

Estimation of Gradients and Coordinate Covariation in Classification

Sayan Mukherjee and Qiang Wu

{SAYAN,QIANG}@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

Editor:

Abstract

We introduce an algorithm that simultaneously estimates a classification function as well as its gradient in the supervised learning framework. The motivation for the algorithm is to find salient variables and estimate how they covary. An efficient implementation with respect to both memory and time is given. The utility of the algorithm is illustrated on simulated data as well as a gene expression data set. An error analysis is given for the convergence of the estimate of the classification function and its gradient to the true classification function and true gradient.

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

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 classification (binary regression) problems.

In a number of these applications classical questions from statistical modeling of which variables are of relevance and how these variables interact arise. For example in analyzing genomic data in addition to providing a classification function an underlying interpretable model of the process may be desired, covarying features would correspond to genes co-regulated in a pathway. This leads to foundational questions in constructing and interpreting statistical models. Estimation of feature covariation is not considered in standard regression or classification methods that allow for variable selection: 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). Statistical models based on shrinkage or regularization were applied to the problem of learning coordinate covariation and relevance for regression problems in Mukherjee and Zhou (2005). We extend this approach to the binary regression or classification setting by simultaneously estimating the classification function as well as its gradient.

1.1 Learning the classification function and gradient

In this subsection we first formulate a solution to estimating a classification function via Tikhonov regularization algorithms or shrinkage estimators. This is done to define notation and basic concepts. We then motivate and introduce an algorithm that simultaneously estimates the classification function and its gradient.

Let X be a compact metric space and $Y = \{1, -1\}$. Let $\rho(x, y)$ be a probability distribution on $Z := X \times Y$ and $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ a random sample independently drawn according to $\rho(x, y)$.

A hypothesis space \mathcal{H} is a set of functions $X \rightarrow \mathbb{R}$. In this paper we will restrict \mathcal{H} to be a Reproducing Kernel Hilbert Space (RKHS) \mathcal{H}_K with an associated Mercer kernel $K : X \times X \rightarrow \mathbb{R}$ that is continuous, symmetric and positive semidefinite. The RKHS is defined (see Aronszajn (1950)) to be the completion of the linear span of the set of functions $\{K_x := K(\cdot, x) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ satisfying $\langle K_u, K_v \rangle_K = K(u, v)$. The reproducing property of \mathcal{H}_K is

$$\langle K_x, f \rangle_K = f(x), \quad \forall x \in X, f \in \mathcal{H}_K. \quad (1)$$

For the problem of binary regression the logistic loss function (Hastie et al., 2001)

$$\phi(yf(x)) = \log(1 + e^{-yf(x)}),$$

has a clear statistical interpretation (modelling the conditional probability $\rho(y|X)$ as a Bernoulli random variable)

$$\text{Prob}(y = \pm 1|x) = \frac{1}{1 + e^{-yf(x)}}.$$

In general, for a convex loss function ϕ we define the expected error of a function f as

$$\mathcal{R}(f) = \int \phi(yf(x)) d\rho(x, y), \quad (2)$$

and the classification function as the function in $L^2_{\rho_X}$, where ρ_X is the marginal distribution on x , that minimizes

$$f_\phi = \arg \min_{f \in L^2_{\rho_X}} \mathcal{R}(f).$$

Under certain conditions (Vapnik, 1998; Wu and Zhou, 2005) $\text{sgn}[f_\phi]$ is a Bayes optimal classifier.

Given a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ the Tikhonov regularization algorithm with logistic loss can be defined as

$$f_{\mathbf{z}} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m \phi(y_i f(x_i)) + \lambda \|f\|_K^2 \right\}, \quad (3)$$

where $\lambda > 0$ is the regularization parameter. The reproducing property tells us that the solution is a generalized linear model (GLM) (Hastie et al., 2001)

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

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 the error of $\text{sgn}(f_{\mathbf{z}})$ converges to the error of a Bayes optimal classifier with respect to the misclassification error:

$$\mathcal{C}(\text{sgn}(f)) = \text{Prob}\{\text{sgn}(f(x)) \neq y\}.$$

In this paper we are interested in simultaneously learning f_ϕ and its gradient from the sample values, $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$. Denote $x = (x^1, x^2, \dots, x^n)^T \in \mathbb{R}^n$. The gradient of f_ϕ is the vector of functions (if the partial derivatives exist)

$$\nabla f_\phi = \left(\frac{\partial f_\phi}{\partial x^1}, \dots, \frac{\partial f_\phi}{\partial x^n} \right)^T. \quad (4)$$

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 $\|\frac{\partial f_\phi}{\partial x^i}\|$ indicates the relevance of this variable, since a small norm implies a small change in the discriminative function f_ϕ with respect to the i -th coordinate,

(b) coordinate covariation: the inner product between partial derivatives $\left\langle \frac{\partial f_\phi}{\partial x^i}, \frac{\partial f_\phi}{\partial x^j} \right\rangle$ indicates the covariance of the i -th and j -th coordinates with respect to variation in f_ϕ .

The derivation of our gradient learning algorithm can be motivated by the Taylor expansion of a function $g(u)$ around the point x

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

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). Given a sparse sample from the classification function $\{(x_i, f_\phi(x_i))\}_{i=1}^m$ one would expect that

$$f_\phi(x_i) \approx f_\phi(x_j) + \nabla f_\phi(x_j) \cdot (x_j - x_i) \quad \text{for } x_i \approx x_j. \quad (5)$$

Our objective will be to estimate the gradient by a vector $\vec{f} \in \mathcal{H}_K^n$ which can be written as a column vector $\vec{f} = (f_1, f_2, \dots, f_n)^T$ with each $f_\ell \in \mathcal{H}_K$ with the following inner product $\langle \vec{f}, \vec{h} \rangle = \sum_{i=1}^n \langle f_i, h_i \rangle_K$. Since we are not given a sparse sample of the classification function $\{(x_i, f_\phi(x_i))\}_{i=1}^m$ but instead are given the sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ we will also have to estimate f_ϕ by a function g . So the condition in equation (5) will be satisfied if the following error is small (given the estimates g and \vec{f}):

$$\phi(y_i(g(x_j) + \vec{f}(x_i) \cdot (x_i - x_j))) \quad \text{for } x_i \approx x_j. \quad (6)$$

The restriction $x_i \approx x_j$ is imposed by a set of weights: $w_{i,j} = w_{i,j}^{(s)} > 0$ with the requirement that $w_{i,j}^{(s)} \rightarrow 0$ as $\|x_i - x_j\|/s \rightarrow \infty$. Throughout this paper we will use a Gaussian with variance s as our weight function:

$$w_{i,j} = w_{i,j}^{(s)} = \frac{1}{s^{n+2}} e^{-\frac{\|x_i - x_j\|^2}{2s^2}} = w(x_i - x_j), \quad i, j = 1, \dots, m. \quad (7)$$

Given condition (6) and weights (7) the following natural empirical error function can be defined.

Definition 1 *Given a sample $\mathbf{z} \in Z^m$, a function $g : X \rightarrow \mathbb{R}$, and a vector-valued function $\vec{f} : X \rightarrow \mathbb{R}^n$, we define the empirical error as follows:*

$$\mathcal{E}_{\mathbf{z}}(g, \vec{f}) = \frac{1}{m^2} \sum_{i,j=1}^m w_{i,j}^{(s)} \phi(y_i(g(x_j) + \vec{f}(x_i) \cdot (x_i - x_j))). \quad (8)$$

Regularizing or shrinking the above empirical error defines the following optimization problem.

Definition 2 *Given a sample $\mathbf{z} \in Z^m$ we can estimate the classification function, $g_{\mathbf{z}}$, and its gradient, $\vec{f}_{\mathbf{z}}$, as follows:*

$$(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}}) = \arg \min_{(g, \vec{f}) \in \mathcal{H}_K^{n+1}} \left\{ \mathcal{E}_{\mathbf{z}}(g, \vec{f}) + \frac{\lambda}{2} (\|g\|_K^2 + \|\vec{f}\|_K^2) \right\}, \quad (9)$$

where $s, \lambda > 0$ are the regularization parameters.

Remark 3 *Our method can be regarded as an extension of algorithms for numerical derivatives in high dimensional spaces. At first thought, a natural approach to computing partial derivatives would be to estimate the classification function and then compute partial derivatives. The problem with this approach is that the partial derivatives are no longer in the RKHS of the classification 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.*

The hypothesis space \mathcal{H}_K^n in the optimization problem (9) may be replaced by some other space of vector-valued functions (Micchelli and Pontil, 2005) in order to learn the gradients.

1.2 Overview

In Section 2, we show that the minimizer of the optimization problem (9) satisfies a representer theorem and then provide a procedure to compute the parameters. In Section 3, we prove the convergence of our estimate of the gradient $\vec{f}_{\mathbf{z}}(x)$ to the gradient of the classification function, ∇f_{ϕ} . The utility of the algorithm is illustrated in section 4 on simulated data as well as gene expression data. We close with a brief discussion in Section 5.

2. Representer theorem and parameter computation

The optimization problem defined by equation (9) is a convex optimization problem in the sense that $\phi(\cdot)$, $\|g\|_K^2$, and $\|\vec{f}\|_K^2$ are all convex functionals. Denote $\mathbb{R}^{p \times q}$ as the space of $p \times q$ matrices.

The following theorem is an analog of the standard representer theorem (Schoelkopf and Smola, 2001; Wahba, 1990) that states the minimizer of the optimization problem defined by equation (9) can be represented as a generalized linear model.

Proposition 4 *Given a sample $\mathbf{z} \in Z^m$ the solution of (9) exists and takes the form*

$$g_{\mathbf{z}}(x) = \sum_{i=1}^m \alpha_{i,\mathbf{z}} K(x, x_i) \quad \text{and} \quad \vec{f}_{\mathbf{z}}(x) = \sum_{i=1}^m c_{i,\mathbf{z}} K(x, x_i) \quad (10)$$

with $c_{\mathbf{z}} = (c_{1,\mathbf{z}}, \dots, c_{m,\mathbf{z}}) \in \mathbb{R}^{n \times m}$ and $\alpha_{\mathbf{z}} = (\alpha_{1,\mathbf{z}}, \dots, \alpha_{m,\mathbf{z}})^T \in \mathbb{R}^m$.

Proof The existence follows from the convexity of ϕ and functionals $\|g\|_K^2$ and $\|\vec{f}\|_K^2$. Suppose $(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}})$ is a minimizer. We can write functions $g_{\mathbf{z}} \in \mathcal{H}_K$ and $\vec{f}_{\mathbf{z}} \in \mathcal{H}_K^n$ as

$$g_{\mathbf{z}} = g_{\parallel} + g_{\perp} \quad \text{and} \quad \vec{f}_{\mathbf{z}} = \vec{f}_{\parallel} + \vec{f}_{\perp}$$

where g_{\parallel} and each element of \vec{f}_{\parallel} is in the span of $\{K_{x_1}, \dots, K_{x_m}\}$, and g_{\perp} and \vec{f}_{\perp} are functions in the orthogonal complement. By the reproducing property, there hold $g_{\mathbf{z}}(x_i) = g_{\parallel}(x_i)$ and $\vec{f}_{\mathbf{z}}(x_i) = \vec{f}_{\parallel}(x_i)$ for all x_i . So the functions g_{\perp} and \vec{f}_{\perp} do not have an effect on $\mathcal{E}_{\mathbf{z}}(g, \vec{f})$. But $\|g_{\mathbf{z}}\|_K^2 = \|g_{\parallel} + g_{\perp}\|_K^2 > \|g_{\parallel}\|_K^2$ and $\|\vec{f}_{\mathbf{z}}\|_K^2 = \|\vec{f}_{\parallel} + \vec{f}_{\perp}\|_K^2 > \|\vec{f}_{\parallel}\|_K^2$ unless $g_{\perp}, \vec{f}_{\perp} = 0$. This implies that $g_{\mathbf{z}} = g_{\parallel}$ and $\vec{f}_{\mathbf{z}} = \vec{f}_{\parallel}$. This results in the representations in equation (10). ■

The optimization in equation (9) can be written in terms of the coefficients $c_{\mathbf{z}}$ and $\alpha_{\mathbf{z}}$. We define a matrix $C = (c_1, c_2, \dots, c_m) \in \mathbb{R}^{n \times m}$ (when optimized these will be the coefficients $c_{\mathbf{z}}$ in the gradient expansion) and the vector $\alpha \in \mathbb{R}^m$ (when optimized the vector will be $\alpha_{\mathbf{z}}$). We denote the kernel matrix K where $K_{ij} = K(x_i, x_j)$ for $i, j = 1, \dots, m$ and the i -th row of the matrix as k_i . The optimization function can be written in matrix form as

$$\Phi(C, \alpha) = \frac{1}{m^2} \sum_{i,j=1}^m w_{i,j} \phi(y_i(k_j \alpha + k_i C^T(x_i - x_j))) + \frac{\lambda}{2} (\alpha^T K \alpha + \text{Tr}(C K C^T)), \quad (11)$$

where $\text{Tr}(M)$ is the trace of a matrix M .

Proposition 5 *If ϕ is differentiable, then the coefficients $c_{\mathbf{z}}$ and $\alpha_{\mathbf{z}}$ can be computed from the equation $\nabla \Phi(\alpha, C) = 0$.*

We can optimize (11) by using Newton's method to solve $\nabla \Phi(\alpha, C) = 0$. The matrix C however is an $n \times m$ matrix and optimizing in \mathbb{R}^{mn} is problematic for applications where $n \gg m$. We will show that the coefficients can be computed by the optimization of an $m \times d$ matrix, where typically $d \ll m$. We will then apply Newton's method in this reduced space.

Define a vector-valued function

$$h = ((h^0)^T, (h_1)^T, \dots, (h_m)^T)^T : \mathbb{R}^{(n+1)m} \rightarrow \mathbb{R}^{(n+1)m}$$

with

$$h^0 = (h_1^0, \dots, h_m^0)^T, \quad h_j^0(\alpha, C) = \frac{1}{m^2} \sum_{i=1}^m w_{i,j} \phi'(y_i(k_j \alpha + k_i C^T(x_j - x_i))) y_i + \lambda \alpha_j$$

and, for $i = 1, \dots, m$,

$$h_i(\alpha, C) = \frac{1}{m^2} \sum_{j=1}^m w_{i,j} \phi'(y_i(k_j \alpha + k_i C^T(x_j - x_i))) y_i (x_i - x_j) + \lambda c_i.$$

The following proposition states that the coefficients $c_{i,\mathbf{z}}$ are in the span of the pairwise differences in the data.

Proposition 6 *If the solution to the equation $h(\alpha, C) = 0$ exists, then the coefficients $c_{\mathbf{z}}$ in the representation of $\vec{f}_{\mathbf{z}}$ satisfy the constraint for every $i = 1, \dots, m$ $c_{i,\mathbf{z}} \in V_{\mathbf{x}} = \text{span}\{x_i - x_j : i, j = 1, \dots, m\}$.*

Proof Solving for the coefficients will give us the result. By direct computation, we have

$$\nabla \Phi(\alpha, C) = \begin{pmatrix} K & 0 \\ 0 & K \otimes I_n \end{pmatrix} h(\alpha, C) \quad (12)$$

where I_n is the $n \times n$ identity matrix. By the assumption, there exists $(\alpha_{\mathbf{z}}, c_{\mathbf{z}})$ solving the equation $h(\alpha, C) = 0$. So $\nabla \Phi(\alpha_{\mathbf{z}}, c_{\mathbf{z}}) = 0$ and $(\alpha_{\mathbf{z}}, c_{\mathbf{z}})$ gives the representation of $g_{\mathbf{z}}$ and $\vec{f}_{\mathbf{z}}$. By the definition of h , we have $h_i(\alpha_{\mathbf{z}}, c_{\mathbf{z}}) = 0$ which implies $c_{i,\mathbf{z}} \in V_{\mathbf{x}}$. This proves the proposition. \blacksquare

Remark 7 *We know the solution $(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}})$ exists and even is unique. This implies the existence of the solution to $\nabla \Phi(\alpha, C) = 0$. But the existence of the solution to $h(\alpha, C) = 0$ is not clear. In fact, this may not be always the case when K is not invertible.*

Proposition 6 allows us to reduce the dimension of the optimization problem for solving the coefficients $c_{\mathbf{z}}$. We will use the well known approach of singular value decomposition to the matrix involving the data \mathbf{x} given by

$$M_{\mathbf{x}} = (x_1 - x_m, x_2 - x_m, \dots, x_{m-1} - x_m, x_m - x_m) \in \mathbb{R}^{n \times m}. \quad (13)$$

Assume the rank of $M_{\mathbf{x}}$ is d . The theory of singular value decomposition tells us that there exists an $n \times n$ orthogonal matrix $V = (V_1, V_2, \dots, V_n)$ and an $m \times m$ orthogonal matrix $U = (U_1, U_2, \dots, U_m)$ such that

$$M_{\mathbf{x}} = V \Sigma U^T = (V_1 \ V_2 \ \dots \ V_n) \begin{pmatrix} \text{diag}\{\sigma_1, \sigma_2, \dots, \sigma_d\} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U_1^T \\ U_2^T \\ \vdots \\ U_m^T \end{pmatrix}. \quad (14)$$

Here $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_d > \sigma_{d+1} = \dots = \sigma_{\min\{m,n\}} = 0$ 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 $i = 1, \dots, d$. The vectors $\{V_i\}_{i=1}^d$ form an orthonormal basis for $V_{\mathbf{x}}$ and denote $V = (V_1, \dots, V_d)$.

Set $\beta_i \in \mathbb{R}^d$ to satisfy $x_i - x_m = V\beta_i$ for $i = 1, \dots, m$. For $\gamma^0 \in \mathbb{R}^m$ and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_m) \in \mathbb{R}^{d \times m}$, define the vector-valued function

$$u = ((u^0)^T, (u_1)^T, \dots, (u_m)^T)^T : \mathbb{R}^{m(d+1)} \rightarrow \mathbb{R}^{m(d+1)}$$

by

$$u^0 = (u_1^0, \dots, u_m^0)^T, \quad u_j^0(\gamma^0, \boldsymbol{\gamma}) = \frac{1}{m^2} \sum_{i=1}^m w_{i,j} \phi'(y_i(k_j \gamma^0 + k_i \boldsymbol{\gamma}^T(\beta_i - \beta_j))) y_i + \lambda \gamma_j^0,$$

and, for $i = 1, \dots, m$,

$$u_i(\gamma^0, \boldsymbol{\gamma}) = \frac{1}{m^2} \sum_{j=1}^m w_{i,j} \phi'(y_i(k_j \gamma^0 + k_i \boldsymbol{\gamma}^T(\beta_i - \beta_j))) y_i(\beta_i - \beta_j) + \lambda \gamma_i.$$

Proposition 8 *If $\gamma_{\mathbf{z}}^0 \in \mathbb{R}^m$ and $\boldsymbol{\gamma}_{\mathbf{z}} = (\gamma_{1,\mathbf{z}}, \dots, \gamma_{m,\mathbf{z}}) \in \mathbb{R}^{d \times m}$ are solutions of the equation $u(\gamma^0, \boldsymbol{\gamma}) = 0$, then $c_{\mathbf{z}}$ and $\alpha_{\mathbf{z}}$ defined by*

$$\alpha_{\mathbf{z}} = \gamma_{\mathbf{z}}^0, \quad c_{i,\mathbf{z}} = V\gamma_{i,\mathbf{z}} \quad \text{for } i = 1, \dots, m,$$

solve $\nabla \Phi(\alpha, C) = 0$ and hence yield a representation of $g_{\mathbf{z}}$ and $\vec{f}_{\mathbf{z}}$ respectively.

Proof By the facts that $c_i = V\gamma_i$ for $i = 1, \dots, m$ defines a one-to-one mapping from $V_{\mathbf{x}}$ onto \mathbb{R}^d and $V^T V = I_d$ the d -dimensional identity matrix, direct computation shows that $u(\gamma_{\mathbf{z}}^0, \boldsymbol{\gamma}_{\mathbf{z}}) = 0$ implies $h(\alpha_{\mathbf{z}}, c_{\mathbf{z}}) = 0$. Then the conclusion follows from Proposition 5 and equation (12). \blacksquare

We now use Proposition 8 to derive the update rule in Newton's method to optimize the coefficients γ^0 and $\boldsymbol{\gamma}$. Let $\boldsymbol{\eta} = ((\gamma^0)^T, (\gamma_1)^T, \dots, (\gamma_m)^T)^T \in \mathbb{R}^{m(d+1)}$ and consider the map $u(\boldsymbol{\eta})$ on $\mathbb{R}^{m(d+1)}$ defined as $u = ((u^0)^T, (u_1)^T, \dots, (u_m)^T)^T$. When ϕ is twice differentiable, we can use Newton's method to solve the equation $u(\boldsymbol{\eta}) = 0$ by the following iterative update rule

$$\boldsymbol{\eta}_{t+1} = \boldsymbol{\eta}_t - (\nabla u(\boldsymbol{\eta}_t))^{-1} u(\boldsymbol{\eta}_t).$$

This leaves us with the computation of the matrix $\nabla u(\boldsymbol{\eta})$:

$$\nabla u(\boldsymbol{\eta}) = \lambda I_{m(d+1)} + \frac{1}{m^2} \begin{pmatrix} \text{diag}(b_1, \dots, b_m)K & A_1 \\ A_2 & \text{diag}(B_1, \dots, B_m)(K \otimes I_d) \end{pmatrix},$$

where $I_{m(d+1)}$ is the $m(d+1)$ square identity matrix, denote

$$t_{i,j} = t_{i,j}(\boldsymbol{\eta}) = \phi''(y_i(k_j \gamma^0 + k_i \boldsymbol{\gamma}^T(\beta_i - \beta_j))),$$

and

$$\begin{aligned}
 A_1 &= \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} \quad \text{with} \quad a_j = \sum_{i=1}^m w_{i,j} t_{i,j} k_i \otimes (\beta_i - \beta_j)^T \\
 A_2 &= \begin{pmatrix} A_{21} \\ \vdots \\ A_{2m} \end{pmatrix} \quad \text{with} \quad A_{2i} = \sum_{j=1}^m w_{i,j} t_{i,j} k_j \otimes (\beta_i - \beta_j), \\
 b_j &= \sum_{i=1}^m w_{i,j} t_{i,j}, \quad B_i = \sum_{j=1}^m w_{i,j} t_{i,j} (\beta_i - \beta_j) (\beta_i - \beta_j)^T.
 \end{aligned}$$

3. Error analysis

In this section, we investigate the statistical performance of the algorithm. We will show that under certain mild conditions, $g_{\mathbf{z}} \rightarrow f_\phi$ and $\vec{f}_{\mathbf{z}} \rightarrow \nabla f_\phi$ as $\lambda, s \rightarrow 0$. Let us first illustrate this by a specific case where $\phi(\cdot)$ is the logistic loss and $(f_\phi, \nabla f_\phi) \in \mathcal{H}_K^{n+1}$ (this case corresponds to the realizable setting in the PAC learning paradigm).

Theorem 9 *Let ϕ be the logistic loss. Assume that for some constants $c_\rho > 0$ and $0 < \theta \leq 1$ the marginal distribution ρ_X satisfies*

$$\rho_X(\{x \in X : d(x, \partial X) < s\}) \leq c_\rho s, \quad (15)$$

and the density $p(x)$ of ρ_X exists and satisfies

$$\sup_{x \in X} p(x) \leq c_\rho \quad \text{and} \quad |p(x) - p(u)| \leq c_\rho |x - u|^\theta, \quad \forall u, x \in X. \quad (16)$$

Suppose that $K \in C^2$ and $(f_\phi, \nabla f_\phi) \in \mathcal{H}_K^{n+1}$. Choose $\lambda = \lambda(m) = m^{-\frac{2\theta}{3(n+2+2\theta)}}$ and $s = s(m) = m^{-\frac{1}{3(n+2+2\theta)}}$. Then there exists a constant $C > 0$ such that for any $0 < \eta < 1$ with confidence $1 - \eta$

$$\begin{aligned}
 \|g_{\mathbf{z}} - f_\phi\|_{L_{\rho_X}^2} &\leq C \log \frac{4}{\eta} \left(\frac{1}{m} \right)^{\frac{\theta}{6(n+2+2\theta)}}, \\
 \|\vec{f}_{\mathbf{z}} - \nabla f_\phi\|_{L_{\rho_X}^2} &\leq C \log \frac{4}{\eta} \left(\frac{1}{m} \right)^{\frac{\theta}{6(n+2+2\theta)}}.
 \end{aligned}$$

Condition (16) means the density of the marginal distribution is Hölder θ . The condition (15) is about the behavior of ρ_X near the boundary of X . When the boundary is piecewise smooth, (16) implies (15).

Theorem 9 is a consequence of the more general Theorem 10 which we prove in Section 3.3.

We first define two quantities that will be used extensively.

$$\kappa = \sup_{x \in X} \sqrt{K(x, x)}; \quad D = \max_{x, u \in X} |x - u|.$$

Note that the reproducing property (1) of the RKHS \mathcal{H}_K implies $\|f\|_\infty \leq \kappa\|f\|_K$ for $f \in \mathcal{H}_K$. This will be used constantly in the following.

For a convex loss function ϕ and $r > 0$, define

$$\begin{aligned} L_r &= \max \{ |\phi'(\kappa(1+D)r)|, |\phi'(-\kappa(1+D)r)| \}, \\ M_r &= \max \{ \phi(\kappa(1+D)r), \phi(-\kappa(1+D)r) \}. \end{aligned}$$

By convexity of ϕ both L_r and M_r increase with r .

Theorem 10 *Let the convex loss function ϕ be twice differentiable and satisfy*

$$q_1(T) = \inf_{|t| \leq T} \phi''(t) > 0, \quad q_2(T) = \sup_{|t| \leq T} \phi''(t) < \infty.$$

Assume ρ satisfies (15) and (16), $K \in C^2$, $(f_\phi, \nabla f_\phi) \in \mathcal{H}_K^{n+1}$. Then there exists a constant \tilde{C} such that for $0 < \delta < 1/2$, $0 < s, \lambda \leq 1$ with probability at least $1 - 2\delta$

$$\max \left\{ \|g_{\mathbf{z}} - f_\phi\|_{L_{\rho_X}^2}^2, \|\vec{f}_{\mathbf{z}} - \nabla f_\phi\|_{L_{\rho_X}^2}^2 \right\} \leq \tilde{C} \left\{ r^2 s^\theta + B_r \left(\frac{L_r r + M_r \log \frac{2}{\delta}}{\sqrt{m} s^{n+2}} + s^2 + \lambda \right) s^{-\theta} \right\},$$

where

$$r = \tilde{c} \left\{ 1 + \frac{s^2}{\lambda} + \left(\frac{L_{\lambda,s}}{\sqrt{\lambda} s^{n+2}} + M_{\lambda,s} \log \frac{2}{\delta} \right) \frac{1}{\sqrt{m} \lambda s^{n+2}} \right\}^{1/2} \quad (17)$$

with some $\tilde{c} \geq 1$, $L_{\lambda,s} = L_{\sqrt{2\phi(0)/\lambda} s^{n+2}}$, and $M_{\lambda,s} = M_{\sqrt{2\phi(0)/\lambda} s^{n+2}}$ and $B_r = \min \left\{ \frac{1}{q_1(c_0 r)}, r \right\}$ with some $c_0 > 0$.

Remark 11 *Theorem 10 applies only to the loss functions satisfying $\phi''(t) > 0$ because of the quantity B_r . But it does not increase very fast with r . We can take $B_r = r$ for logistic loss and exponential loss where $q_1(T)$ decays exponentially fast with T . While for the square loss, $B_r = 1$ for $q_1(T) \equiv 1$.*

The idea behind the proof of Theorem 10 is to first bound the $L_{\rho_X}^2$ differences by the excess error in Section 3.1 and then bound the excess error in Section 3.2. It will be done in Section 3.3. Before that, let us prove Theorem 9 by using Theorem 10.

Proof of Theorem 9. Note that for logistic loss, $\phi'(t) = \frac{-e^{-t}}{1+e^{-t}} \in (-1, 1)$. So $L_r \leq 1$ and $L_{\lambda,s} \leq 1$. Since $\phi(t) \leq \phi(0) + |t| < 1 + |t|$, we have $M_r \leq (1 + \kappa(1+D))r$ when $r \geq 1$ and so $M_{\lambda,s} \leq 2(1 + \kappa(1+D))(\lambda s^{n+2})^{-1}$. Also, $\phi''(t) = \frac{2e^{-t}}{(1+e^{-t})^2}$ implies $\frac{1}{q_1(r)} \geq c_0 r$, $q_2(c_0 r) \leq 1/2$. Substitute $L_{\lambda,s}$ and $M_{\lambda,s}$ into (17). The choice of λ, s ensures that $r \leq r_0$ with $r_0 > 1$ an absolute constant. Since B_r, L_r, M_r are increasing with respect to r , so is the upper bound in Theorem 10. Substituting r_0 into this upper bound and by the choice of λ, s , we obtain with confidence at least $1 - 2\delta$

$$\max \left\{ \|g_{\mathbf{z}} - f_\phi\|_{L_{\rho_X}^2}^2, \|\vec{f}_{\mathbf{z}} - \nabla f_\phi\|_{L_{\rho_X}^2}^2 \right\} \leq C \log \frac{2}{\delta} \left(\frac{1}{m} \right)^{\frac{\theta}{6(n+2+2\theta)}},$$

where

$$C = \tilde{C} \left((r_0)^2 + B_{r_0} (L_{r_0} r_0 + M_{r_0} + 2) \right).$$

Setting $\delta = \frac{\eta}{2}$ finishes the proof. ■

3.1 Bounding $L^2_{\rho_X}$ differences by the excess error

Recall the empirical error (Definition 1) for $(g, \vec{f}) : X \rightarrow \mathbb{R}^{n+1}$

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

One can similarly defined the expected error

$$\mathcal{E}(g, \vec{f}) = \int_{\mathcal{Z}} \int_X w(x-u) \phi(y(g(u) + \vec{f}(x) \cdot (x-u))) d\rho_X(u) d\rho(x, y).$$

Unlike the standard setting of classification and regression $\mathcal{E}(g, \vec{f})$ and $\mathcal{E}_{\mathbf{z}}(g, \vec{f})$ are not respectively the expected and empirical mean of a random variable. This is due to the extra $d\rho_X$ in the expected error term. However, since

$$\mathbb{E}_{\mathbf{z}}[\mathcal{E}_{\mathbf{z}}(g, \vec{f})] = \frac{1}{ms^{n+2}} \mathcal{R}(g) + \frac{m-1}{m} \mathcal{E}(g, \vec{f}),$$

the empirical and expected errors should be close to each other if the empirical error concentrates with m increasing.

Define

$$\mathcal{R}_s = \int_X \int_{\mathcal{Z}} w(x-u) \phi(y f_{\phi}(x)) d\rho(x, y) d\rho_X(u).$$

We will use the *excess error*, $\mathcal{E}(g, \vec{f}) - \mathcal{R}_s$, to bound the $L^2_{\rho_X}$ differences.

For $r > 0$, denote

$$\mathcal{F}_r = \left\{ (g, \vec{f}) \in \mathcal{H}_K^{n+1} : \|g\|_K^2 + \|\vec{f}\|_K^2 \leq r^2 \right\}.$$

Theorem 12 *Assume ρ_X satisfies the conditions (15) and (16) and $(f_{\phi}, \nabla f_{\phi}) \in \mathcal{H}_K^{n+1}$. For $(g, \vec{f}) \in \mathcal{F}_r$ with some $r > 1$, there exist constants $C_0, C_1 > 0$ such that*

$$\|g - f_{\phi}\|_{L^2_{\rho_X}}^2 \leq C_0 \left(s^{\theta} r^2 + s^{2-\theta} B_r(\mathcal{E}(g, \vec{f}) - \mathcal{R}_s) \right)$$

and

$$\|f - \nabla f_{\phi}\|_{L^2_{\rho_X}}^2 \leq C_1 \left(s^{\theta} r^2 + s^{-\theta} B_r(\mathcal{E}(g, \vec{f}) - \mathcal{R}_s) \right),$$

where $B_r = \min \left\{ \frac{1}{q_1(c_0 r)}, r \right\}$ with some $c_0 > 0$.

To prove Theorem 12 we will need the following several lemmas which require the definition of the following quantities.

Definition 13 *Define for $(g, \vec{f}) : X \rightarrow \mathbb{R}^{n+1}$ the square error functional*

$$\mathcal{Q}(g, \vec{f}) = \int_X \int_X w(x-u) \left(g(x) - f_{\phi}(x) + (\vec{f}(x) - \nabla f_{\phi}(x)) \cdot (x-u) \right)^2 d\rho_X(u) d\rho_X(x),$$

the border set

$$X_s = \left\{ x \in X : d(x, \partial X) > s \text{ and } p(x) \geq (1 + c_\rho)s^\theta \right\},$$

and the moments for $0 \leq p < \infty$,

$$N_p = \int_{\{t \in \mathbb{R}^n : |t| \leq 1\}} e^{\frac{|t|^2}{2}} |t|^p dt, \quad \text{and} \quad \tilde{N}_p = \int_{\mathbb{R}^n} e^{\frac{|t|^2}{2}} |t|^p dt.$$

Lemma 14 *Under assumptions of Theorem 12*

$$\frac{N_0}{s^{2-\theta}} \int_{X_s} (g(x) - f_\phi(x))^2 d\rho_X(x) + \frac{N_2 s^\theta}{n} \int_{X_s} |\vec{f}(x) - \nabla f_\phi(x)|^2 d\rho_X(x) \leq \mathcal{Q}(g, \vec{f}).$$

Proof For $x \in X_s$, $\{u \in X : |u - x| \leq s\} \subset X$ since $d(x, \partial X) > s$. For $u \in X$ such that $|u - x| \leq s$

$$p(u) = p(x) - (p(x) - p(u)) \geq (1 + c_\rho)s^\theta - c_\rho|u - x|^\theta \geq s^\theta.$$

Therefore,

$$\begin{aligned} \mathcal{Q}(g, \vec{f}) &\geq \int_{X_s} \int_{|u-x| \leq s} w(x-u) \left(g(x) - f_\phi(x) + (\vec{f}(x) - \nabla f_\phi(x)) \cdot (x-u) \right)^2 p(u) du d\rho_X(x) \\ &\geq s^\theta \int_{X_s} \int_{|u-x| \leq s} w(x-u) \left(g(x) - f_\phi(x) + (\vec{f}(x) - \nabla f_\phi(x)) \cdot (x-u) \right)^2 du d\rho_X(x) \\ &= s^\theta \int_{X_s} \int_{|u-x| \leq s} w(x-u) (g(x) - f_\phi(x))^2 du d\rho_X(x) \\ &\quad + 2s^\theta \int_{X_s} \int_{|u-x| \leq s} w(x-u) (g(x) - f_\phi(x)) ((\vec{f}(x) - \nabla f_\phi(x)) \cdot (x-u)) du d\rho_X(x) \\ &\quad + s^\theta \int_{X_s} \int_{|u-x| \leq s} w(x-u) ((\vec{f}(x) - \nabla f_\phi(x)) \cdot (x-u))^2 du d\rho_X(x) \\ &: = J_1 + J_2 + J_3. \end{aligned}$$

It can be verified that

$$J_1 = \frac{1}{s^{2-\theta}} \int_{X_s} (g(x) - f_\phi(x))^2 \int_{|t| \leq 1} e^{-\frac{|t|^2}{2}} dt d\rho_X(x) = \frac{N_0}{s^{2-\theta}} \int_{X_s} |g(x) - f_\phi(x)|^2 d\rho_X(x).$$

For every $i \in \{1, \dots, n\}$

$$\int_{|u-x| \leq s} w(x-u) (x^i - u^i) du = \frac{1}{s} \int_{|t| \leq 1} e^{\frac{|t|^2}{2}} t^i dt = 0.$$

It follows that $J_2 = 0$.

Note that $((\vec{f}(x) - \nabla f_\phi(x)) \cdot (x-u))^2$ equals to

$$\sum_{i=1}^n \sum_{j=1}^n \left(f^i(x) - \frac{\partial f_\phi}{\partial x^i}(x) \right) \left(f^j(x) - \frac{\partial f_\phi}{\partial x^j}(x) \right) (x^i - u^i)(x^j - u^j).$$

But when $j \neq i$,

$$\int_{|u-x| \leq s} w(x-u) (x^i - u^i)(x^j - u^j) du = \int_{|t| \leq 1} e^{\frac{|t|^2}{2}} t^i t^j dt = 0.$$

Therefore

$$J_3 = s^\theta \sum_{i=1}^n \int_{X_s} (f^i(x) - \frac{\partial f_\phi}{\partial x^i}(x))^2 \int_{|t| \leq 1} e^{-\frac{|t|^2}{2}} (t^i)^2 dt d\rho_X(x) = \frac{N_2 s^\theta}{n} \int_{X_s} |\vec{f}(x) - \nabla f_\phi(x)|^2 d\rho_X(x).$$

Plugging J_1 and J_3 into the inequality completes the proof. \blacksquare

In Lemma 17 below we will bound $\mathcal{Q}(g, \vec{f})$ by the excess error $\mathcal{E}(g, \vec{f}) - \mathcal{R}_s$. For this purpose, we prove two facts which we state in Lemmas 15 and 16 and define the local error function of $t \in \mathbb{R}$ at $x \in X$ as

$$\text{err}_x(t) = \mathbb{E}_{y \sim Y} [\phi(yt)] = \phi(t)P(1|x) + \phi(-t)P(-1|x),$$

which is a twice differentiable, univariate convex function for every $x \in X$.

Lemma 15 *For almost every $x \in X$, the following hold*

- (i) $f_\phi(x)$ is a minimizer of the function $\text{err}_x(t)$, i.e., $f_\phi(x) = \arg \min_{t \in \mathbb{R}} \text{err}_x(t)$;
- (ii) if $T > \max\{|t|, \|f_\phi\|_\infty\}$, then

$$\frac{1}{2}q_1(T)(t - f_\phi(x))^2 \leq \text{err}_x(t) - \text{err}_x(f_\phi(x)) \leq \frac{1}{2}q_2(T)(t - f_\phi(x))^2.$$

- (iii) If $T \geq \{|t|, 3\|f_\phi\|_\infty\}$ there exists a constant $c_1 > 0$ such that

$$\text{err}_x(t) - \text{err}_x(f_\phi(x)) \geq c_1 \max\left\{q_1(T), \frac{1}{T}\right\} (t - f_\phi(x))^2.$$

Proof The first conclusion is a direct consequence of the fact

$$\mathcal{R}(f) = \int_X \text{err}_x(f(x)) d\rho_X(x).$$

Note that $(\text{err}_x)'(f_\phi(x)) = 0$ since $f_\phi(x)$ is a minimizer of $\text{err}_x(t)$. By a Taylor series expansion, there exists t_0 between t and $f_\phi(x)$ such that

$$\text{err}_x(t) - \text{err}_x(f_\phi(x)) = \frac{1}{2}(\text{err}_x)''(t_0)(t - f_\phi(x))^2.$$

Since $(\text{err}_x)''(t_0) = \phi''(t_0)P(1|x) + \phi''(-t_0)P(-1|x)$ and $|t_0| \leq T$ the following holds

$$q_1(T) \leq \phi''(t_0), \phi''(-t_0) \leq q_2(T).$$

It follows $q_1(T) \leq (\text{err}_x)''(t_0) \leq q_2(T)$ which proves (ii).

To show (iii), write

$$\text{err}_x(t) - \text{err}_x(f_\phi(x)) = \int_{f_\phi(x)}^t \int_{f_\phi(x)}^r (\text{err}_x)''(a) da dr.$$

Since $(\text{err}_x)''(a)$ is positive, if $t \geq 3\|f_\phi\|_\infty := 3M_\phi$, then

$$\text{err}_x(t) - \text{err}_x(f_\phi(x)) \geq \int_{2M_\phi}^t \int_{|f_\phi(x)|}^{2M_\phi} (\text{err}_x)''(a) \text{d}a \text{d}r \geq q_1(2M_\phi)M_\phi(|t| - 2M_\phi)$$

and, if $t \leq -3M_\phi$, then

$$\text{err}_x(t) - \text{err}_x(f_\phi(x)) \geq \int_{-2M_\phi}^t \int_{-|f_\phi(x)|}^{-2M_\phi} (\text{err}_x)''(a) \text{d}a \text{d}r \geq q_1(2M_\phi)M_\phi(|t| - 2M_\phi).$$

So, if $|t| > 3M_\phi$,

$$\text{err}_x(t) - \text{err}_x(f_\phi(x)) \geq q_1(2M_\phi)M_\phi(|t| - 2M_\phi) \geq \frac{3q_1(2M_\phi)M_\phi}{16T}(t - f_\phi(x))^2,$$

where we have used the facts $|t| - 2M_\phi \geq \frac{1}{4}|t - f_\phi(x)|$ and $|t - f_\phi(x)| \leq T + M_\phi \leq \frac{4}{3}T$. On the other hand, by (ii), if $|t| \leq 3M_\phi$

$$\text{err}_x(t) - \text{err}_x(f_\phi(x)) \geq \frac{1}{2}q_1(3M_\phi)(t - f_\phi(x))^2 \geq \frac{3q_1(3M_\phi)M_\phi}{2T}(t - f_\phi(x))^2.$$

Hence for all $|t| \leq T$,

$$\text{err}_x(t) - \text{err}_x(f_\phi(x)) \geq \frac{3q_1(3M_\phi)M_\phi}{16T}(t - f_\phi(x))^2.$$

Together with (ii), we obtain

$$\text{err}_x(t) - \text{err}_x(f_\phi(x)) \geq c_1 \max \left\{ q_1(T), \frac{1}{T} \right\} (t - f_\phi(x))^2$$

with $c_1 = \min \left\{ \frac{1}{2}, \frac{3q_1(3M_\phi)M_\phi}{16} \right\}$. ■

Lemma 16 *If $K \in C^2$, then there exists a constant $c_K > 0$ depending only on K such that*

$$|f(x) - f(u)| \leq c_K \|f\|_K |x - u|, \quad \forall f \in \mathcal{H}_K, \quad x, u \in X.$$

Proof It follows from the reproducing property that

$$|f(x) - f(u)| = |\langle f, K(x, \cdot) - K(u, \cdot) \rangle| \leq \|f\|_K \sqrt{K(x, x) - 2K(x, u) + K(u, u)}.$$

Denote $\nabla_1 K(x, u)$ as the gradient of $K(x, u)$ with respect to the first variable x . Since $K \in C^2$, we have

$$\begin{aligned} & K(x, x) - 2K(x, u) + K(u, u) \\ &= \int_0^1 (\nabla_1(K(u + t(x - u), x) - \nabla_1 K(u + t(x - u), y)) \cdot (x - u) \text{d}t \\ &\leq \int_0^1 |\nabla_1(K(u + t(x - u), x) - \nabla_1 K(u + t(x - u), y))| |x - u| \text{d}t \\ &\leq (c_K)^2 |x - u|^2 \end{aligned}$$

with

$$(c_K)^2 = \max \left\{ \left\| \frac{\partial^2 K}{\partial x^i \partial u^j} \right\|_\infty, i, j = 1, \dots, n \right\}.$$

Hence the conclusion is true. \blacksquare

Lemma 17 *Under the assumptions of Theorem 12, there exists a constant $c_2 > 0$ such that*

$$\mathcal{Q}(g, \vec{f}) \leq c_2 \left(r^2 s^2 + B_r (\mathcal{E}(g, \vec{f}) - \mathcal{R}_s) \right),$$

where B_r is defined as in Theorem 12 with $c_0 = \kappa \max \{3\|f_\phi\|_K, (1+D)\}$.

Proof For $(g, \vec{f}) \in \mathcal{F}_r$ and $u, x \in X$, we have

$$|g(u) + \vec{f}(x)(x-u)| \leq \kappa \|g\|_K + \kappa D \|\vec{f}\|_K \leq c_0 r.$$

Since $c_0 r \geq 3\kappa \|f_\phi\|_K \geq 3\|f_\phi\|_\infty$, by Lemma 15 (iii),

$$\begin{aligned} \mathcal{E}(g, \vec{f}) - R_s &= \int_X \int_X w(x-u) \left(\text{err}_x(g(u) + \vec{f}(x) \cdot (x-u)) - \text{err}_x(f_\phi(x)) \right) d\rho_X(x) d\rho_X(u) \\ &\geq \frac{c_1}{c_0} \frac{1}{B_r} \int_X \int_X w(x-u) \left(g(u) + \vec{f}(x) \cdot (x-u) - f_\phi(x) \right)^2 d\rho_X(x) d\rho_X(u), \end{aligned}$$

Denote

$$\begin{aligned} t_1 &= f(u) - f_\phi(u) + (\vec{f}(u) - \nabla f_\phi(u)) \cdot (x-u), \\ t_2 &= (f_\phi(u) - f_\phi(x) + \nabla f_\phi(u) \cdot (x-u)) + (\vec{f}(x) - \vec{f}(u)) \cdot (x-u). \end{aligned}$$

We have

$$\mathcal{Q}(g, \vec{f}) = \int_X \int_X (t_1)^2 d\rho_X(x) d\rho_X(u).$$

Note that

$$\left(f(u) + \vec{f}(x) \cdot (x-u) - f_\phi(x) \right)^2 = (t_1 + t_2)^2 \geq (t_1)^2 + 2t_1 t_2 \geq (t_1)^2 - 2|t_1||t_2|.$$

There holds

$$\frac{c_0}{c_1} B_r (\mathcal{E}(g, \vec{f}) - R_s) \geq \mathcal{Q}(g, \vec{f}) - 2 \int_X \int_X |t_1||t_2| d\rho_X(x) d\rho_X(u).$$

By the fact $\nabla f_\phi \in \mathcal{H}_K^n$ and Lemma 16, we have

$$|t_2| \leq c_K (\|\nabla f_\phi\|_K + \|\vec{f}\|_K) |x-u|^2 \leq c_K (\|\nabla f_\phi\|_K + r) |x-u|^2.$$

Together with the assumption $p(x) \leq c_\rho$ we obtain

$$\begin{aligned} \int_X \int_X |t_1||t_2| d\rho_X(x) d\rho_X(u) &\leq \sqrt{\mathcal{Q}(g, \vec{f})} \left(\int_X \int_X |t_2|^2 d\rho_X(x) d\rho_X(u) \right)^{1/2} \\ &\leq c_K (\|\nabla f_\phi\|_K + r) \sqrt{\mathcal{Q}(g, \vec{f})} \left(c_\rho \int_X \int_{\mathbb{R}^n} w(x-u) |x-u|^4 dx d\rho_X(u) \right)^{1/2} \\ &\leq c_K (\|\nabla f_\phi\|_K + r) \sqrt{c_\rho \tilde{N}_4 s} \sqrt{\mathcal{Q}(g, \vec{f})}. \end{aligned}$$

Combining the above arguments we obtain

$$\mathcal{Q}(g, \vec{f}) - 2c_K (\|\nabla f_\phi\|_K + r) \sqrt{c_\rho \tilde{N}_4 s} \sqrt{\mathcal{Q}(g, \vec{f})} \leq \frac{1}{c_1} \min \left\{ \frac{1}{q_1(c_0 r)}, c_0 r \right\} (\mathcal{E}(f, \vec{f}) - \mathcal{R}_s).$$

Solving this inequality gives

$$\sqrt{\mathcal{Q}(g, \vec{f})} \leq 2c_K (\|\nabla f_\phi\|_K + r) \sqrt{c_\rho \tilde{N}_4 s} + \sqrt{\frac{1}{c_1} \min \left\{ \frac{1}{q_1(c_0 r)}, c_0 r \right\} (\mathcal{E}(f, \vec{f}) - \mathcal{R}_s)}.$$

This implies the conclusion with $c_2 = 2 \max \left\{ 2(c_K)^2 (\|\nabla f_\phi\|_K + 1)^2 c_\rho \tilde{N}_4, \frac{c_0}{c_1} \right\}$. \blacksquare

Proof of Theorem 12. Write

$$\|g - f_\phi\|_{L^2_{\rho_X}}^2 = \int_{X \setminus X_s} (g(x) - f_\phi(x))^2 d\rho_X(x) + \int_{X_s} (g(x) - f_\phi(x))^2 d\rho_X(x). \quad (18)$$

We have

$$\rho_X(X \setminus X_s) \leq c_\rho s + (1 + c_\rho) c_\rho |X| s^\theta \leq (c_\rho + (1 + c_\rho) c_\rho |X|) s^\theta,$$

where $|X|$ is the Lebesgue measure of X . So the first term on the right of (18) is bounded by

$$\kappa^2 (r + \|f_\phi\|_K)^2 (c_\rho + (1 + c_\rho) c_\rho |X|) s^\theta.$$

By Lemmas 14 and 17, the second term on the right of (18) is bounded by

$$\frac{s^{2-\theta}}{N_0} c_2 \left(r^2 s^2 + B_r (\mathcal{E}(f, \vec{f}) - \mathcal{R}_s) \right)$$

Combing these two estimates finishes the proof of the first claim with

$$C_0 = \kappa^2 (1 + \|f_\phi\|_K)^2 (c_\rho + (1 + c_\rho) c_\rho |X|) + \frac{c_2}{N_0}.$$

Similarly, we can show the second claim with

$$C_1 = \kappa^2 (1 + \|\nabla f_\phi\|_K)^2 (c_\rho + (1 + c_\rho) c_\rho |X|) + \frac{nc_2}{N_2}. \quad \blacksquare$$

In order to apply Theorem 12 to $(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}})$, we need a bound on $\|g_{\mathbf{z}}\|_K^2 + \|\vec{f}_{\mathbf{z}}\|_K^2$. We first state a rough bound.

Lemma 18 *For every $s > 0$ and $\lambda > 0$, $\|g_{\mathbf{z}}\|_K^2 + \|\vec{f}_{\mathbf{z}}\|_K^2 \leq \frac{2\phi(0)}{\lambda s^{n+2}}$.*

Proof The conclusion follows from the fact

$$\frac{\lambda}{2} \left(\|g_{\mathbf{z}}\|_K^2 + \|\vec{f}_{\mathbf{z}}\|_K^2 \right) \leq \mathcal{E}_{\mathbf{z}}(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}}) + \frac{\lambda}{2} \left(\|g_{\mathbf{z}}\|_K^2 + \|\vec{f}_{\mathbf{z}}\|_K^2 \right) \leq \mathcal{E}_{\mathbf{z}}(0, \vec{0}) + 0 = \frac{\phi(0)}{s^{n+2}}. \quad \blacksquare$$

Remark 19 *Using this quantity the bound in Theorem 12 is at least of order $\mathcal{O}(\frac{1}{\lambda s^{n+2-\theta}})$ which tends to ∞ as $s \rightarrow 0$ and $\lambda \rightarrow 0$. So a sharper bound is needed. We will obtain such a bound in Section 3.3.*

3.2 Bounding the excess error

In this section, we bound the quantity $\mathcal{E}(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}}) - \mathcal{R}_s$. Let

$$(g_\lambda, \vec{f}_\lambda) = \arg \min_{(g, \vec{f}) \in \mathcal{H}_K^{n+1}} \left\{ \mathcal{E}(g, \vec{f}) + \frac{\lambda}{2} (\|g\|_K^2 + \|\vec{f}\|_K^2) \right\}.$$

Theorem 20 *If $(f_\phi, \nabla f_\phi) \in \mathcal{H}_K^{n+1}$, $(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}})$ and $(g_\lambda, \vec{f}_\lambda)$ are in \mathcal{F}_r for some $r \geq 1$, then with confidence $1 - \delta$*

$$\mathcal{E}(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}}) - \mathcal{R}_s \leq C_2 \left(\frac{L_r r + M_r \log \frac{2}{\delta}}{\sqrt{m} s^{n+2}} + s^2 + \lambda \right),$$

where $C_2 > 0$ is a constant depending on ϕ and ρ but not on r, s and λ .

By a standard decomposition procedure, we have the following result.

Proposition 21 *The following hold*

$$\mathcal{E}(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}}) - \mathcal{R}_s + \frac{\lambda}{2} (\|g_{\mathbf{z}}\|_K^2 + \|\vec{f}_{\mathbf{z}}\|_K^2) \leq \mathcal{S}(\mathbf{z}) + \mathcal{A}(\lambda)$$

where

$$\mathcal{S}(\mathbf{z}) = \left(\mathcal{E}(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}}) \right) + \left(\mathcal{E}_{\mathbf{z}}(g_\lambda, \vec{f}_\lambda) - \mathcal{E}(g_\lambda, \vec{f}_\lambda) \right)$$

and

$$\mathcal{A}(\lambda) = \inf_{(g, \vec{f}) \in \mathcal{H}_K^{n+1}} \left\{ \mathcal{E}(g, \vec{f}) - \mathcal{R}_s + \frac{\lambda}{2} (\|g\|_K^2 + \|\vec{f}\|_K^2) \right\}.$$

The quantity $\mathcal{S}(\mathbf{z})$ is called the sample error and can be bound by controlling the quantity

$$S(\mathbf{z}, r) := \sup_{(g, \vec{f}) \in \mathcal{F}_r} |\mathcal{E}_{\mathbf{z}}(g, \vec{f}) - \mathcal{E}(g, \vec{f})|.$$

In fact, if both $(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}})$ and $(g_\lambda, \vec{f}_\lambda)$ are in \mathcal{F}_r for some $r > 0$, then

$$\mathcal{S}(\mathbf{z}) \leq 2S(\mathbf{z}, r). \tag{19}$$

Again $\mathcal{E}_{\mathbf{z}}(g, \vec{f})$ and $\mathcal{E}(g, \vec{f})$ are not the empirical and expected means of a random variable. We will use McDiarmid's inequality (McDiarmid, 1989) to bound $S(\mathbf{z}, r)$.

Lemma 22 *For every $r > 0$*

$$\text{Prob} \{ |S(\mathbf{z}, r) - \mathbb{E} S(\mathbf{z}, r)| > \varepsilon \} \leq 2 \exp \left(-\frac{m \varepsilon^2 s^{2(n+2)}}{2M_r^2} \right).$$

Proof Denote by \mathbf{z}'_i the sample which coincides with \mathbf{z} except for the i -th pair (x_i, y_i) replaced by (x'_i, y'_i) . It is easy to verify that

$$\begin{aligned} S(\mathbf{z}, r) - S(\mathbf{z}'_i, r) &= \sup_{(g, \vec{f}) \in \mathcal{F}_r} (\mathcal{E}_{\mathbf{z}}(g, \vec{f}) - \mathcal{E}(g, \vec{f})) - \sup_{(g, \vec{f}) \in \mathcal{F}_r} (\mathcal{E}_{\mathbf{z}'_i}(g, \vec{f}) - \mathcal{E}(g, \vec{f})) \\ &\leq \sup_{(g, \vec{f}) \in \mathcal{F}_r} (\mathcal{E}_{\mathbf{z}}(g, \vec{f}) - \mathcal{E}_{\mathbf{z}'_i}(g, \vec{f})) \leq \frac{2m-1}{m^2} \frac{M_r}{s^{n+2}}. \end{aligned}$$

Interchanging the roles of \mathbf{z} and \mathbf{z}'_i gives $|S(\mathbf{z}, r) - S(\mathbf{z}'_i, r)| \leq \frac{2M_r}{ms^{n+2}}$. By McDiarmid's inequality we obtain the desired estimate. \blacksquare

Lemma 23 *For every $r > 0$*

$$\mathbb{E} S(\mathbf{z}, r) \leq \frac{8L_r(\kappa(1+2D)r + \phi(0))}{s^{n+2}\sqrt{m}} + \frac{2M_r}{ms^{n+2}}.$$

In order to prove this lemma, we need Rademacher complexities. We refer to Koltchinskii and Panchenko (2000) and van der Vaart and Wellner (1996) for definitions and properties.

Proof Denote $\xi(x, y, u) = w(x-u)\phi(y(g(u) + \vec{f}(x) \cdot (x-u)))$ for simplicity. Then $\mathcal{E}(g, \vec{f}) = \mathbb{E}_u \mathbb{E}_{(x,y)} \xi(x, y, u)$ and $\mathcal{E}_{\mathbf{z}}(g, \vec{f}) = \sum_{i,j=1}^m \xi(x_i, y_i, x_j)$. One can easily check that

$$\begin{aligned} S(\mathbf{z}, r) &\leq \sup_{(g, \vec{f}) \in \mathcal{F}_r} \left| \mathcal{E}(g, \vec{f}) - \frac{1}{m} \sum_{j=1}^m \mathbb{E}_{(x,y)} \xi(x, y, x_j) \right| + \sup_{(g, \vec{f}) \in \mathcal{F}_r} \left| \frac{1}{m} \sum_{j=1}^m \mathbb{E}_{(x,y)} \xi(x, y, x_j) - \mathcal{E}_{\mathbf{z}}(g, \vec{f}) \right| \\ &\leq \mathbb{E}_{(x,y)} \sup_{(g, \vec{f}) \in \mathcal{F}_r} \left| \mathbb{E}_u \xi(x, y, u) - \frac{1}{m} \sum_{i=1}^m \xi(x, y, x_i) \right| \\ &\quad + \frac{1}{m} \sum_{j=1}^m \sup_{(g, \vec{f}) \in \mathcal{F}_r} \sup_{u \in X} \left| \mathbb{E}_{(x,y)} \xi(x, y, u) - \frac{1}{m-1} \sum_{\substack{i=1 \\ i \neq j}}^m \xi(x_i, y_i, u) \right| \\ &\quad + \frac{1}{m} \sum_{j=1}^m \left(\frac{1}{m} \xi(x_j, y_j, x_j) + \frac{1}{m(m-1)} \sum_{\substack{i=1 \\ i \neq j}}^m \xi(x_i, y_i, x_j) \right) \\ &:= S_1 + S_2 + S_3. \end{aligned}$$

Let $\varepsilon_i, i = 1, \dots, m$ be independent Rademacher variables. Denote

$$G_{(x,y)} = \left\{ h(u) = y(g(u) + \vec{f}(x) \cdot (x-u)) : (g, \vec{f}) \in \mathcal{F}_r \right\}$$

for every $(x, y) \in Z$. For S_1 , by using the properties of Rademacher complexities, we have

$$\begin{aligned}
 \mathbb{E} S_1(\mathbf{z}) &= \mathbb{E}_{(x,y)} \mathbb{E} \sup_{h \in G_{x,y}} \left| \mathbb{E}_u [w(x-u)\phi(h(u))] - \frac{1}{m} \sum_{j=1}^m w(x-x_j)\phi(h(x_j)) \right| \\
 &\leq 2 \sup_{(x,y) \in Z} \mathbb{E} \sup_{h \in G_{(x,y)}} \left| \frac{1}{m} \sum_{j=1}^m \varepsilon_j w(x-x_j)\phi(h(x_j)) \right| \\
 &\leq \frac{4}{s^{n+2}} \sup_{(x,y) \in Z} \mathbb{E} \sup_{h \in G_{(x,y)}} \left| \frac{1}{m} \sum_{j=1}^m \varepsilon_j \phi(h(x_j)) \right| \\
 &\leq \frac{4L_r}{s^{n+2}} \left(\sup_{(x,y) \in Z} \mathbb{E} \sup_{h \in G_{(x,y)}} \left| \frac{1}{m} \sum_{j=1}^m \varepsilon_j h(x_j) \right| + \frac{\phi(0)}{\sqrt{m}} \right) \\
 &\leq \frac{4L_r}{s^{n+2}} \left(\mathbb{E} \sup_{\|g\|_K^2 \leq r^2} \left| \sum_{j=1}^m \varepsilon_j g(x_j) \right| + 2\kappa r \sup_{x \in X} \mathbb{E} \left| \sum_{j=1}^m \varepsilon_j \|x-x_j\| \right| + \frac{\phi(0)}{\sqrt{m}} \right) \\
 &\leq \frac{4L_r(\kappa(1+2D)r + \phi(0))}{s^{n+2}\sqrt{m}}.
 \end{aligned}$$

Similarly, we can verify

$$\mathbb{E} S_2(\mathbf{z}) \leq \frac{4L_r(\kappa(1+2D)r + \phi(0))}{s^{n+2}\sqrt{m-1}}.$$

Obviously $S_3 \leq \frac{2M_r}{ms^{n+2}}$. Combining the estimates for S_1 , S_2 , and S_3 completes the proof. \blacksquare

Proposition 24 *Assume $r > 1$. There exists a constant $c_2 > 0$ such that with confidence at least $1 - \delta$*

$$\mathcal{S}(\mathbf{z}) \leq c_3 \frac{L_r r + M_r \log \frac{2}{\delta}}{\sqrt{m} s^{n+2}}.$$

Proof The result is a direct application of inequality (19) and Lemmas 22 and 23. \blacksquare

We now bound the approximation error $\mathcal{A}(\lambda)$.

Proposition 25 *If $(f_\phi, \nabla f_\phi) \in \mathcal{H}_K^{n+1}$, then $\mathcal{A}(\lambda) \leq c_4(s^2 + \lambda)$ for some $c_4 > 0$.*

Proof By the definition of $\mathcal{A}(\lambda)$ and the fact that $(f_\phi, \nabla f_\phi) \in \mathcal{H}_K^{n+1}$

$$\mathcal{A}(\lambda) \leq \mathcal{E}(f_\phi, \nabla f_\phi) - \mathcal{R}_s + \frac{\lambda}{2} (\|f_\phi\|_K^2 + \|\nabla f_\phi\|_K^2).$$

By Lemma 15 (ii), we have

$$\begin{aligned}
 \mathcal{E}(f_\phi, \nabla f_\phi) - \mathcal{R}_s &= \int_X \int_X w(x-u) \left(\text{err}_x(f_\phi(u) + \nabla f_\phi(x) \cdot (x-u)) - \text{err}_x(f_\phi(x)) \right) d\rho_X(u) d\rho_X(x) \\
 &\leq q_2(\tilde{M}_\phi) \int_X \int_X w(x-u) \left(f_\phi(u) - f_\phi(x) + \nabla f_\phi(x) \cdot (x-u) \right)^2 d\rho_X(u) d\rho_X(x) \\
 &\leq q_2(\tilde{M}_\phi) c_\rho \int_X \int_X w(x-u) |x-u|^4 du d\rho_X(x) \leq q_2(\tilde{M}_\phi) c_\rho \tilde{N}_4 s^2,
 \end{aligned}$$

where $\tilde{M}_\phi = \kappa \|f_\phi\|_K + \kappa D \|\nabla f_\phi\|_K$. Taking

$$c_4 = \max\{q_2(\tilde{M}_\phi)c_\rho\tilde{N}_4, \frac{1}{2}(\|f_\phi\|_K^2 + \|\nabla f_\phi\|_K^2)\},$$

the result follows. \blacksquare

Theorem 20 follows directly from Propositions 21, 24 and 25.

3.3 Proof of Theorem 10

We will use Theorems 12 and 20 to prove Theorem 10.

Notice that both theorems need a bound r so that $(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}})$ and $(g_\lambda, \vec{f}_\lambda)$ are in \mathcal{F}_r . In Lemma 18 we have shown

$$\|g_{\mathbf{z}}\|_K^2 + \|\vec{f}_{\mathbf{z}}\|_K^2 \leq \frac{2\phi(0)}{\lambda s^{n+2}}.$$

Similarly we can show $\|g_\lambda\|_K^2 + \|\vec{f}_\lambda\|_K^2$ is also bounded by $\frac{2\phi(0)}{\lambda s^{n+2}}$. So $\sqrt{\frac{2\phi(0)}{\lambda s^{n+2}}}$ is a universal bound for r such that $(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}})$ and $(g_\lambda, \vec{f}_\lambda)$ are in \mathcal{F}_r . However, this bound is not sharp enough to be useful for Theorem 12 (see Remark 19).

A sharper bound will be given below. This bound also improves the sample error estimate and the estimate in Theorem 20.

Lemma 26 *Under the assumptions of Theorem 10*

$$\|g_\lambda\|_K^2 + \|\vec{f}_\lambda\|_K^2 \leq 2c_4 \left(\frac{s^2}{\lambda} + 1 \right).$$

Proof Since $\mathcal{E}(g, \vec{f}) - \mathcal{R}_s$ is non-negative for all pairs (g, \vec{f}) , we have

$$\frac{\lambda}{2}(\|g_\lambda\|_K^2 + \|\vec{f}_\lambda\|_K^2) \leq \mathcal{E}(g_\lambda, \vec{f}_\lambda) - \mathcal{R}_s + \frac{\lambda}{2}(\|g_\lambda\|_K^2 + \|\vec{f}_\lambda\|_K^2) = \mathcal{A}(\lambda).$$

This in conjunction with Proposition 25 implies the conclusion. \blacksquare

Lemma 27 *Under the assumptions of Theorem 10 with confidence at least $1 - \delta$*

$$\|g_{\mathbf{z}}\|_K^2 + \|\vec{f}_{\mathbf{z}}\|_K^2 \leq c_5 \left\{ 1 + \frac{s^2}{\lambda} + \left(\frac{L_{\lambda,s}}{\sqrt{\lambda s^{n+2}}} + M_{\lambda,s} \log \frac{2}{\delta} \right) \frac{1}{\sqrt{m\lambda s^{n+2}}} \right\}$$

for some $c_5 > 0$.

Proof By the fact $\mathcal{E}(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}}) - \mathcal{R}_s > 0$ and Proposition 21 we have

$$\frac{\lambda}{2} \left(\|g_{\mathbf{z}}\|_K^2 + \|\vec{f}_{\mathbf{z}}\|_K^2 \right) \leq \mathcal{S}(\mathbf{z}) + \mathcal{A}(\lambda).$$

Since both $(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}})$ and $(g_\lambda, \vec{f}_\lambda)$ are in $\mathcal{F}_{\sqrt{2\phi(0)/\lambda s^{n+2}}}$, we apply Proposition 24 to get with probability at least $1 - \delta$

$$\mathcal{S}(\mathbf{z}) \leq c_3 \left(L_{\lambda,s} \sqrt{\frac{2\phi(0)}{\lambda s^{n+2}}} + M_{\lambda,s} \log \frac{2}{\delta} \right) \frac{1}{\sqrt{m\lambda s^{n+2}}}.$$

Together with Proposition 25, we obtain the desired estimate with $c_5 = 2 \max\{c_3, c_4\}$. ■

We now prove Theorem 10.

Proof of Theorem 10. By Theorems 12 and 20 we have with probability at least $1 - \delta$ both $\|g_{\mathbf{z}} - f_{\phi}\|_{L^2_{\rho_X}}$ and $\|\vec{f}_{\mathbf{z}} - \nabla f_{\phi}\|_{L^2_{\rho_X}}$ are bounded by

$$\max\{C_0, C_1\} \left\{ r^2 s^{\theta} + C_2 B_r \left(\frac{L_r r + M_r \log \frac{2}{\delta}}{\sqrt{m} s^{n+2}} + s^2 + \lambda \right) s^{-\theta} \right\}, \quad (20)$$

if both $(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}})$ and $(g_{\lambda}, \vec{f}_{\lambda})$ are in \mathcal{F}_r for some $r > 1$. By Lemmas 26 and 27 we can state that both $\{(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}}) \in \mathcal{F}_r\}$ and $\{(g_{\lambda}, \vec{f}_{\lambda}) \in \mathcal{F}_r\}$ with probability at least $1 - \delta$ if

$$r^2 = \max(c_4, c_5, 1) \left\{ 1 + \frac{s^2}{\lambda} + \left(\frac{L_{\lambda, s}}{\sqrt{\lambda} s^{n+2}} + M_{\lambda, s} \log \frac{2}{\delta} \right) \frac{1}{\sqrt{m} \lambda s^{n+2}} \right\}.$$

Substituting the above r into (20) gives us the desired bound with confidence at least $1 - 2\delta$. ■

4. Simulated data and gene expression data

In this section we apply the gradient learning algorithm (9) to the problem of estimating a classification function and simultaneously selecting relevant variables and measuring their covariance. The idea is to rank the importance of variables according to the norm of their partial derivatives $\|\frac{\partial f_{\phi}}{\partial x^{\ell}}\|$, since a small norm implies small changes of the classification function with respect to this variable. By our error analysis, we expect $\vec{f}_{\mathbf{z}} \approx \nabla f_{\phi}$. So we shall use the norms of the components of $\vec{f}_{\mathbf{z}}$ to rank the variables.

Definition 28 *The relative magnitude of the norm for the variables is defined as*

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

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

Definition 29 *The empirical gradient matrix (EGM), $F_{\mathbf{z}}$, is the $n \times m$ matrix whose columns are $\vec{f}_{\mathbf{z}}(x_j)$ with $j = 1, \dots, m$. The empirical covariance matrix (ECM), $\Xi_{\mathbf{z}}$, is the $n \times n$ matrix of inner products of the directional derivative of two coordinates*

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

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 two datasets are artificial ones which we use to illustrate the procedure. The third is a cancer classification problem that has been well studied and serves as further confirmation of the utility of the method.

Apology 1 *The experimental validation of this procedure and how to use it to analyze high dimensional expression data is at least as important as the formulation and analysis of the method. However, including a proper discussion of the experimental aspects of the method would result in a paper that would be very onerous to read. These aspects will be addressed in Goh et al. (2006).*

4.1 Linearly separable simulation

Linearly separable data is drawn from two classes in an $n = 80$ dimensional space. Samples from class +1 were drawn from

$$\begin{aligned} x^j &\sim \mathcal{N}(1.5, 1), \text{ for } j = 1, \dots, 10, \\ x^j &\sim \mathcal{N}(-3, 1), \text{ for } j = 11, \dots, 20, \\ x^j &\sim \mathcal{N}(0, .2), \text{ for } j = 21, \dots, 80, \end{aligned}$$

where $\mathcal{N}(\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ . Samples from class -1 were drawn from

$$\begin{aligned} x^j &\sim \mathcal{N}(1.5, 1), \text{ for } j = 41, \dots, 50, \\ x^j &\sim \mathcal{N}(-3, 1), \text{ for } j = 51, \dots, 60, \\ x^j &\sim \mathcal{N}(0, .2), \text{ for } j = 1, \dots, 40, 61, \dots, 80. \end{aligned}$$

Drawing twenty samples from the two respective classes results in a design matrix \mathbf{x} that is 80×40 where the first twenty samples belong to class +1 and the remaining to class -1. A draw of this matrix is displayed in figure (1a). In figure (1d) we display the conditional likelihoods obtained by the classification function on the training data. A leave-one-out analysis yields similar results.

In figure (1b) we plot the norm of each component of the estimate of the gradient, $\{\|(\vec{f}_{\mathbf{z}})_\ell\|_K\}_{\ell=1}^{80}$. 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, $\ell = 1, \dots, 20, 41, \dots, 60$. Figure (1c) displays the empirical covariance matrix. The blocking structure of this matrix indicates the covariance of coordinates.

4.2 Nonlinearly separable simulation

Data is drawn from two classes in an $n = 200$ dimensional space that are nonlinearly separable in the first two dimensions. Samples from class +1 were drawn from

$$\begin{aligned} (x^1, x^2) &= (r \sin(\theta), r \cos(\theta)), \text{ where } r \sim U[0, 1] \text{ and } \theta \sim U[0, 2\pi], \\ x^j &\sim \mathcal{N}(0.0, .2), \text{ for } j = 3, \dots, 200, \end{aligned}$$

where $U[a, b]$ is the uniform distribution with support on the interval $[a, b]$. Samples from class -1 were drawn from

$$\begin{aligned} (x^1, x^2) &= (r \sin(\theta), r \cos(\theta)), \text{ where } r \sim U[2, 3] \text{ and } \theta \sim U[0, 2\pi], \\ x^j &\sim \mathcal{N}(0.0, .2), \text{ for } j = 3, \dots, 200. \end{aligned}$$

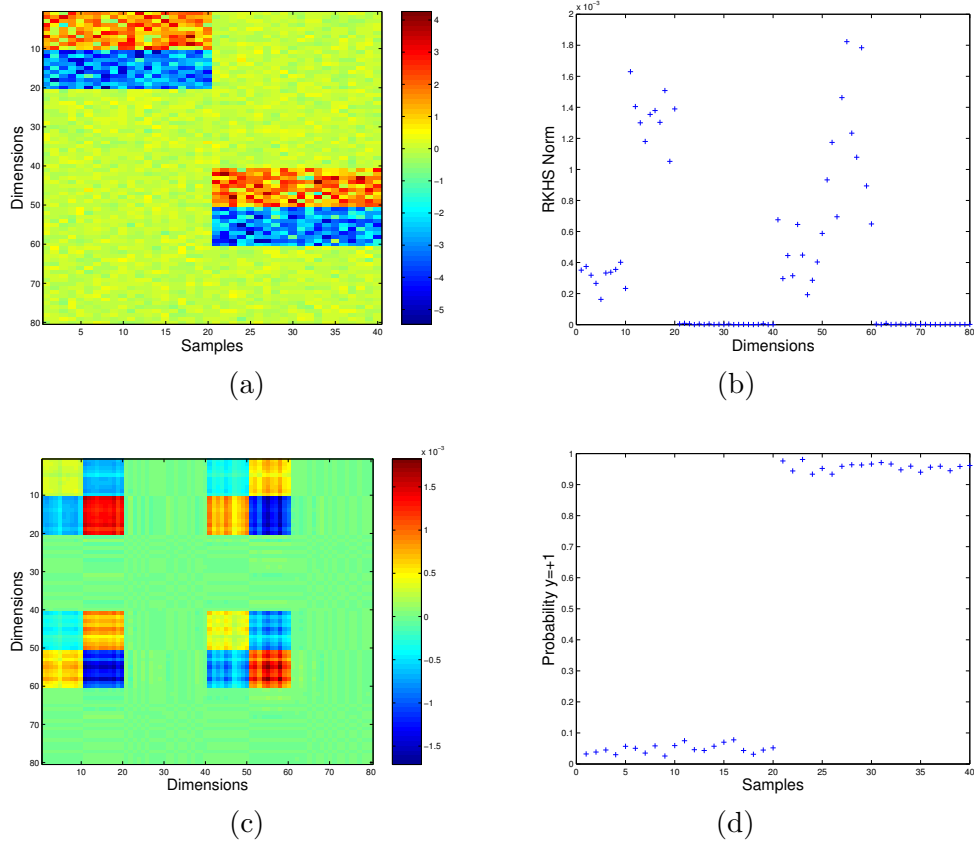


Figure 1: a) The data matrix \mathbf{x} where each sample corresponds to a column and the first twenty samples correspond to class +1 and the second twenty to class -1, b) the RKHS norm for each dimension, c) the empirical covariance matrix, d) the predicted class probabilities on the training data.

Note that the data can be separated by a circle in the first two dimensions.

Drawing thirty samples from the two respective classes results in a design matrix \mathbf{x} that is 200×60 where the first thirty samples belong to class +1 and the remaining to class -1. A draw of the first two dimensions of the data is displayed in figure (2a). Since a linear function cannot accurately classify the data we used a Gaussian kernel

$$K(u, v) = e^{-|u-v|/2\sigma^2},$$

where σ was set to the median pairwise distances between points. In figure (2c,d) we plot the norm of each component of the estimate of the gradient. The first two coordinates are the only ones with nonzero norm as expected. In figure (2b) we plot the ECM. The blocking structure of the ECM indicates the covariance of the first two coordinates. In figure (2e) we display the conditional likelihoods obtained by the classification function on the training data without any feature selection. The classification accuracy improves when we rerun our algorithm using only the dimensions with nonzero norms (2f). The classification results are comparable to what would be obtained by using regularized logistic regression.

4.3 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 drug 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.

4.4 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. In the same spirit as recursive feature elimination (RFE) we iteratively run our procedure on the training data and remove all variables except for the \mathcal{S} with the largest norm, s_ℓ^ϕ . In Table 1 we report test errors for various values of \mathcal{S} that result from the following procedure:

1. given training data \mathbf{z}_{7129} and test data \mathbf{tz}_{7129} compute the number of errors on the test data $\text{ter}_{7129}(\mathbf{tz}_{7129}) = |\text{sign}[g_{\mathbf{z}_{7129}}(\mathbf{tz}_{7129})] \neq ty|$ and the vector of norms $\{s_\ell^\phi\}_{\ell=1}^{7129}$

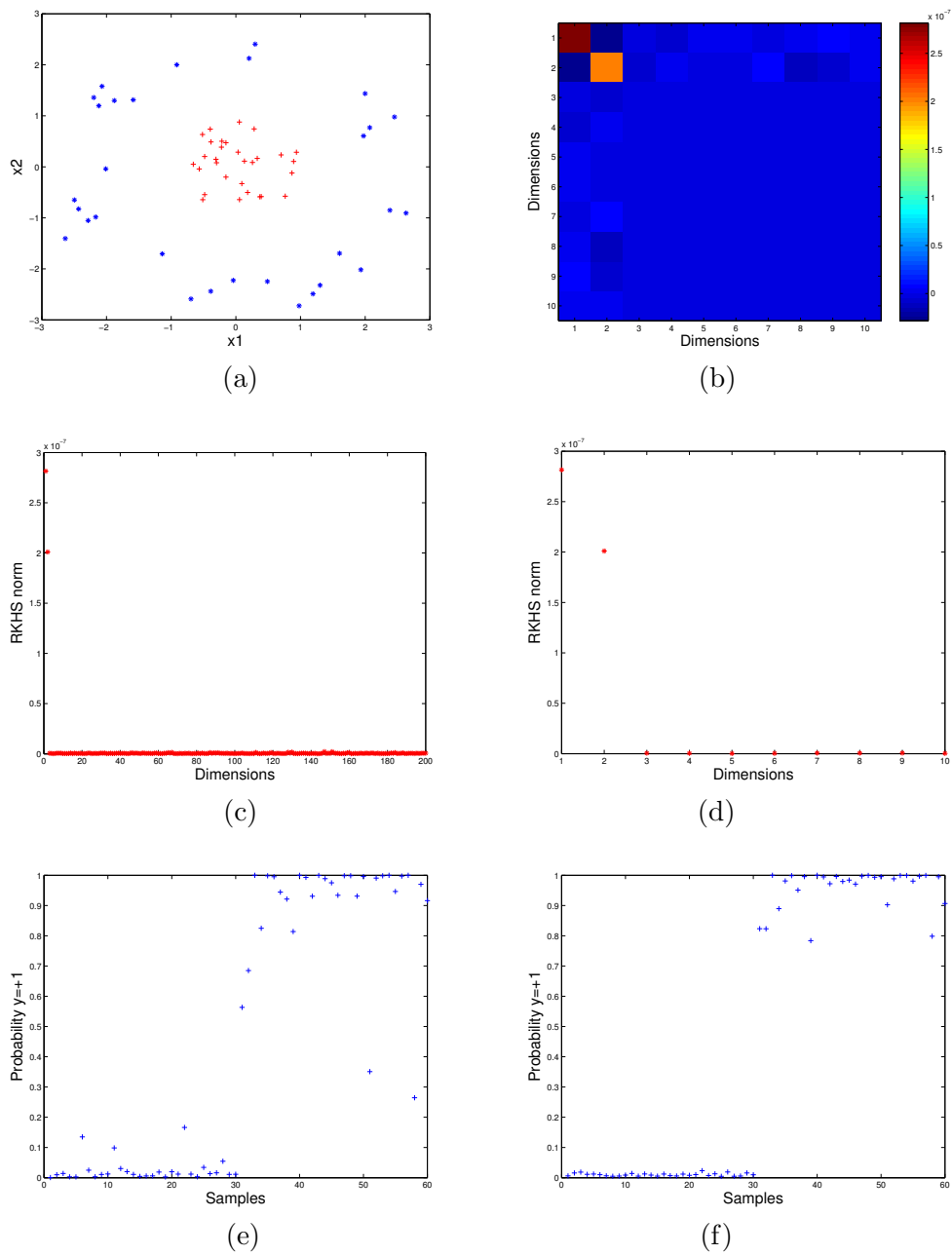


Figure 2: a) The first two dimensions of the data matrix class +1 is in red and class -1 is in blue, b) the empirical covariance matrix for the first 10 dimensions, c) the RKHS norm for all dimensions, d) the RKHS norm for the first 10 dimensions, e) the predicted class probabilities on the training data with no feature selection, f) the predicted class probabilities on the training data with feature selection.

2. for $\mathcal{S} = 3000, 1000, 500, 400, 300, 200, 100, 50$ repeat steps 3,4
3. project the test and training data into the dimensions corresponding to the top \mathcal{S} values of $\{s_\ell^\phi\} : \mathbf{z}_\mathcal{S}$ and $\mathbf{tz}_\mathcal{S}$
4. given the training data $\mathbf{z}_\mathcal{S}$ and test data $\mathbf{tz}_\mathcal{S}$ compute the number of errors on the test data $\text{ter}_\mathcal{S}(\mathbf{tz}_\mathcal{S}) = |\text{sign}[g_{\mathbf{z}_\mathcal{S}}(\mathbf{tz}_\mathcal{S})] \neq ty|$ and the vector of norms $\{s_\ell^\phi\}$.

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)	50	100	200	300	400	500	1,000	3,000	7,129
test errors	2	1	1	1	1	1	1	1	2

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 (3a-d) we plot the relative magnitude sequence s_ℓ^ϕ for the genes. On this dataset the decay in the ranked scores $s_{(\ell)}^\phi$ 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 (3a-d) displays the relative decay of $s_{(\ell)}^\phi$ and $s_{(\ell)}^F$ over various numbers of dimensions. In all plots it is apparent that the decay rate of $s_{(\ell)}^\phi$ 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 or the solution of regularized linear logistic regression 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.

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. In figure (4) we plot the ECM for the 50 dimensions that resulted from the iterative procedure outlined above. This matrix indicates how the dimensions covary and can be used to construct clusters of genes.

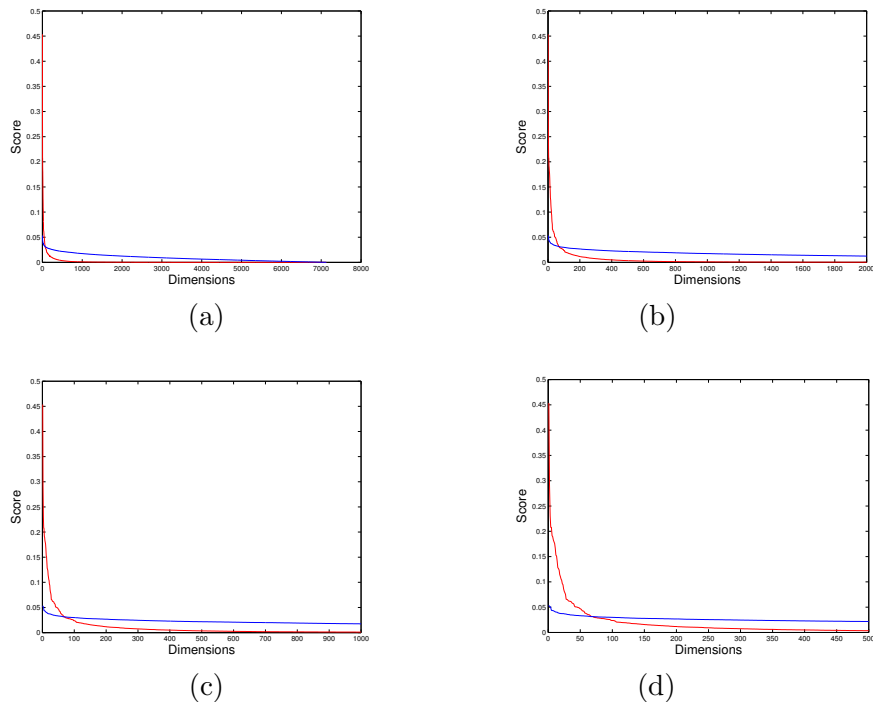


Figure 3: The decay of $s_{(\ell)}^{\phi}$ (red) and $s_{(\ell)}^F$ (blue) over: a) all the genes/dimensions, b) the top 3000 genes/dimensions, c) the top 1000 genes/dimensions, d) the top 500 genes/dimensions.

5. Discussion

We introduce an algorithm that learns a classification function and its gradient from sample data in the binary regression framework. The relevance of this method for variable selection is motivated. An error analysis is given for the convergence of the estimated classification function and gradient to the true ones respectively. This method also places the problem of variable selection into the powerful framework of Tikhonov regularization. There are many extensions and refinements and open questions regarding this method which we discuss below:

1. Accuracy of classification function: It seems intuitive that the classification function obtained by our method should be strictly worse than that obtained by standard regularized logistic regression. This is simply a corollary of very useful dictum proposed by Vladimir Vapnik (Vapnik, 1998), “When solving a given problem, try to avoid solving a more general problem as an intermediate step.” Although we strongly expect our classification function to be less accurate than that provided by regularized logistic regression we need to do more empirical work to confirm this.
2. Logistic regression models: An alternative optimization problem was proposed in Mukherjee and Zhou (2005) for estimating the gradient $\vec{f}_{\mathbf{z}}$ in the binary regression

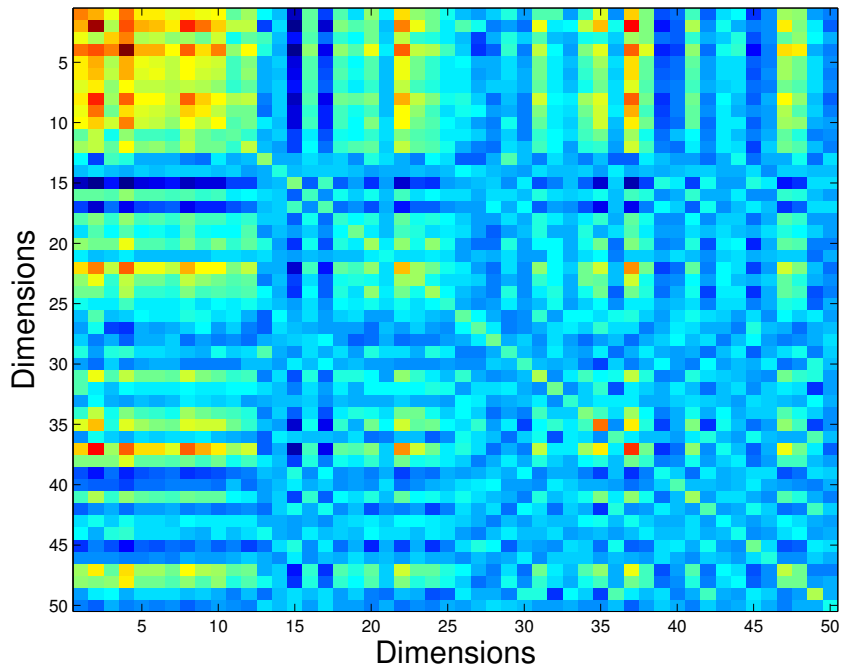


Figure 4: The ECM for the top 50 dimensions.

problem

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

This optimization problem does not follow from the Taylor expansion since in general y_j need not be close to $f_\phi(x_j)$, only the signs of the two functions need agree. This formulation does have an interesting interpretation for variable selection in that variables that are relevant in the classification problem will have a large gradient norms and those not relevant will have norms near zero. In practice, for large values of λ the gradient estimates of the above formulation will be similar to those given by the optimization in (9).

3. 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 (Liang et al., 2006). The Bayesian RB framework can be extended to develop a proper probability model for the gradient learning problem. The optimization procedure (9) 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.
4. Intrinsic dimension: In Theorem 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.

5. 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 $u \gg m$. 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 minimizing over $(g, \vec{f}) \in \mathcal{H}_K^{n+1}$

$$\left\{ \mathcal{E}_{\mathbf{z}}(g, \vec{f}) + \frac{\mu}{(m+u)^2} \sum_{i,j=1}^{m+u} W_{i,j} |f(x_i) - f(x_j)|_{\ell^2(\mathbb{R}^n)}^2 + \lambda \|\vec{f}\|_K^2 \right\}, \quad (21)$$

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

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. *Machine Learning*, 56(1-3):209–239, 2004.
- 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.
- T. Evgeniou, M. Pontil, and T. Poggio. Regularization Networks and Support Vector Machines. *Advances in Computational Mathematics*, 13:1–50, 2000.
- L. Goh, Q. Wu, S. Mukherjee, and T. Furey. Discovering co-oexpressed genes and pathways from gene expression data using coordinate covariance estimates, 2006. in preparation.
- 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.

- T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001. HAS t 01:1 1.Ex.
- V.I. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In J. Wellner E. Giné, D. Mason, editor, *High Dimensional Probability II*, pages 443–459, 2000.
- 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.
- F. Liang, S. Mukherjee, and M. West. Understanding the use of unlabelled data in predictive modeling. *Statistical Science*, 2006. accepted.
- C. McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics*, volume 141, pages 148–188. LMS Lecture Notes Series, 1989.
- C.A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17:177–204, 2005.
- S. Mukherjee and D.X. Zhou. Learning coordinate covariances via gradients. Technical report, 05-11, Institute of Statistics and Decision Sciences, Duke University, 2005.
- 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.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *J Royal Stat Soc B*, 58(1): 267–288, 1996.
- A. van der Vaart and J. Wellner. *Weak convergence and empirical processes*. Springer-Verlag, New York, 1996.
- V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
- G. Wahba. *Splines Models for Observational Data*. Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.
- Q. Wu and D.X. Zhou. Support vector machine classifiers: linear programming versus quadratic programming. *Neural Computation*, 17:1160–1187, 2005.