\|\nabla f_\rho\|_K + \frac{s}{\lambda}\kappa c'_\rho J_3 \leq 3\|\nabla f_\rho\|_K + (\kappa c_\rho)^{2/\tau}\kappa c'_\rho J_3.$$

Finally, we apply (33) of Theorem 14 and know that for a constant $C'_\rho > 0$, with confidence $1 - \delta$,

$$\|\vec{f}_{\mathbf{z},\lambda} - \vec{f}_\lambda\|_K \leq C'_\rho\Big\{\frac{\log(2/\delta)}{\sqrt{m}\lambda s^{1+n/2}} + \frac{1}{m}\Big\} \leq C'_\rho \log(2/\delta)\Big\{\Big(\frac{1}{m}\Big)^{\frac{1}{2}-\gamma-\frac{\gamma}{2}(1+\frac{n}{2})}(\kappa c_\rho)^{-\frac{2}{\tau}(1+\frac{n}{2})} + \frac{1}{m}\Big\}.$$

which is bounded by $C''_\rho \log\big(\frac{2}{\delta}\big)\big(\frac{1}{m}\big)^{-\frac{\tau}{2(n+2+3\tau)}}$ with a constant C''_ρ. This is true for $m \geq (\kappa c_\rho)^{2(n+2+3\tau)/\tau}$. Replacing the constant C''_ρ by a new one enables us to bound errors for the finitely many terms with $m < (\kappa c_\rho)^{2(n+2+3\tau)/\tau}$. Thus Proposition 9 is proved. ■

5. Simulated Data and Gene Expression Data

In this section we apply the least-squares gradient algorithm (7) to the variable selection and variable covariance problems. Our idea is to rank the importance of variables according to the norm of their partial derivatives $\|\frac{\partial f_\rho}{\partial x^\ell}\|$, since a small norm implies small changes on the function with respect to this variable. By our error analysis, we expect $\vec{f}_{\mathbf{z},\lambda} \approx \nabla f_\rho$. So we shall use the norms of the components of $\vec{f}_{\mathbf{z},\lambda}$ to rank the variables.

Definition 21 *The relative magnitude of the norm for the variables is defined as*

$$s_\ell^\rho = \frac{\|\big(\vec{f}_{\mathbf{z},\lambda}\big)_\ell\|_K}{\big(\sum_{j=1}^n \|\big(\vec{f}_{\mathbf{z},\lambda}\big)_j\|_K^2\big)^{1/2}}, \qquad \ell = 1, \ldots, n.$$

In the same way, we can study coordinate covariances by the variance of an empirical matrix.

Definition 22 *The* **empirical gradient matrix** *(EGM), $F_{\mathbf{z}}$, is the $n \times m$ matrix whose columns are $\vec{f}_{\mathbf{z},\lambda}(x_j)$ with $j = 1, \ldots, m$. The* **empirical covariance matrix** *(ECM), $\Xi_{\mathbf{z}}$, is the $n \times n$ matrix of inner products of the directional derivative of two coordinates*

$$Cov(\vec{f}_{\mathbf{z},\lambda}) := \Big[\langle\big(\vec{f}_{\mathbf{z},\lambda}\big)_p, \big(\vec{f}_{\mathbf{z},\lambda}\big)_q\rangle_K\Big]_{p,q=1}^n = \sum_{i,j=1}^m c_{i,\mathbf{z}}c_{j,\mathbf{z}}^T K(x_i, x_j).$$

The ECM gives us the covariance between the coordinates while the EGM gives us information as how the variables differ over different sections of the space.

We apply our idea to three data sets. The first data set is an artificial one which we use to illustrate the procedure. The second is a cancer classification problem that has been well studied and serves as further confirmation of the utility of the method. The third data set provides a gold standard as to relevant variables.

5.1 Artificial Data

We construct a function in an $n = 80$ dimensional space which consists of three linear functions over different partitions of the space. We generate 30 samples as follows:

1. For samples $\{x_i\}_{i=1}^{10}$
$$x^j \sim \mathcal{N}(1,\sigma_x), \text{ for } j = 1,\ldots,10; \qquad x^j \sim \mathcal{N}(0,\sigma_x), \text{ for } j = 11,\ldots,80.$$
2. For samples $\{x_i\}_{i=11}^{20}$
$$x^j \sim \mathcal{N}(1,\sigma_x), \text{ for } j = 11,\ldots,20; \qquad x^j \sim \mathcal{N}(0,\sigma_x), \text{ for } j = 1,\ldots,10,21,\ldots,80.$$
3. For samples $\{x_i\}_{i=21}^{30}$
$$x^j \sim \mathcal{N}(1,\sigma_x), \text{ for } j = 41,\ldots,50; \qquad x^j \sim \mathcal{N}(0,\sigma_x), \text{ for } j = 1,\ldots,40,51,\ldots,80.$$

A draw of this x matrix is shown in figure (1a). Three vectors with support over different dimensions were constructed as follows:

$$\begin{aligned}
w_1 &= 2+.5\sin(2\pi i/10) \text{ for } i=1,...,10 \text{ and } 0 \text{ otherwise},\\
w_2 &= -2-.5\sin(2\pi i/10) \text{ for } i=11,...,20 \text{ and } 0 \text{ otherwise},\\
w_3 &= -2-.5\sin(2\pi i/10) \text{ for } i=41,...,50 \text{ and } 0 \text{ otherwise}.
\end{aligned}$$

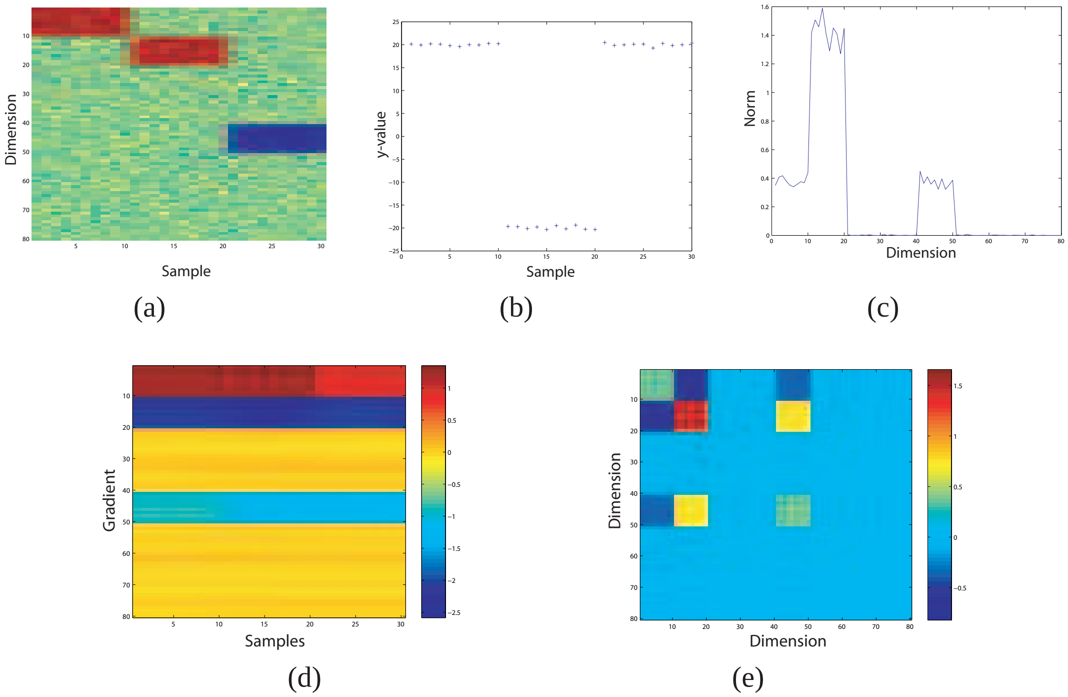
The values for $\{y_i\}_{i=1}^{30}$ were given by the following linear equations

1. For samples $\{y_i\}_{i=1}^{10}$
$$y_i = x_i \cdot w_1 + \mathcal{N}(0,\sigma_y),$$
2. For samples $\{y_i\}_{i=11}^{20}$
$$y_i = x_i \cdot w_2 + \mathcal{N}(0,\sigma_y),$$
3. For samples $\{y_i\}_{i=21}^{30}$
$$y_i = x_i \cdot w_3 + \mathcal{N}(0,\sigma_y).$$

A draw of the y values is shown in figure (1b).

In figure (1c) we plot the norm of each component of the estimate of the gradient, $\{\|(\vec{f}_{\mathbf{z},\lambda})_\ell\|_K\}_{\ell=1}^{80}$ for $\sigma_x = .05$ and $\sigma_y = .30$. The norm of each component gives an indication of the importance of a variable and variables with small norms can be eliminated. Note that the coordinates with nonzero norm are the ones we expect, $\ell = 1,\ldots,20,41,\ldots,50$.

Perhaps more interesting is that we can evaluate the gradient at each sample $\{x_i\}_{i=1}^{m}$. This leads to an estimate of the covariation of the variables. In figure (1d) we plot the EGM, while the ECM is displayed in figure (1e). The blocking structure of the ECM indicates the coordinates that covary.



(a)

(b)

(c)

(d)

(e)

Figure 1: a) The data matrix x where each sample corresponds to a column, b) the vector of y values generated by sampling the function, c) the RKHS norm for each dimension, d) an estimate of the gradient at each sample, the samples correspond to columns, e) the empirical covariance matrix.

5.2 Gene Expression Data

In computational biology, specifically in the subfield of gene expression analysis variable selection and estimation of covariation is of fundamental importance. Microarray technologies enable experimenters to measure the expression level of thousands of genes, the entire genome, at once. The expression level of a gene is proportional to the number of copies of mRNA transcribed by that gene. This readout of gene expression is considered a proxy of the state of the cell. The goals of gene expression analysis include using the expression level of the genes to predict classes, for example tissue morphology or treatment outcome, or real-valued quantities such as toxicity or sensitivity. Fundamental to understanding the biology giving rise to the outcome or toxicity is determining which genes are most relevant for the prediction.

5.2.1 LEUKEMIA CLASSIFICATION

We apply our procedure to a well studied expression data set. The data set is a result of a study using expression data to discriminate acute myeloid leukemia (AML) from acute lymphoblastic leukemia (ALL) (Golub et al., 1999; Slonim et al., 2000) and estimating the genes most relevant to this discrimination. The data set contains 48 samples of AML and 25 samples of ALL. Expression levels of $n = 7,129$ genes and expressed sequence tags (ESTs) were measured via an oligonucleotide microarray for each sample. This data set was split into a training set of 38 samples and a test set of 35 samples.

Various variable selection algorithms have been applied to this data set by using the training set specified in Golub et al. (1999) to select variables and build a classification model and then compute the classification error on the test set. We estimate $\vec{f}_{\mathbf{z},\lambda}$ from the training data and then select the $\mathcal{S}$ variables with the largest s_ℓ^ρ. We then use a linear Support Vector Machine (SVM) to build a classification model and compute the accuracy on the test set. Table 1 reports test errors for various values of $\mathcal{S}$. The classification accuracy is very similar to other feature selection algorithms such as recursive feature elimination (RFE) (Guyon et al., 2002; Lee et al., 2004) and radius-margin bound (RMB) (Chapelle et al., 2002) both of which were developed specifically for SVMs.

genes (S)	5	55	105	155	205	255	305	355	405	455
test errors	1	3	2	1	1	1	1	1	1	1

Table 1: Number of errors in classification for various values of $\mathcal{S}$ using the genes corresponding to dimensions with the largest norms. A linear SVM was used for classification.

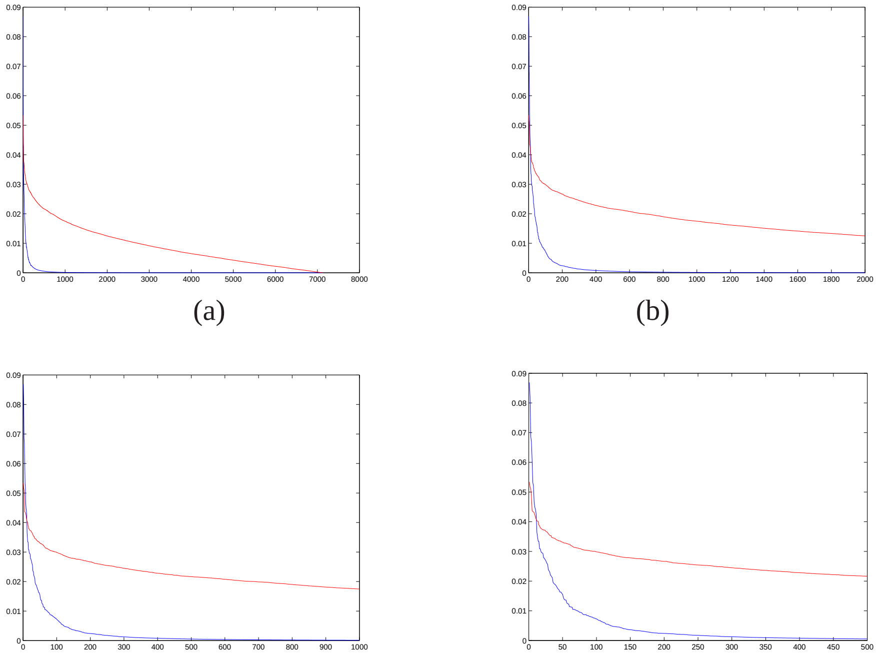
In figure (2a-d) we plot the relative magnitude sequence s_ℓ^ρ for the genes. On this data set the decay in the ranked scores $s_{(\ell)}^\rho$ is steeper than that for most statistics that have been previously used on this data. To illustrate this we compared the gradient score to the Fisher score Slonim et al. (2000) for each gene

$$t_\ell = \frac{|\hat{\mu}_\ell^{\text{AML}} - \hat{\mu}_\ell^{\text{ALL}}|}{\hat{\sigma}_\ell^{\text{AML}} + \hat{\sigma}_\ell^{\text{ALL}}},$$

where $\hat{\mu}_\ell^{\text{AML}}$ is the mean expression level for the AML samples in the ℓ-th gene, $\hat{\mu}_\ell^{\text{ALL}}$ is the mean expression level for the ALL samples in the ℓ-th gene, $\hat{\sigma}_\ell^{\text{AML}}$ is the standard deviation of the expression level for the AML samples in the ℓ-th gene, and $\hat{\sigma}_\ell^{\text{ALL}}$ is the standard deviation of the expression level for the ALL samples in the ℓ-th gene. We then normalize these scores

$$s^F_\ell = \frac{t_\ell}{\left(\sum_{p=1}^n t_p^2\right)^{1/2}}.$$

Figure (2a-d) displays the relative decay of $s^\rho_{(\ell)}$ and $s^F_{(\ell)}$ over various numbers of dimensions. In all plots it is apparent that the decay rate of $s^\rho_{(\ell)}$ is much steeper. Plotting the decay of the elements for the normalized hyperplane $\frac{w^0}{\|w^0\|}$ that is the solution of a linear SVM results in a plot much more like that of the Fisher score than the gradient statistic. Whether and how this steepness (sparsity) has an implication on the generalization error is an open question.



(a)

(b)

(c)

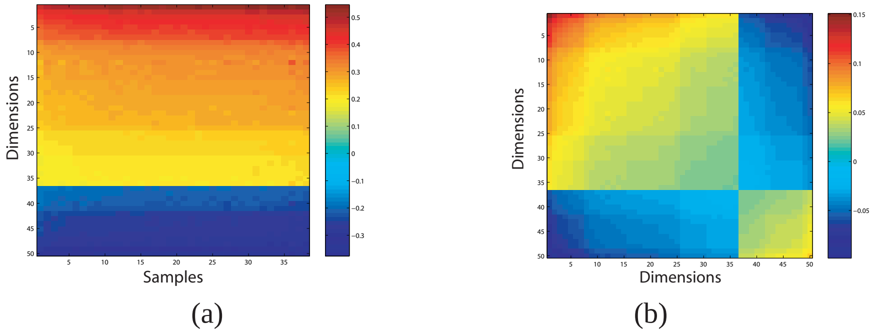
(d)

Figure 2: The decay of $s^\rho_{(\ell)}$ (blue) and $s^F_{(\ell)}$ (red) over: a) all the genes/dimensions, b) the top 2000 genes/dimensions, c) the top 1000 genes/dimensions, d) the top 500 genes/dimensions.

We can also examine the EGM and the ECM. The EGM in this case is a $7,129 \times 38$ matrix and the ECM is $7,129 \times 7,129$ matrix. We plot the EGM in the space of the dimensions corresponding to the top 50 norms ordered by a clustering metric in figure (3a). The covariation in the coordinates is plotted for the top 50 dimensions ordered in the same way as the EGM (see figure (3b)). The blocking structure of the matrix gives us coordinate covariance.

5.2.2 GENDER: "A GOLD STANDARD"

In this section we assess the accuracy of the algorithm with respect to a data set for which a priori biological knowledge gives us a set of important variables. This serves as a gold standard.

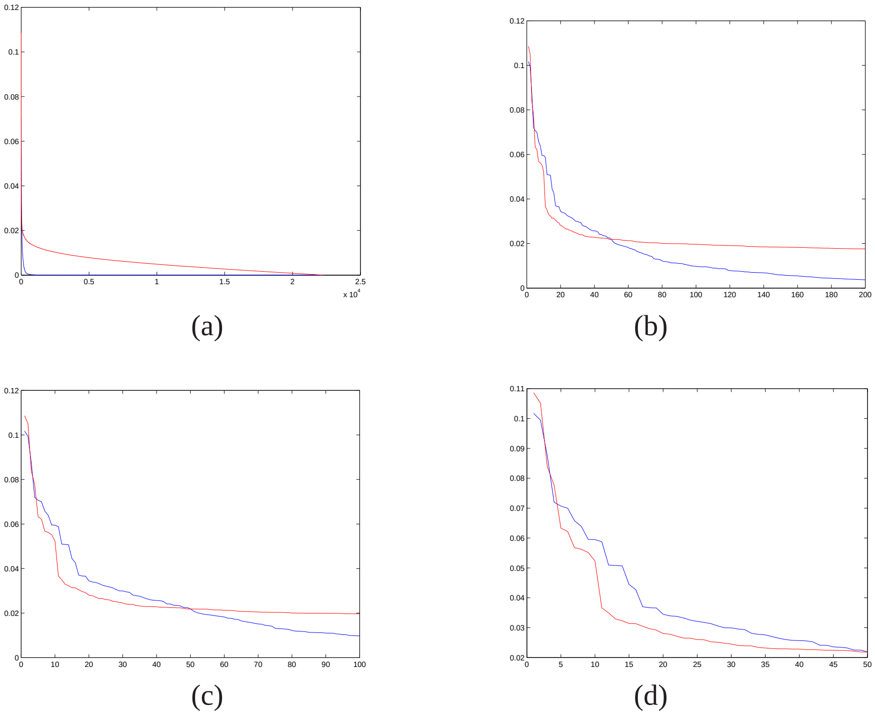


(a) (b)

Figure 3: The a) EGM for the top 50 dimensions ordered by clustering the EGM and b) the ECM for the top 50 dimensions ordered in the same way.

We examine a gene expression data set with 15 male and 17 females samples from lymphblastoid cell lines (unpublished). Expression levels of $n = 22,283$ probes corresponding to genes and expressed sequence tags (ESTs) were measured via an oligonucleotide microarray for each sample.

In figure (4a-d) we plot the relative magnitude sequence s_ℓ^ρ for the genes as compared to those of the relative Fisher score s_ℓ^F and we see again the quicker decay for the gradient norms.



(a) (b) (c) (d)

Figure 4: The decay of $s_{(\ell)}^\rho$ (blue) and $s_{(\ell)}^F$ (red) over: a) all the genes/dimensions, b) the top 200 genes/dimensions, c) the top 100 genes/dimensions, d) the top 50 genes/dimensions.

From a priori biological knowledge we would predict that the most discriminative genes for gender would be those on the Y chromosome as well as genes on the X chromosome known to escape X inactivation. The reason that all the genes on the X chromosome would not be expected to be discriminative is due to dosage compensation in expression which takes compensates for the fact that women have two X chromosomes and men have one. The mechanism for this compensation is X inactivation. However, there are genes known to escape X inactivation and these should be differentially expressed. We obtained a list of such genes by combining lists reported in two sources (Carrel et al., 1999; Disteche et al., 2002). There were 35 probes in the X inactivation set and 66 probes corresponding to genes on the Y chromosome.

An important caveat is that while these 101 probes would be expected to be differentially expressed they would not all be expected to rank at the top of a list of genes that are differentially expressed. This is due to the fact that in the cell line or tissue of question there may be other genes that are more strongly differentially expressed due to local conditions. This is why the term gold standard is quoted.

We first used a standard variation filter (Slonim et al., 2000) which reduced the number of probes to about $12,000$. This data set was then standardized (the expression values for each gene was recentered and scaled to be zero mean and standard deviation of one). We then iteratively ran our procedure 20 times, each time removing the bottom 10% of the probes. We found that 16 of the 101 probes appeared in the top ranked 500 probes. Ranking by the Fisher score we found 22 of the top 101 probes in the top ranked 500 probes. Using the logistic loss may result in more like the Fisher score since it is a more appropriate model for classification. Both results are significant with respect to a hypergeometric distribution as the model for the null hypothesis. However, the assumptions of independence in the model which gives rise to the hypergeometric distribution are completely inappropriate in this problem (the probes tend to be strongly correlated). There are statistical tests that account for the correlations but this topic is beyond the scope of this paper (Sweet-Cordero et al., 2005; Subramanian et al., 2005).

6. Discussion

We introduce an algorithm that learns gradients from samples of function values and show its relevance to variable selection. An error analysis is given for the convergence of the estimated gradient to the true gradient. This method also places the problem of variable selection into the powerful framework of Tikhonov regularization. There are many extensions and refinements to this method which we discuss below:

1. Logistic regression model: In Definition 2 we state an algorithm for classification. As many applications of this method are for classification problems it is important to implement a reduced matrix version of this algorithm as was done for regression by Algorithm 1. In addition, an error analysis for the classification setting is also necessary.

2. Fully Bayesian model: The Tikhonov regularization framework coupled with the use of an RKHS allows us to implement a fully Bayesian version of the procedure in the context of Bayesian radial basis (RB) models Liao et al. (2005); Liao (2005). The Bayesian RB framework can be extended to develop a proper probability model for the gradient learning problem. The optimization procedures 1 and 2 would be replaced by Markov Chain Monte-carlo methods and the full posterior rather than the maximum a posteriori estimate would be computed. A very useful result of this is that in addition to the point estimates for the gradient we would also be able to compute confidence intervals.

3. Intrinsic dimension: In Proposition 9 the rate of convergence of the gradient has the form of $O(m^{-1/n})$ which can be extremely slow if n is large. However, in most data sets and when variable selection is meaningful the data are concentrated on a much lower dimensional manifold embedded in the high dimensional space. In this setting an analysis that replaces the ambient dimension n with the intrinsic dimension of the manifold $n_{\mathcal{M}}$ would be of great interest.

4. Semi-supervised setting: Intrinsic properties of the manifold X can be further studied by unlabelled data. This is one of the motivations of semi-supervised learning. In many applications, it is much easier to obtain unlabelled data with a larger sample size $u >> m$. For our purpose, unlabelled data $\mathbf{x} = (x_i)_{i=m+1}^{m+u}$ can be used to reduce the dimension or correlation. Since we learn the gradient by $\vec{f}$, it is natural to use the unlabelled data to control the approximate norm of $\vec{f}$ in some Sobolev spaces and introduce a semi-supervised learning algorithm as

$$\begin{aligned}\vec{f}_{\mathbf{z},\mathbf{x},\lambda,\mu} \quad &= \arg\min_{\vec{f}\in\mathcal{H}_K^n} \left\{ \tfrac{1}{m^2} \textstyle\sum_{i,j=1}^m w_{i,j}^{(s)} \left(y_i - y_j + \vec{f}(x_i)\cdot(x_j - x_i) \right)^2 \right. \\ &\qquad \left. + \tfrac{\mu}{(m+u)^2} \textstyle\sum_{i,j=1}^{m+u} W_{i,j} |\vec{f}(x_i) - \vec{f}(x_j)|_{\ell^2(\mathbb{R}^n)}^2 + \lambda \|\vec{f}\|_K^2 \right\}, \qquad (40)\end{aligned}$$

where $\{W_{i,j}\}$ are edge weights in the data adjacency graph, μ is another regularization parameter and often satisfies $\lambda = o(\mu)$.

Acknowledgments

We would like to thank André Elisseeff and Misha Belkin for useful discussions. We would like to thank Aravind Subramanian for help with the X inactivation sets. We would like to acknowledge support for this project from the National Science Foundation and from the University Grants Council of Hong Kong (Project No. CityU 103303) and the Institute for Genome Sciences & Policy at Duke.

Appendix A

Let $\mathcal{S}(X)$ denote the class of all sequences $f = (f_0, f_1, \ldots)$ of Bochner-integrable random vectors in X with $f_0 \equiv 0$, defined on a probability space. Let $\mathcal{M}(X)$ denote the class of all sequences $f_j \in \mathcal{S}(X)$ that are martingales. The following theorem can be found in (Pinelis, 1994) as Theorem 3.3 with a correction made in Pinelis (1999). Note that Hilbert spaces are $(2,D)$-smooth Banach spaces with $D = 1$.

Theorem 23 *(Pinelis, 1994) Suppose that $f \in \mathcal{M}(X)$, X is a $(2,D)$-smooth separable Banach space and*

$$\left\| \sum_{j=1}^{\infty} E_{j-1} \|f_j - f_{j-1}\|^m \right\|_\infty \leq m!\Gamma^{m-2} B^2/(2D^2)$$

for some $\Gamma > 0, B > 0$ and $m = 2,3,\ldots$ Then for all $r \geq 0$,

$$Prob(\sup_j \|f_j\| \geq r) \leq 2\exp\left(-\frac{r^2}{\Gamma r + B^2 + B\sqrt{B^2 + 2\Gamma r}} \right).$$

Appendix B

We give the proof for Theorem 7.
Proof We divide our approximation in three steps.

Step 1. Approximate $c_{\mathbf{z}}$ by $\widetilde{c}_{\mathbf{z}}$ which is defined by

$$\widetilde{c}_{\mathbf{z}} = \left(m^2\lambda I_{mn} + \mathrm{diag}\{B_1,\cdots,B_m\}\left[K(x_i,x_j)I_n\right]_{i,j=1}^m\right)^{-1}(\widetilde{\mathcal{Y}}_1,\ldots,\widetilde{\mathcal{Y}}_m)^T,$$

where $\widetilde{\mathcal{Y}}_i = \sum_{j=1}^m w_{i,j}(y_j-y_i)\sum_{\ell=1}^{\mathcal{S}}\sigma_\ell\left(U_\ell^j - U_\ell^i\right)V_\ell$. For each i,

$$\left|\widetilde{\mathcal{Y}}_i - Y_i\right|_{\ell^2(\mathbb{R}^n)} \le \sum_{j=1}^m w_{i,j}2M\sigma_{\mathcal{S}+1}\left(\sum_{\ell=\mathcal{S}+1}^d \left(U_\ell^j-U_\ell^i\right)^2\right)^{1/2}.$$

Since the matrix U is orthogonal, we know that $\sum_{\ell=1}^m\left(U_\ell^j\right)^2 = 1$ for each j and $\sum_{j=1}^m\left(U_\ell^j\right)^2 = 1$ for each ℓ. By the Schartz inequality, $\left|\widetilde{\mathcal{Y}}_i - Y_i\right|^2_{\ell^2(\mathbb{R}^n)}$ is bounded by

$$(4M\sigma_{\mathcal{S}+1})^2\left\{\sum_{j=1}^m\left(w_{i,j}\right)^2\cdot\sum_{j=1}^m\sum_{\ell=\mathcal{S}+1}^d\left(U_\ell^j\right)^2 + \left(\sum_{j=1}^m w_{i,j}\right)^2\sum_{\ell=\mathcal{S}+1}^d\left(U_\ell^i\right)^2\right\}.$$

It follows that

$$\left|c_{\mathbf{z}} - \widetilde{c}_{\mathbf{z}}\right|_{\ell^2(\mathbb{R}^{mn})} \le \frac{1}{m^2\lambda}\left\{\sum_{i=1}^m\left|\widetilde{\mathcal{Y}}_i - Y_i\right|^2_{\ell^2(\mathbb{R}^n)}\right\}^{1/2} \le \frac{4M\sigma_{\mathcal{S}+1}\sqrt{d-\mathcal{S}}}{m^2\lambda}\sqrt{\Delta_m}.$$

Step 2. Approximate $\widetilde{c}_{\mathbf{z}}$ by $\widetilde{b}_{\mathbf{z}}$ which is defined by

$$\widetilde{b}_{\mathbf{z}} = \left(m^2\lambda I_{mn} + \mathrm{diag}\{\widetilde{B}_1,\cdots,\widetilde{B}_m\}\left[K(x_i,x_j)I_n\right]_{i,j=1}^m\right)^{-1}(\widetilde{\mathcal{Y}}_1,\ldots,\widetilde{\mathcal{Y}}_m)^T,$$

where

$$\widetilde{B}_i = \sum_{j=1}^m w_{i,j}\sum_{\ell=1}^{\mathcal{S}}\sum_{p=1}^{\mathcal{S}}\sigma_\ell\sigma_p\left(U_\ell^j - U_\ell^i\right)\left(U_p^j - U_p^i\right)V_\ell V_p^T.$$

For $b \in \mathbb{R}^n$, the vector $\left(B_i - \widetilde{B}_i\right)b$ equals

$$\left\{\sum_{\ell=\mathcal{S}+1}^d\sum_{p=1}^{\mathcal{S}} + \sum_{\ell=1}^d\sum_{p=\mathcal{S}+1}^d\right\}\sigma_\ell\sigma_p\sum_{j=1}^m w_{i,j}\left(U_\ell^j - U_\ell^i\right)\left(U_p^j - U_p^i\right)(V_p^T b)V_\ell.$$

By the Schwartz inequality, the $\ell^2(\mathbb{R}^n)$ norm of the first term above is bounded by

$$\begin{aligned} &\sum_{j=1}^m w_{i,j}\sigma_{\mathcal{S}+1}\left\{\sum_{\ell=\mathcal{S}+1}^d\left(U_\ell^j - U_\ell^i\right)^2\left(\sum_{p=1}^{\mathcal{S}}\sigma_p\left(U_p^j - U_p^i\right)(V_p^T b)\right)^2\right\}^{1/2} \\ \le\ &\sum_{j=1}^m w_{i,j}\sigma_{\mathcal{S}+1}\left\{\sum_{\ell=\mathcal{S}+1}^d\left(U_\ell^j - U_\ell^i\right)^2\right\}^{1/2}|b|_{\ell^2(\mathbb{R}^n)}\left\{\sum_{p=1}^{\mathcal{S}}\sigma_p^2\left(U_p^j - U_p^i\right)^2\right\}^{1/2}. \end{aligned}$$

This is at most $2\sigma_{\mathcal{S}+1}\sigma_1|b|_{\ell^2(\mathbb{R}^n)}\sum_{j=1}^m w_{i,j}$. The $\ell^2(\mathbb{R}^n)$ norm of the second term in the expression of $\big(B_i - \widetilde{B}_i\big)b$ can be bounded in the same way and we thus have

$$\left\|\big(B_i - \widetilde{B}_i\big)b\right\|_{\ell^2(\mathbb{R}^n)} \le 4\sigma_{\mathcal{S}+1}\sigma_1|b|_{\ell^2(\mathbb{R}^n)}\sum_{j=1}^m w_{i,j}.$$

Then we have the following estimate for the operator norm of the difference of the diagonal operators

$$\left\|\operatorname{diag}\{B_1,\cdots,B_m\} - \operatorname{diag}\{\widetilde{B}_1,\cdots,\widetilde{B}_m\}\right\| \le 4\sigma_{\mathcal{S}+1}\sigma_1 \max_{1\le i\le m}\sum_{j=1}^m w_{i,j}.$$

It follows that

$$\begin{aligned}&\left\|\operatorname{diag}\{B_1,\cdots,B_m\}\left[K(x_i,x_j)I_n\right]_{i,j=1}^m - \operatorname{diag}\{\widetilde{B}_1,\cdots,\widetilde{B}_m\}\left[K(x_i,x_j)I_n\right]_{i,j=1}^m\right\| \\ &\le 4\kappa^2 m\sigma_{\mathcal{S}+1}\sigma_1 \max_{1\le i\le m}\sum_{j=1}^m w_{i,j}.\end{aligned}$$

Notice that for two invertible operators L_1, L_2 on a Hilbert space, there holds

$$L_1^{-1} - L_2^{-1} = L_1^{-1}(L_2 - L_1)L_2^{-1}.$$

Hence

$$\|L_1^{-1} - L_2^{-1}\| \le \|L_1^{-1}\|\ \|L_2 - L_1\|\ \|L_2^{-1}\|.$$

Applying this to our setting, we have

$$\left|\widetilde{b}_{\mathbf{z}} - \widetilde{c}_{\mathbf{z}}\right|_{\ell^2(\mathbb{R}^{mn})} \le \frac{4\kappa^2 m\sigma_{\mathcal{S}+1}\sigma_1}{(m^2\lambda)^2}\left\{\max_{1\le i\le m}\sum_j w_{i,j}\right\}\left\|(\widetilde{\mathcal{Y}}_1,\ldots,\widetilde{\mathcal{Y}}_m)^T\right\|_{\ell^2(\mathbb{R}^{mn})}.$$

For each i, we have

$$\left|\widetilde{\mathcal{Y}}_i\right|_{\ell^2(\mathbb{R}^n)} \le 2M\sum_{j=1}^m w_{i,j}\left\{\left(\sum_{\ell=1}^{\mathcal{S}}\sigma_\ell^2\big(U_\ell^j\big)^2\right)^{1/2} + \left(\sum_{\ell=1}^{\mathcal{S}}\sigma_\ell^2\big(U_\ell^i\big)^2\right)^{1/2}\right\}.$$

It follows that

$$\left|\widetilde{b}_{\mathbf{z}} - \widetilde{c}_{\mathbf{z}}\right|_{\ell^2(\mathbb{R}^{mn})} \le \frac{8M\kappa^2 m\sigma_{\mathcal{S}+1}\sigma_1^2\sqrt{\mathcal{S}}}{(m^2\lambda)^2}\Delta_m.$$

Step 3. Find the coefficients $\widetilde{b}_{\mathbf{z}}$. The linear system it satisfies is

$$m^2\lambda\widetilde{b}_{i,\mathbf{z}} + \sum_{q=1}^m\sum_{j=1}^m w_{i,j}\sum_{\ell=1}^{\mathcal{S}}\sum_{p=1}^{\mathcal{S}}\sigma_\ell\sigma_p\left(U_\ell^j - U_\ell^i\right)\left(U_p^j - U_p^i\right)V_\ell V_p^T K(x_i,x_q)\widetilde{b}_{q,\mathbf{z}} = \widetilde{\mathcal{Y}}_i,$$

where $i = 1,\ldots,m$. Since $\widetilde{\mathcal{Y}}_i$ lies in $\operatorname{span}\{V_\ell\}_{\ell=1}^{\mathcal{S}}$, we know that each $\widetilde{b}_{i,\mathbf{z}}$ also lies in this subspace of $\mathbb{R}^n$. That is, there is a vector $b_{i,\mathbf{z}}^* \in \mathbb{R}^{\mathcal{S}}$ such that

$$\widetilde{b}_{i,\mathbf{z}} = \sum_{\ell=1}^{\mathcal{S}} b_{i,\mathbf{z}}^{*\ell}V_\ell, \qquad i = 1,\ldots,m.$$

Substituting this expression into the linear system for $\widetilde{b}_{\mathbf{z}}$, we know that $b^*_{\mathbf{z}}$ can be solved by the linear system

$$m^2\lambda b^{*\ell}_{i,\mathbf{z}} + \sum_{q=1}^{m}\sum_{j=1}^{m} w_{i,j} \sum_{p=1}^{\mathcal{S}} \sigma_\ell \sigma_p \left(U^j_\ell - U^i_\ell\right)\left(U^j_p - U^i_p\right) K(x_i, x_q) b^{*p}_{q,\mathbf{z}}$$
$$= \sum_{j=1}^{m} w_{i,j}(y_j - y_i)\sigma_\ell \left(U^j_\ell - U^i_\ell\right), \qquad 1 \le \ell \le \mathcal{S}, 1 \le i \le m.$$

This is exactly the linear system (19). Therefore, $\hat{b}_{i,\mathbf{z}} = b^*_{i,\mathbf{z}}$ for each i and $\widetilde{b}_{\mathbf{z}} = b_{\mathbf{z}}$ is the desired coefficients for the function $\vec{f}_{\mathbf{z},\lambda,\mathcal{S}}$. ■

Appendix C

The following is Matlab® code that implements algorithm (1), the approximation algorithm with reduced matrix size. The code could be made more efficient by exploiting the vector nature of Matlab. However, we include the version with loops for transparency.

```
% a matrix x that is dim by m where m is the number of samples
% a vector y that is m by 1
% eps is a constraint on the ratio of the top s eigenvalues to the sum over
%         all eigenvalues
% lambda is the regularization constant
% sigma is the variance of the weight matrix computed automatically from the
%         data
% F is the gradient evaluated at each sample again a dim by m matrix
% nrm is the RKHS norm for each dimension


function [F,nrm,sigma] =
solveder(x,y,lambda,eps)

[dim,m] = size(x);

% this subroutine computes distances between all pairs and sets sigma to the
%         median
a = zeros(m,m);
 for i=1:m
   for j=1:m
       a(i,j) = norm(x(:,i)-x(:,j));
   end
 end
 sigma = median(median(a));

% this subroutine computes the weight matrix
a = zeros(m,m);
```

```
 for i=1:m
   for j=1:m
       a(i,j) = (1/(sigma*sqrt(2*pi)))*exp(-norm(x(:,i)-x(:,j))^2/(2*sigma^2));
   end
 end

% the kernel matrix is computed will add nonlinear version
K = zeros(m,m); K = transpose(x)*x;

% constructs the matrix of differences between all points
M = zeros(dim,m); for i=1:m
     M(:,i) = x(:,i)-x(:,m);
end


 % computes the eigenvalues and eigenvectors of M^t M
 % and keeps s eigenvectors as specified by eps
 d = eig(K);
 W = transpose(M)*M;
 [V,d] = eig(W);
 d = diag(d);
 vals = cumsum(d);
 inds = find(vals/vals(m) < eps);
 s = m-max(inds);

 % since matlab indexes eigenvalues from smallest to largest we reverse
 U = zeros(m,m);
 dp = zeros(m,1);
 for i=1:m
    U(:,m-i) = V(:,i);
    dp(i) = d(m-i);
 end

 % projects of the paired differences into the subspace of the s eigenfunctions
t = zeros(s,m); for i=1:m
     t(:,i) = sqrt(dp(1:s)).*transpose(U(i,1:s));
 end

 Ktilde = zeros(m*s,m*s);
 ytilde = zeros(m*s,1);

 % computes the Ktilde matrix and the vector script Y
 for i=1:m
    Bmat = zeros(s,s);
    yv = zeros(s,1);
    for j=1:m
        Bmat = Bmat+a(i,j)* (t(:,j)-t(:,i))*(transpose(t(:,j)-t(:,i)));
        yv = yv + a(i,j)*(y(j)-y(i))*(t(:,j)-t(:,i));
    end

    ytilde((i-1)*s+1:i*s,1) = yv;

    for j=1:m
        Ktilde((i-1)*s+1:i*s,(j-1)*s+1:j*s) = K(i,j)*Bmat;
    end
 end


 % solves the linear system for coefficients c
 I = eye(m*s);
 c = (mˆ2*lambda*I+Ktilde)\ytilde;

 % uwraps the coefficients into a vector for each sample
 Cmat = zeros(dim,m);
 for i = 1:m
    vec=zeros(dim,1);
     for j=1:s
         vec = vec+(c((i-1)*s+j,1)/sqrt(dp(j,1)))*M*U(:,j);
     end
     Cmat(:,i) = vec;
 end

 % computes the gradient for each sample
 F = zeros(dim,m);
 F = Cmat*K;

 %computes the norm for each dimension
 nrm = zeros(dim,1);
 for i=1:dim
   nrm(i) = Cmat(i,:)*K*transpose(Cmat(i,:));
 end
```

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 686:337–404, 1950.

M. Belkin and P. Niyogi. Semi-Supervised Learning on Riemannian Manifolds. *Machine Learning*, 56(1-3):209–239, 2004.

I. Carrel, A. Cottle, K. Coglin, and H. Willard. A first-generation X-incativation profile of the human X chromosome. *Proc. Natl. Acad. Sci. USA*, 96:14440–14444, 1999.

O. Chapelle, V. N. Vapnik, O. Bousquet, and S. Mukherjee. Choosing Multiple Parameters for Support Vector Machines. *Machine Learning*, 46(1-3):131–159, 2002.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.

C. Cortes and V. N. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39: 1–49, 2001.

C. Disteche, G. Flippova, and K. Tsuchiya. Escape from X inactivation. *Cytogenet. Genome Res.*, 99:35–43, 2002.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization Networks and Support Vector Machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines: theory and applications to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Society*, 99:67–81, 2004.

M. Liao. *Bayesian estimation of gene expression index and Bayesian kernel models*. PhD thesis, Duke University, Durham, NC, 2005.

M. Liao, F. Liang, S. Mukherjee, and M. West. Bayesian kernel regression and radial basis function models. Preprint, 2005.

C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17: 177–204, 2005.

I. Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Probab.*, 22:1679–1706, 1994.

I. Pinelis. Correction: ”Optimum bounds for the distributions of martingales in Banach spaces”. *Ann. Probab.*, 27:2119, 1999.

T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990.

B. Schoelkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

D. K. Slonim, P. Tamayo, J. P. Mesirov, T. R. Golub, and E. S. Lander. Class prediction and discovery using gene expression data. In *Proc. of the 4th Annual International Conference on Computational Molecular Biology (RECOMB)*, pages 263–272, 2000.

S. Smale and D. X. Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 24, 2006a.

S. Smale and D. X. Zhou. Shannon sampling II. Connections to learning theory. *Appl. Comput. Harmonic Anal.*, 19:285–302, 2006b.

S. Smale and D. X. Zhou. Shannon sampling and function reconstruction from point values. *Bull. Amer. Math. Soc.*, 41:279–305, 2004.

S. Smale and D. X. Zhou. Estimating the approximation error in learning theory. *Anal. Appl.*, 1: 17–41, 2003.

A. Subramanian, P. Tamayo, VK. Mootha, S. Mukherjee, BL. Ebert, MA. Gillette, A. Paulovich, SL. Pomeroy, TR. Golub, ES. Lander, and JP. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 2005.

A. Sweet-Cordero, S. Mukherjee, A. Subramanian, H. You, J. J. Roix, C. Ladd-Acosta, J. P. Mesirov, T. R. Golub, and T. Jacks. An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nature Genetics*, 37:48–55, 2005.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J Royal Stat Soc B*, 58(1):267–288, 1996.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning. *Foundat. Comput. Math.*, 5:59–85, 2005.

G. Wahba and J. Wendelberger. Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Rev.*, 108:1122–1145, 1980.

M. West. Bayesian factor regression models in the “large p, small n” paradigm. In J. M. Bernardo et al., editor, *Bayesian Statistics 7*, pages 723–732. Oxford, 2003.

Q. Wu and D. X. Zhou. Support vector machine classifiers: linear programming versus quadratic programming. *Neural Computation*, 17:1160–1187, 2005.

T. Zhang. Leave-one-out bounds for kernel methods. *Neural Computation*, 15(6):1397–1437, 2003.

D. X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Trans. Inform. Theory*, 49:1743–1752, 2003.