

Learning Coordinate Covariances via Gradients

Sayan Mukherjee

*Institute for Genome Sciences and Policy
Institute of Statistics and Decision Sciences
Duke University
Durham, NC 27708, USA*

SAYAN@STAT.DUKE.EDU

Ding-Xuan Zhou

*Department of Mathematics
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong, China*

MAZHOU@CITYU.EDU.HK

Editor: Leslie Pack Kaelbling

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 datasets. For square loss we provide a very efficient implementation with respect to both memory and time.

Keywords: Tikhonov regularization, Variable selection, Reproducing Kernel Hilbert Space, Generalization bounds

1. Introduction

The advent of datasets 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 modeling 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 the same 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 $X \rightarrow Y \subset \mathbb{R}$. If we assign a penalty functional $\Omega : \mathcal{H} \rightarrow \mathbb{R}_+$ on \mathcal{H} and choose a loss function $V : \mathbb{R}^2 \rightarrow \mathbb{R}_+$, the Tikhonov regularization scheme in \mathcal{H} associated with (V, Ω) is defined for a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in (X \times Y)^m$ and $\lambda > 0$ as

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

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 $K : X \times X \rightarrow \mathbb{R}$ be continuous, symmetric and positive semidefinite, i.e., for any finite set of distinct points $\{x_1, \dots, x_\ell\} \subset X$, the matrix $(K(x_i, x_j))_{i,j=1}^\ell$ 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 closure 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

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

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 $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ 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

$$f_\rho(x) = \int_Y y d\rho(y|x), \quad x \in X \quad (3)$$

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, 2005b)) 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 $\lambda = \lambda(m) \rightarrow 0$ as $m \rightarrow \infty$.

For (binary) classification, we take $Y = \{-1, 1\}$. A real valued function $f : X \rightarrow \mathbb{R}$ induces a classifier $\text{sgn}(f) : X \rightarrow Y$. In this case, one uses a (convex) loss function $\phi : \mathbb{R} \rightarrow$

\mathbb{R}_+ to measure the empirical error $\phi(t)$, $t = yf(x)$, when $\text{sgn}(f(x))$ is applied to predict $y \in Y$. Examples of such a convex loss function ϕ include the logistic loss

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

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 $\text{sgn}(f_{\mathbf{z}}^V)$ converges to the Bayes rule $\text{sgn}(f_{\rho})$ with respect to the misclassification error $\mathcal{R}(\text{sgn}(f)) = \text{Prob}\{\text{sgn}(f(x)) \neq y\}$.

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 $x = (x^1, x^2, \dots, x^n)^T \in \mathbb{R}^n$. The gradient of f_{ρ} is the vector of functions (if the partial derivatives exist)

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

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 $f(u)$ around the point x gives us

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

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

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

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

$f(u)$ and its expansion $f(x) + \int_{\Delta x \in \Gamma_x} \langle \nabla f, \Delta x \rangle \approx f(x) + \nabla f(x) \cdot (u - x)$ 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 $\vec{f} = (f_1, f_2, \dots, f_n)^T$ with $f_\ell \in \mathcal{H}_K$. Define $\langle \vec{f}, \vec{g} \rangle_K = \sum_{\ell=1}^n \langle f_\ell, g_\ell \rangle_K$. 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

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

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 $w_{i,j}^{(s)} \rightarrow 0$ as $|x_i - x_j|/s \rightarrow \infty$.

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

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

For regression we define the following algorithm 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 algorithm is defined for the sample $\mathbf{z} \in Z^m$ as*

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

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

A similar algorithm can be defined for classification with a convex loss function like the hinge or logistic loss.

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

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

Remark 3 *Algorithms for numerical derivatives 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 to compute 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 high dimensional space.

1.3 Overview

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

The regularization parameters in (7) depend on m : $\lambda = \lambda(m)$, $s = s(m)$ and usually $\lambda(m), s(m) \rightarrow 0$ as m becomes large. In Sections 4 and 5, we shall present an error analysis for a special choice (6) of the weights. It shows how the choice of these two regularization parameters lead to rates of the convergence of $\vec{f}_{\mathbf{z},\lambda}$ to ∇f_ρ .

The utility of the algorithm is demonstrated in Section 6 in applications to simulated data as well as gene expression data. We close with a brief discussion in Section 7.

2. Representer Theorem

The algorithm (7) can be solved by a linear system. Denote $\mathbb{R}^{p \times q}$ as the space of $p \times q$ matrices, I_n the $n \times n$ identity matrix, and $\text{diag}\{B_1, B_2, \dots, B_m\}$ the $m \times m$ block diagonal matrix with each $B_i \in \mathbb{R}^{n \times n}$. To save space, we use $c = (c_1^T, c_2^T, \dots, c_m^T)^T$ to express an mn column vector with $c_i \in \mathbb{R}^n$.

Theorem 4 For $i = 1, \dots, m$, let B_i

$$B_i = \sum_{j=1}^m w_{i,j} (x_j - x_i)(x_j - x_i)^T \in \mathbb{R}^{n \times n}, \quad Y_i = \sum_{j=1}^m w_{i,j} (y_j - y_i)(x_j - x_i) \in \mathbb{R}^n. \quad (9)$$

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

$$\left\{ m^2 \lambda I_{mn} + \text{diag}\{B_1, B_2, \dots, B_m\} [K(x_i, x_j) I_n]_{i,j=1}^m \right\} c = (Y_1^T, Y_2^T, \dots, Y_m^T)^T. \quad (10)$$

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 $x \cdot y = \sum_{i=1}^n x^i y^i = x^T y$ for $x, y \in \mathbb{R}^n$. 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

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

For $q \in \{1, \dots, m\}$, $k \in \{1, \dots, n\}$,

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

Then we know that $c_{\mathbf{z}} = \{c_{i,\mathbf{z}}\}_{i=1}^m$ is the same as the solution to the linear system

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

Since $(x_j - x_i)^T c_p$ is 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

$$B_i \sum_{p=1}^m K(x_i, x_p) c_p + m^2 \lambda c_i = Y_i, \quad i = 1, \dots, m. \quad (11)$$

This is exactly the system in (10). ■

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 (10) for implementing the learning algorithm (7) is not satisfactory, since the size of the linear system (10) is $(mn) \times (mn)$.

Observe that each term in the summation defining B_i in (9) 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 (10). In this section, we show how to reduce this size to $(dm) \times (dm)$ with $d \leq m - 1$. Denote $V_{\mathbf{x}} = \text{span}\{x_j - x_m\}_{j=1}^{m-1}$, the subspace of \mathbb{R}^n generated by the vectors $\{x_j - x_m\}$.

Theorem 5 *Let an $n \times d$ matrix $V = (V_1, \dots, V_d)$ have linearly independent column vectors and its column space $\text{span}\{V_\ell\}_{\ell=1}^d$ contains $V_{\mathbf{x}}$. Write $x_j - x_m = \sum_{\ell=1}^d \tilde{x}_j^\ell V_\ell = V \tilde{x}_j$ with $\tilde{x}_j \in \mathbb{R}^d$ for each j . Then $\vec{f}_{\mathbf{z},\lambda} = \sum_{i=1}^m \left\{ \sum_{\ell=1}^d \tilde{c}_{i,\mathbf{z}}^\ell V_\ell \right\} K_{x_i}$ with $\tilde{c}_{\mathbf{z}} = (\tilde{c}_{1,\mathbf{z}}^T, \dots, \tilde{c}_{m,\mathbf{z}}^T)^T \in \mathbb{R}^{md}$ satisfying*

$$\left\{ m^2 \lambda I_{md} + \text{diag}\{\tilde{B}_1, \tilde{B}_2, \dots, \tilde{B}_m\} [K(x_i, x_j) I_d]_{i,j=1}^m \right\} \tilde{c} = (\tilde{Y}_1^T, \tilde{Y}_2^T, \dots, \tilde{Y}_m^T)^T. \quad (12)$$

where $\tilde{B}_i = \sum_{j=1}^m w_{i,j} (\tilde{x}_j - \tilde{x}_i) (x_j - x_i)^T V \in \mathbb{R}^{d \times d}$, $\tilde{Y}_i = \sum_{j=1}^m w_{i,j} (y_j - y_i) (\tilde{x}_j - \tilde{x}_i) \in \mathbb{R}^d$.

Proof Since $B_i K(x_i, x_p) c_p \in V_{\mathbf{x}}$ and $Y_i \in V_{\mathbf{x}}$, we see for $c_{\mathbf{z}}$ satisfied by equation (11) $c_{i,\mathbf{z}} \in V_{\mathbf{x}}$ for each i . Write $c_{i,\mathbf{z}} = \sum_{\ell=1}^d \tilde{c}_i^\ell V_\ell$ with $\tilde{c}_i = (\tilde{c}_i^1, \dots, \tilde{c}_i^d)^T \in \mathbb{R}^d$. Substituting this

and the following expression for $x_j - x_i = (x_j - x_m) - (x_i - x_m)$ into equation (11) results in the following linear system

$$\begin{aligned} & \sum_{p=1}^m \sum_{j=1}^m w_{i,j} \sum_{\ell=1}^d (\tilde{x}_j^\ell - \tilde{x}_i^\ell) V_\ell K(x_i, x_p) \sum_{q=1}^d \tilde{c}_p^q (x_j - x_i)^T V_q + m^2 \lambda \sum_{\ell=1}^d \tilde{c}_i^\ell V_\ell \\ &= \sum_{j=1}^m w_{i,j} (y_j - y_i) \sum_{\ell=1}^d (\tilde{x}_j^\ell - \tilde{x}_i^\ell) V_\ell, \quad i = 1, \dots, m. \end{aligned}$$

Note that $\sum_{q=1}^d \tilde{c}_p^q (x_j - x_i)^T V_q = (x_j - x_i)^T (V_1, \dots, V_d) \tilde{c}_p = (x_j - x_i)^T V \tilde{c}_p$. By the linear independence of $\{V_\ell\}$, the above equation is equivalent to

$$\sum_{j=1}^m w_{i,j} (\tilde{x}_j^\ell - \tilde{x}_i^\ell) (x_j - x_i)^T V \sum_{p=1}^m K(x_i, x_p) \tilde{c}_p + m^2 \lambda \tilde{c}_i^\ell = \sum_{j=1}^m w_{i,j} (y_j - y_i) (\tilde{x}_j^\ell - \tilde{x}_i^\ell)$$

with $\ell = 1, \dots, d, i = 1, \dots, m$. That is,

$$\sum_{j=1}^m w_{i,j} (\tilde{x}_j - \tilde{x}_i) (x_j - x_i)^T V \sum_{p=1}^m K(x_i, x_p) \tilde{c}_p + m^2 \lambda \tilde{c}_i = \sum_{j=1}^m w_{i,j} (y_j - y_i) (\tilde{x}_j - \tilde{x}_i)$$

with $i = 1, \dots, m$. Observe that $\sum_{p=1}^m K(x_i, x_p) \tilde{c}_p$ is the i -th block of $[K(x_i, x_j) I_d]_{i,j=1}^m \tilde{c}$. Our conclusion follows. \blacksquare

We may take d to be the dimension of $V_{\mathbf{x}}$. In this case, the coefficient matrix of the linear system is $(dm) \times (dm)$ with $d \leq m - 1$.

The expression for $\vec{f}_{\mathbf{z}, \lambda}$ given by (10) is a special form of that derived from (12): take $V = I_n$, then $\{V_1, \dots, V_n\}$ form a canonical basis of \mathbb{R}^n and $\tilde{x}_j = x_j - x_m$, $\tilde{x}_j - \tilde{x}_i = x_j - x_i$, hence $\tilde{B}_i = B_i$, $\tilde{Y}_i = Y_i$, and (12) becomes (10). Thus, Theorem 2 is more general than Theorem 1. However, we keep Theorem 1 because it yields simple tools for the error analysis in the next section.

Denote $M_{\mathbf{x}} = (x_1 - x_m, x_2 - x_m, \dots, x_{m-1} - x_m) \in \mathbb{R}^{n \times (m-1)}$. Its column space is exactly $V_{\mathbf{x}}$ used in Theorem 5. One natural basis is from a singular value decomposition and is formed by the eigenvectors of the $(m-1) \times (m-1)$ symmetric positive semidefinite matrix $M_{\mathbf{x}}^T M_{\mathbf{x}}$ associated with positive eigenvalues. We consider this special choice of the basis for Theorem 5 to reduce the matrix size further and develop an efficient approximation algorithm based on an eigenvalue truncation.

Corollary 6 *Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d > 0$ be all the positive eigenvalues of $M_{\mathbf{x}}^T M_{\mathbf{x}}$ and $\{U_j\}_{j=1}^d$ be corresponding eigenvectors forming an orthonormal system in \mathbb{R}^{m-1} . Then $\{V_j = \frac{1}{\sqrt{\lambda_j}} M_{\mathbf{x}} U_j\}_{j=1}^d$ is an orthonormal basis of $V_{\mathbf{x}}$ and $\vec{f}_{\mathbf{z}, \lambda} = \sum_{i=1}^m \{\sum_{\ell=1}^d \tilde{c}_{i, \mathbf{z}}^\ell V_\ell\} K_{x_i}$ where $\tilde{c}_{\mathbf{z}}$ satisfies (12) with $\tilde{B}_i = \sum_{j=1}^m w_{i,j} (\sqrt{\lambda_1} (U_1^j - U_1^i), \dots, \sqrt{\lambda_d} (U_d^j - U_d^i))^T (\sqrt{\lambda_1} (U_1^j - U_1^i), \dots, \sqrt{\lambda_d} (U_d^j - U_d^i))$ and $\tilde{Y}_i = \sum_{j=1}^m w_{i,j} (y_j - y_i) (\sqrt{\lambda_1} (U_1^j - U_1^i), \dots, \sqrt{\lambda_d} (U_d^j - U_d^i))$. Here $U_\ell^m = 0$ for $1 \leq \ell \leq d$.*

Proof We apply a singular value decomposition. Let $\lambda_{d+1} = \dots = \lambda_{m-1} = 0$ and $\{U_j\}_{j=d+1}^{m-1}$ be an orthonormal basis of the null space of $M_{\mathbf{x}}^T M_{\mathbf{x}}$. If we denote $V_{d+1} = \dots = V_{m-1} = 0$, then $M_{\mathbf{x}} U_j = \sqrt{\lambda_j} V_j$ for each $j \in \{1, \dots, m-1\}$ and

$$M_{\mathbf{x}}(U_1, U_2, \dots, U_{m-1}) = (\sqrt{\lambda_1} V_1, \sqrt{\lambda_2} V_2, \dots, \sqrt{\lambda_d} V_d, 0, \dots, 0).$$

Since the matrix $U = (U_1, U_2, \dots, U_{m-1})$ is orthogonal, we have $U^{-1} = U^T$ and thereby

$$M_{\mathbf{x}} = (\sqrt{\lambda_1} V_1, \sqrt{\lambda_2} V_2, \dots, \sqrt{\lambda_d} V_d, 0, \dots, 0) U^T$$

is the same as

$$(\sqrt{\lambda_1} V_1, \sqrt{\lambda_2} V_2, \dots, \sqrt{\lambda_d} V_d) (U_1^T, U_2^T, \dots, U_d^T)^T.$$

In particular,

$$x_j - x_m = (\sqrt{\lambda_1} V_1, \sqrt{\lambda_2} V_2, \dots, \sqrt{\lambda_d} V_d) (U_1^j, U_2^j, \dots, U_d^j)^T.$$

So the vector $x_j - x_m$ can be expressed as

$$x_j - x_m = \sum_{\ell=1}^d \sqrt{\lambda_\ell} U_\ell^j V_\ell = V (\sqrt{\lambda_1} U_1^j, \sqrt{\lambda_2} U_2^j, \dots, \sqrt{\lambda_d} U_d^j)^T, \quad 1 \leq j \leq m-1 \quad (13)$$

and $\{V_j\}_{j=1}^d$ spans $V_{\mathbf{x}}$. Moreover, $V_i \cdot V_j = V_i^T V_j = \frac{1}{\sqrt{\lambda_i \lambda_j}} U_i^T M_{\mathbf{x}}^T M_{\mathbf{x}} U_j = \sqrt{\frac{\lambda_j}{\lambda_i}} U_i^T U_j = \delta_{i,j}$.

This means $\{V_j\}_{j=1}^d$ is an orthonormal basis of $V_{\mathbf{x}}$, and $V^T V = I_d$. Set $U_\ell^m = 0$ for $1 \leq \ell \leq d$, then (13) also holds for $j = m$.

By (13), we know that $\tilde{x}_j = (\sqrt{\lambda_1} U_1^j, \sqrt{\lambda_2} U_2^j, \dots, \sqrt{\lambda_d} U_d^j)^T$. It follows that (12) holds and the expression for \tilde{B}_i is seen from the fact that $(x_j - x_i)^T V = (V(\tilde{x}_j - \tilde{x}_i))^T V = (\tilde{x}_j - \tilde{x}_i)^T$. \blacksquare

The representation given in Corollary 6 enables us to reduce the matrix size by solving an approximation to the linear system derived from the eigenvalues. A strong correlation among the vectors $\{x_i\}$ would result in a large number of small eigenvalues. If we ignore the small eigenvalues $\lambda_{S+1}, \dots, \lambda_d$, the error is proportional to $\sqrt{\lambda_{S+1}}$, as shown in the following.

Corollary 7 Assume $|y| \leq M$ almost surely. Denote $\kappa = \sup_{x \in X} \sqrt{K(x, x)}$. In the setting of

Corollary 6, let $1 \leq S < d$, $t_j = (\sqrt{\lambda_1} U_1^j, \dots, \sqrt{\lambda_S} U_S^j)^T \in \mathbb{R}^S$, $\mathcal{B}_i = \sum_{j=1}^m w_{i,j} (t_j - t_i)(t_j - t_i)^T$ and $\mathcal{Y}_i = \sum_{j=1}^m w_{i,j} (y_j - y_i)(t_j - t_i)$. Solve the linear system

$$\left\{ m^2 \lambda I_{mS} + \text{diag}\{\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_m\} [K(x_i, x_j) I_S]_{i,j=1}^m \right\} \hat{b} = (\mathcal{Y}_1^T, \mathcal{Y}_2^T, \dots, \mathcal{Y}_m^T)^T. \quad (14)$$

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

$$\|b_{\mathbf{z}} - c_{\mathbf{z}}\|_{\ell^2(\mathbb{R}^{mS})} \leq \frac{4M\sqrt{\lambda_{S+1}}}{m^2\lambda} \left\{ \sqrt{d\Delta_m} + \frac{4\kappa^2\sqrt{S}\lambda_1}{m\lambda} \Delta_m \right\}, \quad (15)$$

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

Proof Expand $t_j \in \mathbb{R}^{\mathcal{S}}$ to a vector $\tilde{t}_j = (\sqrt{\lambda_1}U_1^j, \dots, \sqrt{\lambda_{\mathcal{S}}}U_{\mathcal{S}}^j, 0, \dots, 0)^T \in \mathbb{R}^d$. Denote $\tilde{\mathcal{Y}}_i = \sum_{j=1}^m w_{i,j}(y_j - y_i)(\tilde{t}_j - \tilde{t}_i) \in \mathbb{R}^d$ and let $\tilde{\mathbf{c}}_{\mathbf{z}} = (\tilde{c}_{1,\mathbf{z}}^T, \dots, \tilde{c}_{m,\mathbf{z}}^T)^T \in \mathbb{R}^{md}$ be the solution of the system

$$\left\{ m^2 \lambda I_{md} + \text{diag}\{\tilde{B}_1, \tilde{B}_2, \dots, \tilde{B}_m\} [K(x_i, x_j) I_d]_{i,j=1}^m \right\} \tilde{\mathbf{c}} = (\tilde{\mathcal{Y}}_1^T, \tilde{\mathcal{Y}}_2^T, \dots, \tilde{\mathcal{Y}}_m^T)^T.$$

Then by comparing $\{\tilde{\mathcal{Y}}_i\}$ with $\{\tilde{Y}_i\}$ defined in Theorem 5, we have

$$|\tilde{\mathbf{c}}_{\mathbf{z}}' - \tilde{\mathbf{c}}_{\mathbf{z}}|_{\ell^2(\mathbb{R}^{md})} \leq \frac{1}{m^2 \lambda} \left\{ \sum_{i=1}^m |\tilde{\mathcal{Y}}_i - \tilde{Y}_i|_{\ell^2(\mathbb{R}^d)}^2 \right\}^{1/2}.$$

For each i , we have

$$\begin{aligned} |\tilde{\mathcal{Y}}_i - \tilde{Y}_i|_{\ell^2(\mathbb{R}^d)} &= \left| \sum_{j=1}^m w_{i,j}(y_j - y_i) \left(0, \dots, 0, \sqrt{\lambda_{\mathcal{S}+1}}(U_{\mathcal{S}+1}^j - U_{\mathcal{S}+1}^i), \dots, \sqrt{\lambda_d}(U_d^j - U_d^i) \right)^T \right|_{\ell^2(\mathbb{R}^d)} \\ &\leq \sum_{j=1}^m w_{i,j} 2M \sqrt{\lambda_{\mathcal{S}+1}} \left\{ \left(\sum_{\ell=\mathcal{S}+1}^d (U_{\ell}^j)^2 \right)^{1/2} + \left(\sum_{\ell=\mathcal{S}+1}^d (U_{\ell}^i)^2 \right)^{1/2} \right\}. \end{aligned}$$

Since the matrix U is orthogonal, we know that $\sum_{\ell=1}^{m-1} (U_{\ell}^j)^2 = 1$ for each $j \in \{1, \dots, m-1\}$ and $\sum_{j=1}^{m-1} (U_{\ell}^j)^2 = 1$ for each $\ell \in \{1, \dots, d\}$. Hence we see by the Schwartz inequality that

$$|\tilde{\mathbf{c}}_{\mathbf{z}}' - \tilde{\mathbf{c}}_{\mathbf{z}}|_{\ell^2(\mathbb{R}^{md})} \leq \frac{2M \sqrt{\lambda_{\mathcal{S}+1}} \sqrt{d - \mathcal{S}}}{m^2 \lambda} \left\{ \left(\sum_{i,j=1}^m (w_{i,j})^2 \right)^{1/2} + \max_{1 \leq i \leq m} \sum_{j=1}^m w_{i,j} \right\}.$$

Let $\tilde{\mathbf{B}}_i = \sum_{j=1}^m w_{i,j}(\tilde{t}_j - \tilde{t}_i)(\tilde{t}_j - \tilde{t}_i)^T$. Then for any $b \in \mathbb{R}^d$,

$$\begin{aligned} (\tilde{B}_i - \tilde{\mathbf{B}}_i)b &= \sum_{j=1}^m w_{i,j} \left(\sqrt{\lambda_1}(U_1^j - U_1^i), \dots, \sqrt{\lambda_{\mathcal{S}}}(U_{\mathcal{S}}^j - U_{\mathcal{S}}^i), 0, \dots, 0 \right)^T \sum_{\ell=\mathcal{S}+1}^d \sqrt{\lambda_{\ell}}(U_{\ell}^j - U_{\ell}^i)b^{\ell} \\ &\quad + \sum_{j=1}^m w_{i,j} \left(0, \dots, 0, \sqrt{\lambda_{\mathcal{S}+1}}(U_{\mathcal{S}+1}^j - U_{\mathcal{S}+1}^i), \dots, \sqrt{\lambda_d}(U_d^j - U_d^i) \right)^T \sum_{\ell=1}^d \sqrt{\lambda_{\ell}}(U_{\ell}^j - U_{\ell}^i)b^{\ell}. \end{aligned}$$

The Schwartz inequality tells us that

$$\left| \sum_{\ell=\mathcal{S}+1}^d \sqrt{\lambda_{\ell}}(U_{\ell}^j - U_{\ell}^i)b^{\ell} \right| \leq \sqrt{\lambda_{\mathcal{S}+1}} |b|_{\ell^2(\mathbb{R}^d)} \left\{ \sum_{\ell=\mathcal{S}+1}^d (U_{\ell}^j - U_{\ell}^i)^2 \right\}^{1/2}$$

and

$$\sum_{\ell=1}^d \sqrt{\lambda_{\ell}}(U_{\ell}^j - U_{\ell}^i)b^{\ell} \leq |b|_{\ell^2(\mathbb{R}^d)} \left\{ \sum_{\ell=1}^d \lambda_{\ell} (U_{\ell}^j - U_{\ell}^i)^2 \right\}^{1/2}.$$

Hence

$$\begin{aligned} \|(\tilde{B}_i - \tilde{\mathcal{B}}_i)b\|_{\ell^2(\mathbb{R}^d)} &\leq 2 \sum_{j=1}^m w_{i,j} \sqrt{\lambda_{S+1}} |b|_{\ell^2(\mathbb{R}^d)} \left\{ \sum_{\ell=S+1}^d (U_\ell^j - U_\ell^i)^2 \right\}^{1/2} \left\{ \sum_{\ell=1}^d \lambda_\ell (U_\ell^j - U_\ell^i)^2 \right\}^{1/2} \\ &\leq 8\sqrt{\lambda_1} \sqrt{\lambda_{S+1}} |b|_{\ell^2(\mathbb{R}^d)} \sum_{j=1}^m w_{i,j}. \end{aligned}$$

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

$$\left\| \text{diag}\{\tilde{B}_1, \tilde{B}_2, \dots, \tilde{B}_m\} - \text{diag}\{\tilde{\mathcal{B}}_1, \tilde{\mathcal{B}}_2, \dots, \tilde{\mathcal{B}}_m\} \right\| \leq 8\sqrt{\lambda_1} \sqrt{\lambda_{S+1}} \max_{1 \leq i \leq m} \sum_{j=1}^m w_{i,j}.$$

It follows that

$$\begin{aligned} &\left\| \text{diag}\{\tilde{B}_1, \tilde{B}_2, \dots, \tilde{B}_m\} [K(x_i, x_j) I_d]_{i,j=1}^m - \text{diag}\{\tilde{\mathcal{B}}_1, \tilde{\mathcal{B}}_2, \dots, \tilde{\mathcal{B}}_m\} [K(x_i, x_j) I_d]_{i,j=1}^m \right\| \\ &\leq 8\kappa^2 m \sqrt{\lambda_1} \sqrt{\lambda_{S+1}} \max_{1 \leq i \leq m} \sum_{j=1}^m w_{i,j}. \end{aligned}$$

Expand the vectors $\hat{b}_{i,\mathbf{z}}$ to those on $\tilde{b}_{i,\mathbf{z}} \in \mathbb{R}^d$ by adding zero entries. Then

$$\vec{f}_{\mathbf{z},\lambda,S} = \sum_{i=1}^m \left\{ \sum_{\ell=1}^d \tilde{b}_{i,\mathbf{z}}^\ell V_\ell \right\} K_{x_i}$$

and the vector $\tilde{b}_{\mathbf{z}} = (\tilde{b}_{1,\mathbf{z}}^T, \dots, \tilde{b}_{m,\mathbf{z}}^T)^T \in \mathbb{R}^{md}$ is given by

$$\tilde{b}_{\mathbf{z}} = \left\{ m^2 \lambda I_{md} + \text{diag}\{\tilde{\mathcal{B}}_1, \tilde{\mathcal{B}}_2, \dots, \tilde{\mathcal{B}}_m\} [K(x_i, x_j) I_d]_{i,j=1}^m \right\}^{-1} (\tilde{\mathcal{Y}}_1^T, \tilde{\mathcal{Y}}_2^T, \dots, \tilde{\mathcal{Y}}_m^T)^T.$$

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

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

Applying this to our setting, we have

$$\|\tilde{b}_{\mathbf{z}} - \tilde{\mathcal{C}}_{\mathbf{z}}\|_{\ell^2(\mathbb{R}^{md})} \leq \left(\frac{1}{m^2 \lambda} \right)^2 8\kappa^2 m \sqrt{\lambda_1} \sqrt{\lambda_{S+1}} \max_{1 \leq i \leq m} \sum_j w_{i,j} \|(\tilde{\mathcal{Y}}_1^T, \tilde{\mathcal{Y}}_2^T, \dots, \tilde{\mathcal{Y}}_m^T)^T\|_{\ell^2(\mathbb{R}^{md})}.$$

For each i , we have

$$\|\tilde{\mathcal{Y}}_i\|_{\ell^2(\mathbb{R}^d)} \leq 2M \sum_{j=1}^m w_{i,j} \left\{ \left(\sum_{\ell=1}^S \lambda_\ell (U_\ell^j)^2 \right)^{1/2} + \left(\sum_{\ell=1}^S \lambda_\ell (U_\ell^i)^2 \right)^{1/2} \right\}.$$

It follows that

$$|\tilde{b}_{\mathbf{z}} - \tilde{c}_{\mathbf{z}}|_{\ell^2(\mathbb{R}^{md})} \leq \left(\frac{1}{m^2\lambda}\right)^2 16M\kappa^2 m\sqrt{\mathcal{S}}\lambda_1\sqrt{\lambda_{\mathcal{S}+1}} \left\{ \max_{1 \leq i \leq m} \left(\sum_{j=1}^m w_{i,j}\right)^2 + \sum_{i,j=1}^m (w_{i,j})^2 \right\}.$$

Combining this with the bound for $|\tilde{c}_{\mathbf{z}} - \tilde{z}_{\mathbf{z}}|_{\ell^2(\mathbb{R}^{md})}$ we conclude that

$$|b_{\mathbf{z}} - c_{\mathbf{z}}|_{\ell^2(\mathbb{R}^{mn})} = |\tilde{b}_{\mathbf{z}} - \tilde{c}_{\mathbf{z}}|_{\ell^2(\mathbb{R}^{md})} \leq \frac{4M\sqrt{\lambda_{\mathcal{S}+1}}}{m^2\lambda} \left\{ \sqrt{d - \mathcal{S}}\sqrt{\Delta_m} + \frac{4\kappa^2\sqrt{\mathcal{S}}\lambda_1}{m\lambda}\Delta_m \right\}.$$

This proves the corollary. \blacksquare

Corollary 7 provides a theoretical foundation of the following approximation algorithm with reduced matrix size that well approximates algorithm (7).

Algorithm 1: Approximation algorithm with reduced matrix size

input : inputs $(x_i)_{i=1}^m$, labels $(y_i)_{i=1}^m$, kernel K , weights $(w_{i,j})$, eigenvalue threshold $\epsilon > 0$

return: coefficients $(b_{i,\mathbf{z}})_{i=1}^m$

$M_{\mathbf{x}} = (x_1 - x_m, x_2 - x_m, \dots, x_{m-1} - x_m) \in \mathbb{R}^{n \times (m-1)}$;

$V_{\mathbf{x}} = M_{\mathbf{x}}^T M_{\mathbf{x}} \in \mathbb{R}^{(m-1) \times (m-1)}$;

Given $V_{\mathbf{x}}$ compute eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{\mathcal{S}} > \epsilon$ and corresponding eigenvectors $\{U_{\ell} = (U_{\ell}^j)_{j=1}^{m-1}\}_{\ell=1}^{\mathcal{S}} \subset \mathbb{R}^{m-1}$;

$t_j = (\sqrt{\lambda_1}U_1^j, \dots, \sqrt{\lambda_{\mathcal{S}}}U_{\mathcal{S}}^j)^T \in \mathbb{R}^{\mathcal{S}}$ for $1 \leq j \leq m-1$ and $t_m = 0$;

$\mathcal{B}_i = \sum_{j=1}^m w_{i,j}(t_j - t_i)(t_j - t_i)^T$ 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 $1 \leq i \leq m$;

$$[\tilde{K}] = \begin{bmatrix} \mathcal{B}_1 K(x_1, x_1) & \mathcal{B}_1 K(x_1, x_2) & \dots & \mathcal{B}_1 K(x_1, x_m) \\ \mathcal{B}_2 K(x_2, x_1) & \mathcal{B}_2 K(x_2, x_2) & \dots & \mathcal{B}_2 K(x_2, x_m) \\ \vdots & \ddots & \dots & \vdots \\ \mathcal{B}_m K(x_m, x_1) & \mathcal{B}_m K(x_m, x_2) & \dots & \mathcal{B}_m K(x_m, x_m) \end{bmatrix}, \quad \vec{\mathcal{Y}} = \begin{bmatrix} \mathcal{Y}_1 \\ \mathcal{Y}_2 \\ \vdots \\ \mathcal{Y}_m \end{bmatrix}$$

$\hat{b}_{\mathbf{z}} = (\hat{b}_{1,\mathbf{z}}^T, \dots, \hat{b}_{m,\mathbf{z}}^T)^T \in \mathbb{R}^{m\mathcal{S}}$ where

$$\left\{ m^2\lambda I_{m\mathcal{S}} + [\tilde{K}] \right\} \hat{b}_{\mathbf{z}} = \vec{\mathcal{Y}} \quad (16)$$

$\vec{f}_{\mathbf{z},\lambda,\mathcal{S}} = \sum_{i=1}^m \left\{ \sum_{\ell=1}^{\mathcal{S}} \hat{b}_{i,\mathbf{z}}^{\ell} \frac{1}{\sqrt{\lambda_{\ell}}} M_{\mathbf{x}} U_{\ell} \right\} K_{x_i}$ is an approximation of $\vec{f}_{\mathbf{z},\lambda}$;

return $(b_{i,\mathbf{z}})_{i=1}^m$

4. Sample Error

In what follows we use a Gaussian (equation (6)) for the weights and estimate error bounds. We show that $\vec{f}_{\mathbf{z},\lambda} \rightarrow \nabla f_\rho$ as $m \rightarrow \infty, \lambda = \lambda(m) \rightarrow 0, s = s(m) \rightarrow 0$ for suitable choices of the regularization parameters. Because we are learning gradients, some regularity conditions on both the marginal distribution and the density are required. Let us mention a simple case to illustrate the idea (it will be a corollary of the error analysis that follows).

Proposition 8 *Assume $|y| \leq M$ almost surely. Suppose that for some $0 < \tau \leq 2/3, c_\rho > 0$, the marginal distribution ρ_X satisfies*

$$\rho_X(\{x \in X : \inf_{u \notin X} |u - x| \leq s\}) \leq c_\rho^2 s^{4\tau}, \quad \forall s > 0, \quad (17)$$

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

$$\sup_{x \in X} p(x) \leq c_\rho, \quad |p(x) - p(u)| \leq c_\rho |u - x|^\tau, \quad \forall u, x \in X. \quad (18)$$

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

$$\|\vec{f}_{\mathbf{z},\lambda} - \nabla f_\rho\|_\rho \leq C_{\rho,K} \log\left(\frac{2}{\delta}\right) \left(\frac{1}{m}\right)^{\frac{\tau}{2(n+2+3\tau)}}. \quad (19)$$

The condition (18) means the density of the marginal distribution is Hölder τ . The condition (17) 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, (18) implies (17).

First we estimate the sample error by means of the sampling operator introduced in Smale and Zhou (2004, 2005b,a).

Definition 9 *The sampling operator $S_{\mathbf{x}} : \mathcal{H}_K^n \rightarrow \mathbb{R}^{mn}$ associated with a discrete subset $\mathbf{x} = \{x_i\}_{i=1}^m$ of X is defined by*

$$S_{\mathbf{x}}(\vec{f}) = (\vec{f}(x_i))_{i=1}^m.$$

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

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

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

Consider the equation (11) 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 (11) implies $(S_{\mathbf{x}}^T D_{\mathbf{x}} S_{\mathbf{x}} + m^2 \lambda I) \vec{f}_{\mathbf{z},\lambda} = S_{\mathbf{x}}^T \vec{Y}$. Therefore,

$$\vec{f}_{\mathbf{z},\lambda} = \left(\frac{1}{m^2} S_{\mathbf{x}}^T D_{\mathbf{x}} S_{\mathbf{x}} + \lambda I\right)^{-1} \frac{1}{m^2} S_{\mathbf{x}}^T \vec{Y}. \quad (20)$$

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

Definition 10 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

$$\mathcal{E}(\vec{f}) = 2\sigma_s^2 + \int_X \int_X w(x-u) \left[f_\rho(x) - f_\rho(u) + \vec{f}(x) \cdot (u-x) \right]^2 d\rho_X(x) d\rho_X(u). \quad (21)$$

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

$$\vec{f}_\lambda = \arg \min_{\vec{f} \in \mathcal{H}_K^n} \{ \mathcal{E}(\vec{f}) + \lambda \|\vec{f}\|_K^2 \}. \quad (22)$$

Taking the functional derivatives, we know from (21) that it can be expressed in terms of an integral operator on the space $(L_{\rho_X}^2)^n$ with norm $\|\vec{f}\|_\rho = (\|f_\ell\|_\rho^2)^{1/2}$ as follows.

Proposition 11 Let $L_{K,s} : (L_{\rho_X}^2)^n \rightarrow (L_{\rho_X}^2)^n$ be the integral operator defined by

$$L_{K,s} \vec{f} = \int_X \int_X w(x-u) (u-x) K_x(u-x)^T \vec{f}(x) d\rho_X(x) d\rho_X(u). \quad (23)$$

It is a positive operator on $(L_{\rho_X}^2)^n$ and we have

$$\vec{f}_\lambda = (L_{K,s} + \lambda I)^{-1} \vec{f}_{\rho,s}. \quad (24)$$

where

$$\vec{f}_{\rho,s} := \int_X \int_X w(x-u) (u-x) K_x(f_\rho(u) - f_\rho(x)) d\rho_X(x) d\rho_X(u). \quad (25)$$

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 12 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 $F : Z^m \rightarrow H$ 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 $\varepsilon > 0$,

$$\text{Prob}_{\mathbf{z} \in Z^m} \{ \|F(\mathbf{z}) - E_{\mathbf{z}}(F(\mathbf{z}))\| \geq \varepsilon \} \leq 2 \exp \left\{ - \frac{\varepsilon^2}{2(\widetilde{M}\varepsilon + \sigma^2)} \right\}, \quad (26)$$

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) to the finite sequence $\{f_j = E_{z_m, \dots, z_j} F - E_{z_m, \dots, z_1} F : j = 1, 2, \dots, m+1\}$. Then $f_1 = 0$ 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 $\sum_{j=2}^{m+1} E_{z_{j-1}} \|f_j - f_{j-1}\| \leq \sigma^2$. The conditions of Theorem 3.3 of Pinelis (1994) hold with $B^2 = \sigma^2$, $\Gamma = \widetilde{M}$. Observe from the proof that there was a minor mistake in that theorem. The probability there should be $2 \exp\left\{-\frac{r^2}{r\Gamma + B^2 + B\sqrt{B^2 + 2r\Gamma}}\right\}$, which is bounded by $2 \exp\left\{-\frac{r^2}{2(r\Gamma + B^2)}\right\}$. Then (26) follows from that theorem.

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

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

We see that 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. \blacksquare

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 $\text{Diam}(X) = \max_{x, t \in X} |x - t|$. Denote the moments of the Gaussian as

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

Theorem 13 *Assume $|y| \leq M$ almost surely.*

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

$$\|\vec{f}_{\mathbf{z}, \lambda} - \vec{f}_{\lambda}\|_K \leq \frac{16\kappa \text{Diam}(X) \log(2/\delta)}{\sqrt{m\lambda s^{n+2}}} \left\{ 2M + \kappa \text{Diam}(X) \|\vec{f}_{\lambda}\|_K \right\} + \frac{1}{m} \|\vec{f}_{\lambda}\|_K. \quad (27)$$

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*

$$\begin{aligned} \|\vec{f}_{\mathbf{z}, \lambda} - \vec{f}_{\lambda}\|_K &\leq \frac{8\kappa \log(2/\delta)}{\sqrt{m\lambda s^{1+n/2}}} \left(\sqrt{c_\rho} + \frac{\text{Diam}(X)}{\sqrt{m s^{1+n/2}}} \right) \\ &\quad \left(2M \sqrt{M_2} + \kappa (\text{Diam}(X) + \sqrt{M_4}) \|\vec{f}_{\lambda}\|_K \right) + \frac{1}{m} \|\vec{f}_{\lambda}\|_K. \end{aligned} \quad (28)$$

Proof By (20), we have

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

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

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

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 (24), $L_{K,s} \vec{f}_\lambda + \lambda \vec{f}_\lambda = \vec{f}_{\rho,s}$. 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

$$\|f\|_\rho \leq \|f\|_\infty \leq \kappa \|f\|_K, \quad \forall f \in \mathcal{H}_K. \quad (29)$$

Then we apply Proposition 12 to the function $F(\mathbf{z})$ to get our error bound.

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

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

Using (29) for \vec{f}_λ and $|x - x_j| \leq \text{Diam}(X)$ for any $x \in X$, we see that

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

1. Apply the trivial bound $\sigma^2 \leq m\widetilde{M}^2$. Then Proposition 12 tells us that for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

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

This proves (27).

2. To improve the variance σ^2 , we bound $\|F(\mathbf{z}) - E_{z_i}(F(\mathbf{z}))\|_K$ by

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

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

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

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

$$E_{z_i}(\|F(\mathbf{z}) - E_{z_i}(F(\mathbf{z}))\|_K^2) \leq \frac{2(m-1)}{m^2} \left\{ 4\kappa M \sqrt{c_\rho} \sqrt{\frac{M_2}{s^{n+2} 2^{1+n/2}}} + 2\kappa^2 \|\vec{f}_\lambda\|_K \sqrt{c_\rho} \sqrt{\frac{M_4}{s^n 2^{2+n/2}}} \right\}.$$

It follows then that for $s \leq 1$, there holds

$$\sigma^2 \leq \frac{16c_\rho \kappa^2}{ms^{n+2}} \left\{ 2^{1/4} M \sqrt{M_2} + \kappa \|\vec{f}_\lambda\|_K \sqrt{M_4} \right\}^2.$$

Then our second statement follows from Proposition 12. ■

5. Regularization Error

In this section, 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 16 and Theorem 18 with $r = 1/2$.

Proposition 14 *Assume (17) and (18). Suppose that $\nabla f_\rho \in \mathcal{H}_K^n$ and for some $c'_\rho > 0$,*

$$|f_\rho(u) - f_\rho(x) - \nabla f_\rho(x) \cdot (u - x)| \leq c'_\rho |u - x|^2, \quad \forall u, x \in X. \quad (30)$$

Then for any $\lambda > 0$ and $0 < s \leq \min\left\{\left\{2\kappa^2 c_\rho (M_{2+\tau} + M_4 + c_\rho M_2)\right\}^{1/\tau} \lambda^{1/\tau}, 1\right\}$, there holds

$$\|\vec{f}_\lambda - \nabla f_\rho\|_\rho \leq (\kappa^2 c'_\rho M_3) \frac{s}{\lambda} + \left\{ 2(V_\rho n (2\pi)^{n/2})^{-1/2} \|\nabla f_\rho\|_K \right\} \sqrt{\lambda}.$$

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

Lemma 15 *Assume that for some $0 < \tau < 1$, the conditions (17) and (18) hold. Denote $V_\rho = \int_X (p(x))^2 dx > 0$. Then $V_\rho \leq c_\rho$ and for any $0 < s \leq 1$ we have*

$$\|L_{K,s} - V_\rho n(2\pi)^{n/2} L_K\|_{\mathcal{H}_K^n \rightarrow \mathcal{H}_K^n} \leq s^\tau \kappa^2 c_\rho (M_{2+\tau} + M_4 + c_\rho M_2), \quad (31)$$

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

$$L_K \vec{f} = \int_X K_x \vec{f}(x) \frac{p(x)}{V_\rho} d\rho_X(x). \quad (32)$$

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

$$\|L_{K,s} - V_\rho n(2\pi)^{n/2} L_K\|_{(L^2_{\rho_X})^n \rightarrow (L^2_{\rho_X})^n} \leq s^\tau \kappa^2 c_\rho (M_{2+\tau} + M_4 + c_\rho M_2), \quad \forall 0 < s \leq 1. \quad (33)$$

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

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

Observe that $n(2\pi)^{n/2} = M_2$ and $\int_{\mathbb{R}^n} w(u-x)(u^i - x^i)(u^j - x^j) du = 0$ when $i \neq j$. Then $\int_{\mathbb{R}^n} \frac{1}{s^n} e^{-\frac{|u-x|^2}{2s^2}} \left(\frac{u-x}{s} \right) \left(\frac{u-x}{s} \right)^T du = M_2 I_n$. Hence

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

It follows that

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

Separate the domain X into $X_s := \{x \in X : \inf_{u \notin X} |u-x| \leq \sqrt{s}\}$, 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 $1 \leq s \left| \frac{u-x}{s} \right|^2$. Hence

$$\int_{\mathbb{R}^n \setminus X} \frac{1}{s^n} e^{-\frac{|u-x|^2}{2s^2}} \left| \frac{u-x}{s} \right|^2 du \leq s \int_{\mathbb{R}^n \setminus X} \frac{1}{s^n} e^{-\frac{|u-x|^2}{2s^2}} \left| \frac{u-x}{s} \right|^4 du \leq s M_4.$$

It follows from (18) that

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

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

For the subset X_s , we use the Schwartz inequality and (18) to obtain

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

This is bounded by $\kappa c_\rho M_2 \sqrt{\rho_X(X_s)} \|\vec{f}\|_\rho$. By (17), $\rho_X(X_s) \leq c_\rho^2 s^{2\tau}$. Thus, for $0 < s \leq 1$,

$$\|\vec{g} - n(2\pi)^{n/2} L_K \vec{f}\|_K \leq s^\tau \kappa c_\rho (M_4 + c_\rho M_2) \|\vec{f}\|_\rho.$$

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

$$\|L_{K,s} \vec{f} - V_\rho n(2\pi)^{n/2} L_K \vec{f}\|_K \leq s^\tau \kappa c_\rho (M_{2+\tau} + M_4 + c_\rho M_2) \|\vec{f}\|_\rho$$

which proves (33) by (29). When $\vec{f} \in \mathcal{H}_K^n$, we have from (29) again that

$$\|L_{K,s} \vec{f} - V_\rho n(2\pi)^{n/2} L_K \vec{f}\|_K \leq s^\tau \kappa^2 c_\rho (M_{2+\tau} + M_4 + c_\rho M_2) \|\vec{f}\|_K.$$

This verifies (31) and proves the lemma. \blacksquare

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 $(L_{\rho_X}^2)^n$ 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 $\vec{f} \in (L_{\rho_X}^2)^n$.

Theorem 16 *Under the assumption (30), we have*

$$\|\vec{f}_\lambda - \nabla f_\rho + \lambda(L_{K,s} + \lambda I)^{-1} \nabla f_\rho\|_K \leq \frac{s}{\lambda} \kappa c'_\rho M_3.$$

Proof By (24), we find that

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

Then

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

which is bounded by $\frac{1}{\lambda} \|\vec{f}_{\rho,s} - L_{K,s} \nabla f_\rho\|_K$. Using (30) 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

$$\|\vec{f}_{\rho,s} - L_{K,s} \nabla f_\rho\|_K \leq \int_X \int_X w(x-u) |u-x| \|K_x\|_K c'_\rho |u-x|^2 d\rho_X(x) d\rho_X(u) \leq s \kappa c'_\rho M_3.$$

This proves the theorem. \blacksquare

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

Lemma 17 *Assume (17) and (18). Denote $c''_\rho = (2\kappa^2 c_\rho (M_{2+\tau} + M_4 + c_\rho M_2))^{1/\tau}$. Then*

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

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

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

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

This in connection with Lemma 15 implies

$$\|(L_{K,s} + \lambda I)^{-1} \vec{f}\| \leq \left\{ 1 - \frac{1}{\lambda} s^\tau \kappa^2 c_\rho (M_{2+\tau} + M_4 + c_\rho M_2) \right\}^{-1} \|[V_\rho n (2\pi)^{n/2} L_K + \lambda I]^{-1} \vec{f}\|.$$

This verifies the lemma. \blacksquare

Lemma 17 yields the convergence of $\|\lambda(L_{K,s} + \lambda I)^{-1} \nabla f_\rho\|$. The convergence rates require some conditions on ∇f_ρ relative to the pair $(L_{\rho_X}^2, \mathcal{H}_K)$. 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 (2005a).

Theorem 18 *Assume (17), (18), and (30). Let $0 < s \leq \min\{c''_\rho \lambda^{1/\tau}, 1\}$. If $\|L_K^{-r} \nabla f_\rho\|_\rho < \infty$ for some $0 < r \leq 1$, then*

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

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

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

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

Proposition 19 *Assume (17), (18), and (30). Then*

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

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

$$\mathcal{K}(\vec{f}, t) = \inf_{\vec{g} \in \mathcal{H}_K^n} \left\{ \|\vec{f} - \vec{g}\|_\rho + t \|\vec{g}\|_K \right\}, \quad t > 0. \quad (34)$$

Let us prove Proposition 8 to show how our error analysis above can be applied.

Proof of Proposition 8. 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 (30) is satisfied for some constant $c'_\rho > 0$.

Since $\lambda = (1/m)^\gamma$ with $\gamma = \frac{\tau}{n+2+3\tau}$ and $s = (\kappa c_\rho)^{2/\tau} \lambda^{1/\tau}$, we see from the fact $M_2 > 1$ that for $m \geq (\kappa c_\rho)^{2(n+2+3\tau)/\tau}$, the restriction $0 < s \leq \min\{2\kappa^2 c_\rho (M_{2+\tau} + M_4 + c_\rho M_2)\}^{1/\tau} \lambda^{1/\tau}, 1\}$ in Proposition 14 and Lemma 17 is satisfied. Then by Proposition 14, since $\frac{1}{\tau} - 1 \geq \frac{1}{2}$, we have for some constant $C_\rho > 0$ that

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

Apply Lemma 17, we know that

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

This in connection with Theorem 16 implies that

$$\|\vec{f}_\lambda\|_K \leq \|\nabla f_\rho\|_K + 2\|\nabla f_\rho\|_K + \frac{s}{\lambda} \kappa c'_\rho M_3 \leq 3\|\nabla f_\rho\|_K + (\kappa c_\rho)^{2/\tau} \kappa c'_\rho M_3.$$

Finally, we apply (28) of Theorem 13 and know that for a constant $C'_\rho > 0$, with confidence $1 - \delta$,

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

which is bounded by $C''_\rho \log(\frac{2}{\delta}) \left(\frac{1}{m} \right)^{\frac{\tau}{2(n+2+3\tau)}}$ 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 $m < (\kappa c_\rho)^{2(n+2+3\tau)/\tau}$. Then Proposition 8 is proved. \blacksquare

6. Simulated data and gene expression data

In this section we apply the least square gradient algorithm (7) to studying variable selection and variable covariances. Our idea is to rank the importance of variables according to the norm of their partial derivatives $\|\frac{\partial f_\rho}{\partial x^i}\|$, 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 20 *The relative magnitude of the norm for the variables is defined as*

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

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

Definition 21 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 $j = 1, \dots, m$. The **empirical covariance matrix (ECM)**, $\Xi_{\mathbf{z}}$, is the $n \times n$ matrix of inner products of the gradient between two coordinates

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

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 datasets. The first dataset 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 dataset provides a gold standard as to relevant variables.

6.1 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 $\{x_i\}_{i=1}^{10}$

$$x^j = \mathcal{N}(1, \sigma_x), \text{ for } j = 1, \dots, 10; \quad x^j = \mathcal{N}(0, \sigma_x), \text{ for } j = 11, \dots, 80.$$

2. For samples $\{x_i\}_{i=11}^{20}$

$$x^j = \mathcal{N}(1, \sigma_x), \text{ for } j = 11, \dots, 20; \quad x^j = \mathcal{N}(0, \sigma_x), \text{ for } j = 1, \dots, 10, 21, \dots, 80.$$

3. For samples $\{x_i\}_{i=21}^{30}$

$$x^j = \mathcal{N}(-1, \sigma_x), \text{ for } j = 41, \dots, 50; \quad x^j = \mathcal{N}(0, \sigma_x), \text{ for } j = 1, \dots, 40, 51, \dots, 80.$$

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

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

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 $\{y_i\}_{i=21}^{30}$

$$y_i = x_i \cdot w_3 + \mathcal{N}(0, \sigma_y).$$

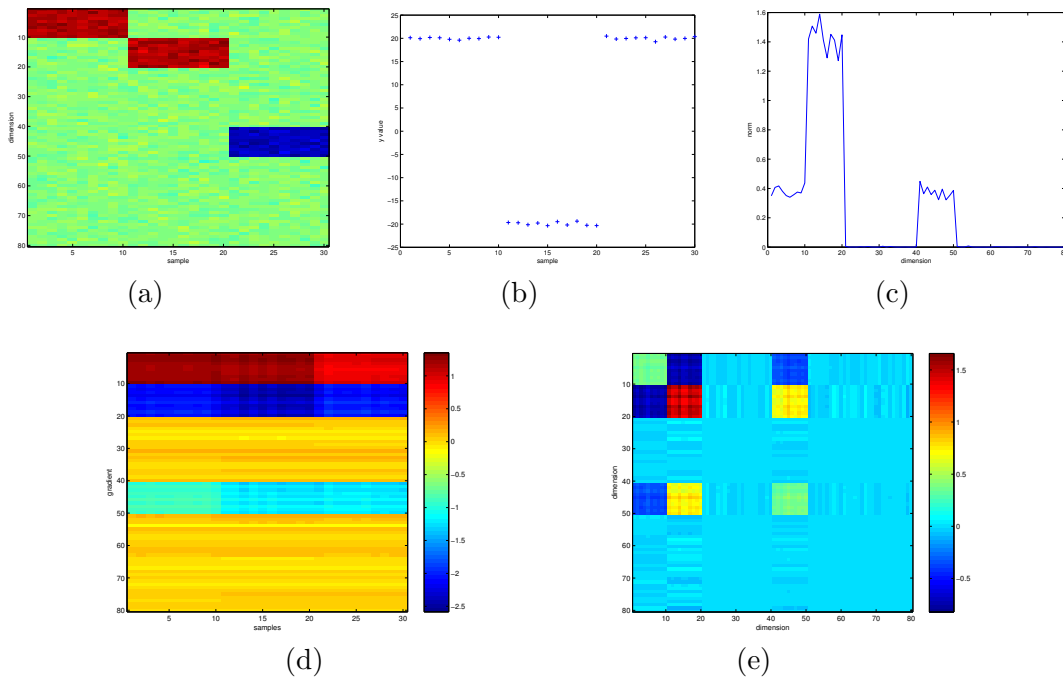


Figure 1: a) The data matrix x where each sample corresponds to a column, b) the vector of y values generated by sampling the function, c) the RKHS norm for each dimension, d) an estimate of the gradient at each sample, the samples correspond to columns, e) the empirical covariance matrix.

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, $\ell = 1, \dots, 20, 41, \dots, 50$.

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.

6.2 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.

6.2.1 LEUKEMIA CLASSIFICATION

We apply our procedure to a well studied expression dataset. The dataset 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 dataset 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 dataset 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 dataset 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 (\mathcal{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 dataset 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

$$s_\ell^F = \frac{t_\ell}{(\sum_{p=1}^n t_p^2)^{1/2}}.$$

Figure (2a-d) displays the relative decay of $s_{(\ell)}^\rho$ and $s_{(\ell)}^F$ over various numbers of dimensions. In all plots it is apparent that the decay rate of $s_{(\ell)}^\rho$ 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.

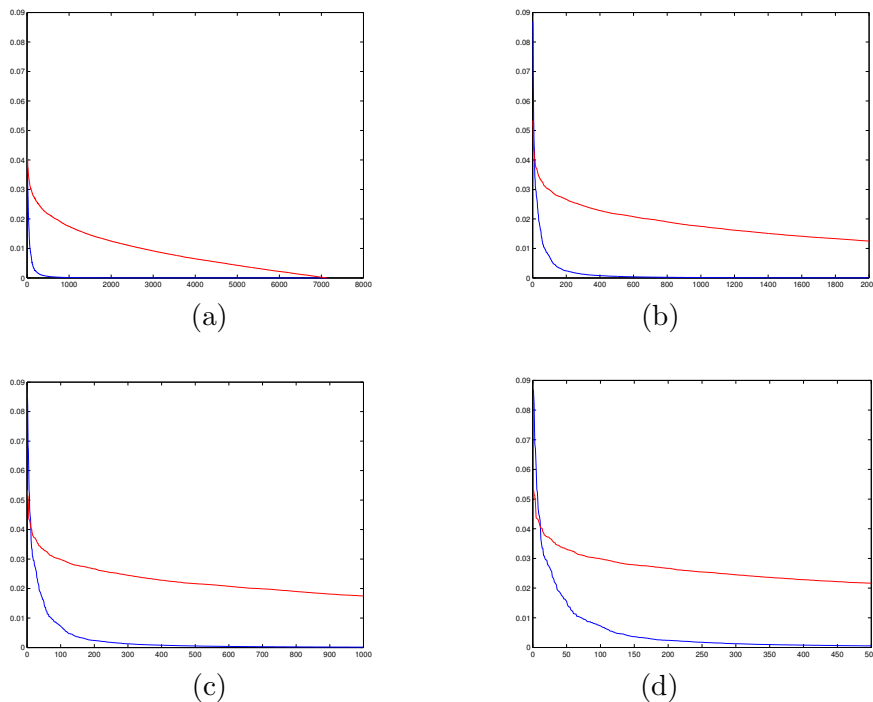


Figure 2: The decay of $s_{(\ell)}^{\rho}$ (blue) and $s_{(\ell)}^F$ (red) over: a) all the genes/dimensions, b) the top 2000 genes/dimensions, c) the top 1000 genes/dimensions, d) the top 500 genes/dimensions.

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 (3e). The covariation in the coordinates is plotted for the top 50 dimensions ordered in the same way as the EGM (see figure (3f)). The blocking structure of the matrix gives us coordinate covariance.

6.2.2 GENDER: “A GOLD STANDARD”

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

We examine a gene expression dataset with 15 male and 17 females samples from lymphoblastoid 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.

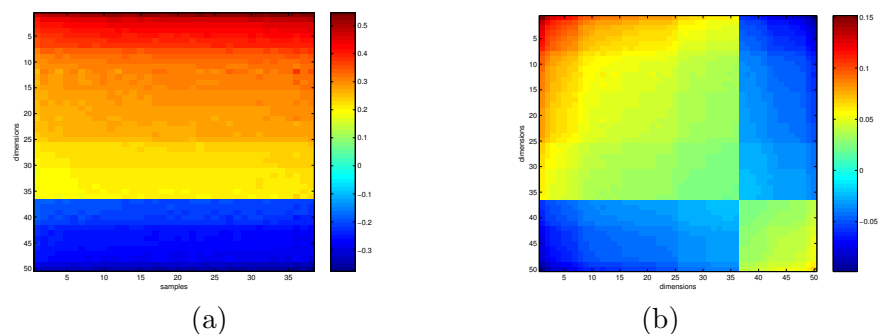


Figure 3: The a) EGM for the top 50 dimensions ordered by clustering the EGM and b) the ECM for the top 50 dimensions ordered in the same way.

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.

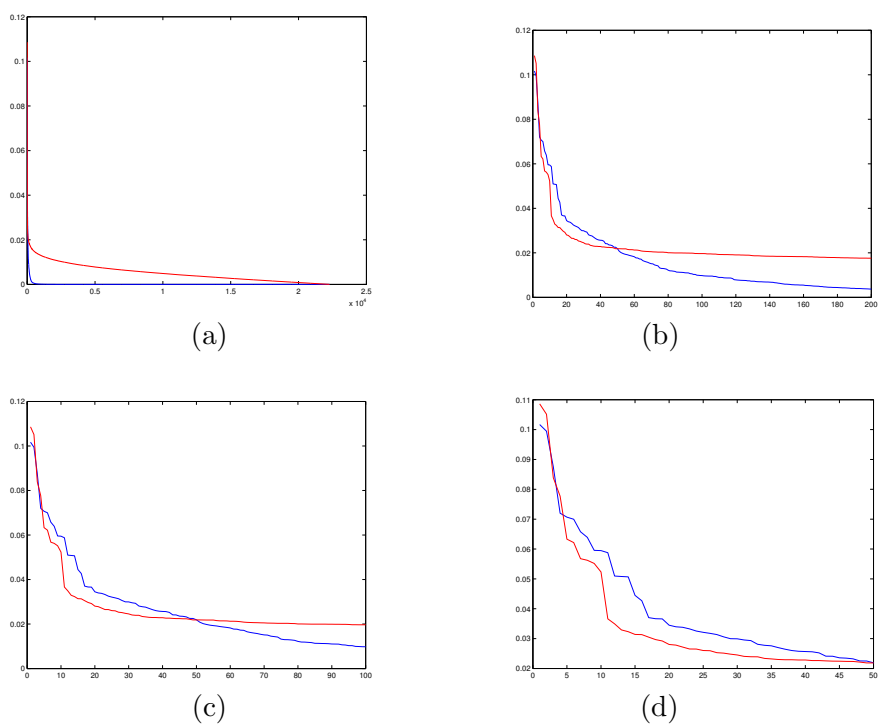


Figure 4: The decay of s_{ℓ}^{ρ} (blue) and s_{ℓ}^F (red) over: a) all the genes/dimensions, b) the top 200 genes/dimensions, c) the top 100 genes/dimensions, d) the top 50 genes/dimensions.

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 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 dataset 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 of the 101 probes ranked in the top 500 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, the 101 probes are ranked uniformly over the original 22,283. 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).

7. 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 8 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 datasets 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 unlabeled data. This is one of the motivations of semi-supervised learning. In many applications, it is much easier to obtain unlabeled data which in some sense increases the sample size $u \gg m$. For our purpose, unlabeled 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 unlabeled data to control the approximate norm of \vec{f} in some Sobolev spaces and introduce a semi-supervised learning algorithm as

$$\begin{aligned} \vec{f}_{\mathbf{z}, \mathbf{x}, \lambda, \mu} = \arg \min_{\vec{f} \in \mathcal{H}_K^n} & \left\{ \frac{1}{m^2} \sum_{i,j=1}^m w_{i,j}^{(s)} \left(y_i - y_j + \vec{f}(x_i) \cdot (x_j - x_i) \right)^2 \right. \\ & \left. + \frac{\mu}{(m+u)^2} \sum_{i,j=1}^{m+u} W_{i,j} |\vec{f}(x_i) - \vec{f}(x_j)|_{\ell^2(\mathbb{R}^n)}^2 + \lambda \|\vec{f}\|_K^2 \right\}, \end{aligned} \quad (35)$$

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, Misha Belkin, and Aravind Subramanian for useful discussions. 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 103704).

Appendix A.

The following is matlab code that implements algorithm (1), the approximation algorithm with reduced matrix size :

```

% 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 is 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-1);
for i=1:m-1
    M(:,i) = x(:,i)-x(:,m);
end

```

```

% computes the eigenvalues and eigenvectors of  $M^t M$ 
% and keeps  $s$  eigenvectors as specified by  $\epsilon$ 
d = eig(K);
W = transpose(M)*M;
[V,d] = eig(W);
d = diag(d);
vals = cumsum(d);
inds = find(vals/vals(m-1) < eps);
s = m-1-max(inds);

% since matlab indexes eigenvalues from smallest to largest we reverse
U = zeros(m-1,m-1);
dp = zeros(m-1,1);
for i=1:m-1
    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-1
    t(:,i) = sqrt(dp(1:s)).*transpose(U(i,1:s));
end
t(:,m) = zeros(s,1);

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
end

```

```

% solves the linear system for coefficients c
I = eye(m*s);
c = (m^2*lambda*I+Ktilde)\ytilde;

% unwraps 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. *Trans. Amer. Math. Soc.*, 686:337–404, 1950.
- M. Belkin and P. Niyogi. Semi-Supervised Learning on Riemannian Manifolds. *Mach. Learn.*, 56(1-3):209–239, 2004.
- I. Carrel, A. Cottle, K. Coglin, and H. Willard. A first-generation x-inactivation 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. *Mach. Learn.*, 46(1-3):131–159, 2002.
- 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. Disteché, 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. *Mach. Learn.*, 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. *J. Amer. Stat. Soc.*, 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.
- I. Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Probab.*, 22:1679–1706, 1994.
- 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. Shannon sampling and function reconstruction from point values. *Bull. Amer. Math. Soc.*, 41:279–305, 2004.
- S. Smale and D. X. Zhou. Learning theory estimates via integral operators and their approximations. Preprint, 2005a.
- S. Smale and D. X. Zhou. Shannon sampling II. Connections to learning theory. *Appl. Comput. Harmonic Anal.*, 2005b. to appear.
- 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.
- 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 Comp.*, 17:1160–1187, 2005.
- T. Zhang. Leave-one-out bounds for kernel methods. *Neural Comput.*, 15(6):1397–1437, 2003.
- D.X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Trans. Inform. Theory*, 49:1743–1752, 2003.