THE OPTIMALITY OF KERNEL CLASSIFIERS IN SOBOLEV SPACE

Jianfa Lai[*]   **Zhifan Li**[*]   **Dongming Huang**   **Qian Lin**

ABSTRACT

Kernel methods are widely used in machine learning, especially for classification problems. However, the theoretical analysis of kernel classification is still limited. This paper investigates the statistical performances of kernel classifiers. With some mild assumptions on the conditional probability $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$, we derive an upper bound on the classification excess risk of a kernel classifier using recent advances in the theory of kernel regression. We also obtain a minimax lower bound for Sobolev spaces, which shows the optimality of the proposed classifier. Our theoretical results can be extended to the generalization error of overparameterized neural network classifiers. To make our theoretical results more applicable in realistic settings, we also propose a simple method to estimate the interpolation smoothness of $2\eta(x) - 1$ and apply the method to real datasets.

1 Introduction

In this paper, we study the problem of binary classification in a reproducing kernel Hilbert space (RKHS). Suppose n i.i.d samples $\{(X_i, Y_i) \in \mathcal{X} \times \{-1, 1\}\}$ are drawn from a joint distribution $(X, Y) \sim \rho$, where the conditional probability of the response variable Y given the predictor variable $X = x$ is denoted by $\eta(x) = \mathbb{P}(Y = 1 | X = x)$. We aim to find a classifier function $f(x) : \mathcal{X} \to [-1, 1]$ that minimizes the classification risk, defined as:

$$\mathcal{L}(\hat{f}) := \mathbb{P}_{(X,Y) \sim \rho}\left[\text{sign}(\hat{f}(X)) \neq Y\right].$$

The minimal classification risk is achieved by the Bayes classifier function corresponding to ρ, which is defined as $f^*_\rho(x) = 2\eta(x) - 1$. Our main focus is on analyzing the convergence rate of the classification excess risk, defined as:

$$\mathcal{E}(\hat{f}) = \mathcal{L}(\hat{f}) - \mathcal{L}(f^*_\rho).$$

This paper studies a class of kernel methods called spectral algorithms (which will be defined in Section 2.3) for constructing estimators of f^*_ρ. The candidate functions are selected from an RKHS $\mathcal{H}$, which is a separable Hilbert space associated with a kernel function K defined on $\mathcal{X}$ [37, 38]. Spectral algorithms, as well as kernel methods, are becoming increasingly important in machine learning because both experimental and theoretical results show that overparameterized neural network classifiers exhibit similar behavior to classifiers based on kernel methods [6]. Therefore, understanding the properties of classification with spectral algorithms can shed light on the generalization of deep learning classifiers.

In kernel methods context, researchers often assume that $f^*_\rho \in \mathcal{H}$, and have obtained the minimax optimality of spectral algorithms [8, 9]. Some researchers have also studied the convergence rate of the generalization error of misspecified spectral algorithms $\left(f^*_\rho \notin \mathcal{H}\right)$, assuming that f^*_ρ falls into the interpolation space $[\mathcal{H}]^s$ with some $s > 0$ [17, 46]. In this line of work, researchers consider the embedding index condition which reflects the capability of $\mathcal{H}$ embedding into L^∞ space. Moreover, [46] extends the boundedness assumption to the cases where $[\mathcal{H}]^s \cap L^\infty(X, \mu) \subsetneq [\mathcal{H}]^s$.

Motivated by the aforementioned studies, we adopt similar assumptions in our study of kernel classifiers trained via the gradient flow. We assume that the Bayes classifier $f^*_\rho \in [\mathcal{H}]^s$ satisfies the boundedness condition $f^*_\rho \in [-1, 1]$. We first derive the upper bound of the classification excess risk, showing that the generalization error of the kernel classifier is highly related to the interpolation smoothness s. To clarify the minimax optimality of kernel classification, we then obtain the minimax lower bound for classification in Sobolev RKHS, which is a novel result in the literature. Our technique is motivated by the connection between kernel estimation and infinite-width neural networks, and our

---

[*]These authors contributed equally to this work.

framework can be applied to neural network supervised learning. Furthermore, we provide a method to estimate the interpolation space smoothness parameter s and also present some numerical results for neural network classification problems through simulation studies and real data analysis.

1.1 Our contribution

In this paper, we study the generalization error of kernel classifiers. We show that

i) We show the generalization error of the gradient flow kernel classifier is bounded by $O(n^{-s\beta/(2s\beta+2)})$ provided that the Bayes classifier $f_\rho^* \in [\mathcal{H}]^s$, where β is the eigenvalue decay rate (EDR) of the kernel. This result is not only applicable to the Sobolev RKHS $\mathcal{H}$ but also to any RKHS with the embedding index $\alpha_0 = 1/\beta$, such as the RKHS with dot-product kernels and the RKHS with shift-invariant periodic kernels.

ii) We establish a minimax lower bound on the classification excess risk in the interpolation space of Sobolev RKHS. Combined with the results in i), the convergence rate of the kernel classifier is minimax optimal in Sobolev space. Before our work, [45] illustrated a similar result of the minimax lower bound for Besov spaces. However, the result has only been proved for $d = 1$ by [25] and the case for $d > 1$ remains unresolved.

iii) To make our theoretical results more applicable in realistic settings, we propose a simple method to estimate the interpolation smoothness s. We apply this method to estimate the relative smoothness of various real datasets with respect to the neural tangent kernels, where the results are in line with our understanding of these real datasets.

1.2 Related works

We study the classification rules derived from a class of real-valued functions in a reproducing kernel Hilbert space (RKHS), which are used in kernel methods such as Support Vector Machines (SVM) [38]. Most of the existing works consider hinge loss as the loss function, i.e. [42, 39, 4, 7] etc. Another kernel method, kernel ridge regression, also known as least-square SVM [38], is investigated by some researchers [44, 35]. Recently, some works have combined the least square loss classification with neural networks [11, 22].

We choose kernel methods because it allows us to use the integral operator tool for analysis [10, 9, 17, 46], while previous SVM works tend to use the empirical process technique [39]. Moreover, we can easily extend the $\mathcal{H}$ to the misspecified model case $[\mathcal{H}]^s$ when true model f_ρ^* belongs to a less-smooth interpolation space. Furthermore, we consider more regularization methods, collectively known as spectral algorithms, which were first proposed and studied by [36, 5, 9]. [46] combined these two ideas and obtained minimax optimality for the regression model. We extend their results to the classification problems.

We study the minimax optimality of Sobolev kernel classification, and before our work, the minimax lower bound of classification excess risk for the RKHS class was seldom considered. [31, 32] have discussed Classification problems in Sobolev space, but they did not consider the lower bound of classification risk. [2, 3, 33] provided some minimax lower bound techniques for classification, but how to solve RKHS remains unknown. Sobolev space (see, e.g., [1]) is known as a vector space of functions equipped with a norm that is a combination of L^2-norms of the function together with its derivatives up to a given order and can be embedded into Hölder class. Inspired by the minimax lower bound for Hölder class classification in [3], we derive the lower bound for the Sobolev class.

Recently, deep neural networks have gained incredible success in classification tasks from image classification [27, 21] to natural language processing [12]. Since [24] introduced the neural tangent kernel, The gradient flow of the training process can be well approximated by a simpler gradient flow associated with the NTK kernel when the width of neural networks is sufficiently large [28, 29]. Therefore, we can analyze the classification risk of neural networks trained by gradient descent.

2 Preliminaries

We observe n samples $\{(X_i, Y_i) \in \mathcal{X} \times \{-1, 1\}\}$ where $\mathcal{X} \subset \mathbb{R}^d$ is compact. Let ρ be an unknown probability distribution on $\mathcal{X} \times \{-1, 1\}$ and μ be the marginal distribution on $\mathcal{X}$. We assume μ has a uniformly bounded density $0 < \mu_{min} \leq \mu(x) \leq \mu_{max}$ for $x \in \mathcal{X}$. The classification task is to predict the unobserved label y given a new input x. The conditional probability is defined as $\eta(x) = \mathbb{P}(Y = 1|X = x)$. For any classifier f, the risk based on the 0-1 loss can be written as

$$\mathcal{L}(f) = \boldsymbol{E}_{(X,Y)\sim\rho}\mathbb{I}\{\text{sign}(f(X)) \neq Y\} = \boldsymbol{E}_X[(1 - \eta(X))\mathbb{I}\{f(X) \geq 0\} + \eta(X)\mathbb{I}\{f(X) < 0\}]. \tag{1}$$

One of the minimizers of the risk has the form $f_\rho^* = 2\eta - 1$. Let $\mathcal{L}^* = \mathcal{L}(f_\rho^*)$. For any classifier $\hat{f}$ learned from data, its accuracy is often characterized by the classification excess risk, which can be formulated as

$$\mathcal{E}(\hat{f}) = \mathcal{L}(\hat{f}) - \mathcal{L}^* = \boldsymbol{E}_X(|f_\rho^*(X)|\mathbb{I}\{\hat{f}(X)f_\rho^*(X) < 0\}). \tag{2}$$

In the rest of this section, we introduce some essential concepts in RKHS and kernel classifiers. In Section 2.1, we review some definitions in the interpolation space of RKHS. The relationship between fractional Sobolev space and Sobolev RKHS is presented in Section 2.2. Section 2.3 presents the explicit formula of the gradient-flow kernel classifier and the corresponding rewritten form through spectral algorithms and filter functions.

2.1 Interpolation Space of RKHS

Denote $L^2(\mathcal{X}) := L^2(\mathcal{X}, \mu)$ as the L^2 space. Throughout the paper, we denote by $\mathcal{H}$ a separable RKHS on $\mathcal{X}$ with respect to a continuous kernel function K. We also assume that $\sup_{x \in \mathcal{X}} K(x, x) \leq \kappa$ for some constant κ. The celebrated Mercer's theorem shows that there exist non-negative numbers $\lambda_1 \geq \lambda_2 \geq \cdots$ and functions $e_1, e_2, \cdots \in L^2(\mathcal{X})$ such that $\langle e_i, e_j \rangle_{L^2(\mathcal{X})} = \delta_{ij}$ and

$$K_d(x, x') = \sum_{j=1}^{\infty} \lambda_j e_j(x) e_j(x'), \tag{3}$$

where the series on the right hand side converges in $L^2(\mathcal{X})$.

Denote the natural embedding inclusion operator by $S_k : \mathcal{H} \to L^2(\mathcal{X}, \mu)$. Moreover, the adjoint operator $S_k^* : L^2(\mathcal{X}, \mu) \to \mathcal{H}$ is an integral operator, i.e., for $f \in L^2(\mathcal{X}, \mu)$ and $x \in \mathcal{X}$, we have

$$(S_k^* f)(x) = \int_{\mathcal{X}} K(x, x') f(x') d\mu(x').$$

It is well-known that S_k and S_k^* are Hilbert-Schmidt operators (and thus compact) and their HS norms (denoted as $\| \cdot \|_2$) satisfy that

$$\|S_k^*\|_2 = \|S_k\|_2 = \|K\|_{L^2(\mathcal{X}, \mu)} := \left(\int_{\mathcal{X}} K(x, x) d\mu(x) \right)^{1/2} \leq \kappa.$$

Next, we define two integral operators as follows:

$$L := S_k S_k^* : L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu), \quad T := S_k^* S_k : \mathcal{H} \to \mathcal{H}.$$

L and T are self-adjoint, positive-definite, and in the trace class (and thus Hilbert-Schmidt and compact). Their trace norms (denoted as $\| \cdot \|_1$) satisfy that $\|L\|_1 = \|T\|_1 = \|S_k\|_2^2 = \|S_k^*\|_2^2$.

For any $s \geq 0$, the fractional power integral operator $L^s : L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu)$ and $T^s : \mathcal{H} \to \mathcal{H}$ are defined as

$$L^s(f) = \sum_{j=1}^{\infty} \lambda_j^s \langle f, e_j \rangle_{L^2} e_j, \quad T^s(f) = \sum_{j=1}^{\infty} \lambda_j^s \left\langle f, \lambda_j^{\frac{1}{2}} e_j \right\rangle_{\mathcal{H}} \lambda_j^{\frac{1}{2}} e_j. \tag{4}$$

The interpolation space $[\mathcal{H}]^s$ is defined as

$$[\mathcal{H}]^s := \left\{ \sum_{j=1}^{\infty} a_j \lambda_j^{s/2} e_j : \sum_j^{\infty} a_j^2 < \infty \right\} \subseteq L^2(\mathcal{X}) \tag{5}$$

It is easy to show that $[\mathcal{H}]^s$ is also a separable Hilbert space with orthogonal basis $\{\lambda_i^{s/2} e_i\}_{i \in N}$. Specially, we have $[\mathcal{H}]^0 \subseteq L^2(\mathcal{X}, \mu)$, $[\mathcal{H}]^1 = \mathcal{H}$ and $[\mathcal{H}]^{s_2} \subsetneq [\mathcal{H}]^{s_1} \subsetneq [\mathcal{H}]^0$ for any numbers $0 < s_1 < s_2$. For the functions in $[\mathcal{H}]^s$ with larger s, we say they have higher (relative) interpolation smoothness with respect to the RKHS (the kernel).

2.2 Fractional Sobolev Space and Sobolev RKHS

For $m \in \mathbb{N}$, we denote the usual Sobolev space $W^{m,2}(\mathcal{X})$ by $H^m(\mathcal{X})$ and $L^2(\mathcal{X})$ by $H^0(\mathcal{X})$. Then the (fractional) Sobolev space for any real number $r > 0$ can be defined through the real interpolation

$$H^r(\mathcal{X}) := \left(L^2(\mathcal{X}), H^m(\mathcal{X}) \right)_{\frac{r}{m}, 2}$$

3

where $m := min\{k \in \mathbb{N} : k > r\}$.

It is well known that when $r > d/2$, H^r is a separable RKHS with respect to a bounded kernel and the corresponding eigenvalue decay rate (EDR) is $\beta = 2r/d$. Furthermore, the interpolation space of $H^r(\mathcal{X})$ under Lebesgue measure is given by

$$[H^r(\mathcal{X})]^s = H^{rs}(\mathcal{X}). \tag{6}$$

It follows that given a Sobolev RKHS $\mathcal{H} = H^r$ for $r > d/2$, if $f \in H^a$ for any $a > 0$, one can find that $f \in [\mathcal{H}]^s$ with $s = a/r$. Thus, in this paper, we will assume that the Bayes classifier f_ρ^* is in the interpolation of the Sobolev RKHS $[\mathcal{H}]^s$.

2.3 Kernel Classifiers: Spectra Algorithm

We then introduce a more general framework known as spectra algorithm [36, 8, 5]. We define the filter function and the spectral algorithms as follows:

Definition 1 (Filter function). *Let $\{\varphi_\nu : [0, \kappa^2] \to R^+ \mid \nu \in \Gamma \subseteq R^+\}$ be a class of functions and $\psi_\nu(z) = 1 - z\varphi_\nu(z)$. If φ_ν and ψ_ν satisfy:*

- *$\forall \alpha \in [0, 1]$, we have $\quad \sup_{z \in [0, \kappa^2]} z^\alpha \varphi_\nu(z) \leq E\nu^{1-\alpha}, \quad \forall \nu \in \Gamma$;*

- *$\exists \tau \geq 1$ s.t. $\forall \alpha \in [0, \tau]$, we have $\quad \sup_{z \in [0, \kappa^2]} |\psi_\nu(z)| z^\alpha \leq F_\tau \nu^{-\alpha}, \quad \forall \nu \in \Gamma$,*

where E, F_τ are absolute constants, then we call φ_ν a filter function. We refer to ν as the regularization parameter and τ as the qualification.

Definition 2 (spectral algorithm). *Let φ_ν be a filter function index with $\nu > 0$. Given the samples Z, a spectral algorithm produces an estimator of f_ρ^* given by $\hat{f}_\nu = \varphi_\nu(T_X) g_Z$.*

The following example shows that $\hat{f}_t(x)$ can be formulated by the spectral algorithms.

Example 1 (Classifier with Gradient flow). *The filter function of gradient flow φ_ν can be defined as $\varphi_\nu^{gf}(z) = \frac{1-e^{-\nu z}}{z}$. The qualification τ could be any positive number, $E = 1$ and $F_\tau = (\tau/e)^\tau$. So that for a test input x, the predicted output is given by $\hat{y} = \text{sign}(\hat{f}_\nu(x))$.*

Other spectral algorithms consist of kernel ridge regression, spectral cut-off, iterated Tikhonov, and so on. For more examples, we refer to [18]. Spectral algorithms differ in $\varphi_\nu(z)$ and $\psi_\nu(z)$, which is corresponding to saturation effect defined in [18]. Moreover, [30] gives a thorough analysis of the saturation effect for kernel ridge regression.

Notations. Denote $B(x, r)$ as a ball, and $\lambda[B(x, r)]$ is denoted as the Lebesgue measure of $B(x, r)$. We use $\| \cdot \|_{\mathscr{B}(B_1, B_2)}$ to denote the operator norm of a bounded linear operator from a Banach space B_1 to B_2, i.e., $\|A\|_{\mathscr{B}(B_1, B_2)} = \sup_{\|f\|_{B_1}=1} \|Af\|_{B_2}$. Without bringing ambiguity, we will briefly denote the operator norm as $\| \cdot \|$. In addition, we use $\text{tr}\, A$ and $\|A\|_1$ to denote the trace and the trace norm of an operator. We use $\|A\|_2$ to denote the Hilbert-Schmidt norm.

3 Main Results

3.1 Assumptions

This subsection lists the standard assumptions for general RKHS $\mathcal{H}$ and clarifies how these assumptions correspond to properties of Sobolev RKHS.

Assumption 1 (Source condition). *For $s > 0$, there is a constant $B > 0$ such that $f_\rho^* \in [\mathcal{H}]^s$ and $\|f_\rho^*\|_{[\mathcal{H}]^s} \leq B$.*

This assumption is weak since s can be small. However, functions in $[\mathcal{H}]^s$ with smaller s are less smooth, which will be harder for an algorithm to estimate.

Assumption 2 (Eigenvalue Decay Rate (EDR)). *The EDR of the eigenvalues $\{\lambda_j\}$ associated to the kernel K is $\beta > 1$, i.e.,*

$$cj^{-\beta} \leq \lambda_j \leq Cj^{-\beta} \tag{7}$$

for some positive constants c and C.

Note that the eigenvalues λ_i and EDR are only determined by the marginal distribution μ and the RKHS $\mathcal{H}$. For Sobolev RKHS H^r equipped with Lebesgue measure ν and bounded domain with smooth boundary $\mathcal{X} \subseteq R^d$, it is well known that when $r > d/2$, H^r is a separable RKHS with respect to a bounded kernel and the corresponding eigenvalue decay rate (EDR) is $\beta = 2r/d$ [16].

Our next assumption is the embedding index. First, we give the definition of embedding property [17]: For $0 < \alpha < 1$, there is a constant $A > 0$ with $\|[H]_\nu^\alpha \hookrightarrow L_\infty(\nu)\| \leq A$. This means $[\mathcal{H}]^\alpha$ is continuously embedded into $L_\infty(\nu)$ and the operator norm of the embedding is bounded by A. The larger α is, the weaker the embedding property is.

Assumption 3 (Embedding index). *Suppose that there exists $\alpha_0 > 0$, such that*

$$\alpha_0 = \inf \left\{ \alpha \in [\frac{1}{\beta}, 1] : \|[\mathcal{H}]^\alpha \hookrightarrow L^\infty(\mathcal{X}, \mu)\| < \infty \right\},$$

and we refer to α_0 as the embedding index of an RKHS $\mathcal{H}$.

This assumption directly implies that all the functions in $[\mathcal{H}]^\alpha$ are μ-a.e bounded for $\alpha > \alpha_0$. Moreover, we will clarify this assumption for Sobolev kernels and dot-product kernels on $\mathbb{S}^{d-1}$ in the appendix.

3.2 Minimax optimality of kernel classifiers

This subsection presents our main results on the minimax optimality of kernel classifiers. We first establish a minimax lower bound for the Sobolev RKHS $H^r(\mathcal{X})$ under the source condition (Assumption 4). We then provide an upper bound based on Assumptions 4, 5, and 6, and we clarify that the Sobolev RKHS satisfies these assumptions. As a result, we demonstrate that the Sobolev kernel classifier is minimax rate optimal.

Theorem 3.1 (Lower Bound). *Suppose $f_\rho^* \in [H^r(\mathcal{X})]^s$ for $s > 0$, where H^r is the Sobolev RKHS. For all learning methods $\hat{f}$, for any fixed $\delta \in (0, 1)$, when n is sufficiently large, there is a distribution $\rho \in \mathcal{P}$ such that, with probability at least $1 - \delta$, we have*

$$\mathcal{E}(\hat{f}) \geq C\delta n^{-\frac{s\beta}{2(s\beta+1)}}, \tag{8}$$

where C is a universal constant.

Theorem 3.1 shows the minimax lower bound on the classification excess risk over the interpolation space of the Sobolev RKHS. Theorem 3.1 also establishes a minimax lower bound at the rate of $n^{-\frac{a}{2a+d}}$ for the Sobolev space H^a with $a = rs$. [45] illustrated a similar result of the minimax lower bound for Besov spaces. However, the result has only been proved for $d = 1$ by [25] and the case for $d > 1$ remains unresolved.

The following theorem presents an upper bound for the kernel classifier.

Theorem 3.2 (Upper Bound). *Suppose that Assumptions 4, 5, and 6 hold for $0 < s \leq 2\tau$, where τ is the qualification of the filter function. By choosing $\nu \asymp n^{\frac{\beta}{s\beta+1}}$, for any fixed $\delta \in (0, 1)$, when n is sufficiently large, with probability at least $1 - \delta$, we have*

$$\mathcal{E}(\hat{f}_\nu) \leq C \left(\ln \frac{4}{\delta} \right) n^{-\frac{s\beta}{2(s\beta+1)}} \tag{9}$$

where C is a constant independent of n and δ.

Combined with Theorem 3.1, Theorem 3.2 shows that by choosing a proper early-stopping time, the Sobolev kernel classifier is minimax rate optimal. Moreover, given the kernel and the decay rate β, the optimal rate is mainly affected by the smoothness s of f_ρ^* with respect to the kernel. Thus, in Section 5, we will introduce how to estimate the smoothness of functions or datasets given a specific kernel.

We emphasize that Theorem 3.2 can be applied to any general RKHS with an embedding index $\alpha_0 = 1/\beta$, such as an RKHS with a shift-invariant periodic kernel and an RKHS with a dot-product kernel. Thanks to the uniform convergence of overparameterized neural networks [28, 29], Theorem 3.2 can also be applied to analyze the generalization error of the neural network classifiers. We will discuss this application in the next section.

4 Applications in Neural Networks

Suppose that we have observed n i.i.d. samples $\{X_i, Y_i\}_{i=1}^n$ from ρ. For simplicity, we further assume that the marginal distribution μ of ρ is the uniform distribution on the unit sphere $\mathbb{S}^{d-1}$. We use a neural network with L hidden layers

and width m to perform the classification on $\{X_i, Y_i\}_{i=1}^n$. The network model $f(x; \theta)$ and the resulting prediction are given by the following equations

$$h^0(x) = x, \quad h^l(x) = \sqrt{\frac{2}{m}}\sigma(W^{l-1}h^{l-1}(x)), \quad l = 1, ..., L$$

$$f(x; \theta) = W^L h^L(x) \quad \text{and} \quad \hat{y} = \text{sign}(f(x; \theta)),$$

(10)

where h^l represents the hidden layer, $\sigma(x) := \max(x, 0)$ is the ReLU activation (applied elementwise), $W^0 \in \mathbb{R}^{m \times d}$ and $W^l \in \mathbb{R}^{m \times m}$ are the parameters of the model. We use θ to represent the collection of all parameters flatten as a column vector. With the mirrored initialization (shown in [29]), we consider the training process given by the gradient flow $\dot{\theta} = -\partial L(\theta)/\partial\theta$, where the squared loss function is adopted $L(\theta) = \frac{1}{2n}\sum_{i=1}^n (Y_i - f(X_i, \theta))^2$.

The consideration for this choice of loss function is that the squared loss function is robust for optimization and more suitable for hard learning scenarios ([23, 11, 26]). [23] showed that the square loss function has been shown to perform well in modern classification tasks, especially in natural language processing while [26] presented the out-of-distribution robustness of the square loss function.

When the network is overparameterized, [29] showed that the trained network $f(x; \theta)$ can be approximated by a kernel gradient method with respect to the following neural tangent kernel

$$K_{ntk}(x, x') = \sum_{r=0}^L \kappa_1^{(r)}(\bar{u}) \prod_{s=r}^{L-1} \kappa_0(\kappa_1^{(s)}(\bar{u}))$$

(11)

where $\bar{u} = \langle x, x' \rangle$, $\kappa_1^{(p)} = \kappa_1 \underbrace{\circ \cdots \circ}_{p \text{ times}} \kappa_1$ represents p times composition of κ_1 and $\kappa_1^{(0)}(u) = u$ by convention; if $r = L$, the product $\prod_{s=r}^{L-1}$ is understood to be 1. Denote $Y_{[n]} = (Y_1, ..., Y_n)^T$, $K(X_{[n]}, X_{[n]})$ as an $n \times n$ matrix of $(K(X_i, X_j))_{i,j \in [n]}$ and $\lambda_{min} = \lambda_{min}(K(X_{[n]}, X_{[n]}))$ The following proposition shows the uniform convergence of $f(x; \theta)$.

Proposition 4.1 (Theorem 1 in [29]). *Suppose $x \in \mathbb{S}^{d-1}$. For any $\epsilon > 0$, any hidden layer $L \geq 2$, and $\delta \in (0, 1)$, when the width $m \geq \text{poly}\left(n, \lambda_{min}^{-1}, ||Y_{[n]}||_2, \ln(1/\delta), \ln(1/\epsilon)\right)$, with probability at least $1 - \delta$ with respect to random initialization, we have*

$$\sup_{t \geq 0} \sup_{x \in \mathcal{X}} |f_t(x; \theta) - f_t^{ntk}(x)| \leq \epsilon.$$

where $f_t^{ntk}(x)$ is defined as in Example 1 but with the kernel K_{ntk}.

Theorem G.5 in [19] showed that the RKHS of the NTK on $\mathbb{S}^{d-1}$ is a Sobolev space. Moreover, the kernel K_{ntk} is a dot-product kernel satisfying a polynomial eigenvalue decay $\beta = d/(d-1)$. Thus, we can obtain the following corollary by combining Theorem 3.2 and Proposition 4.1.

Corollary 4.2. *Suppose that $x \in \mathbb{S}^{d-1}$ and Assumption 4 holds for $\mathcal{H}$ being the RKHS of the kernel K_{ntk} and $s > 0$. Suppose $t \asymp n^{\frac{\beta}{s\beta+1}}$. For any fixed $\delta \in (0, 1)$, when $m \geq \text{poly}\left(n, \lambda_{min}^{-1}, ||Y_{[n]}||_2, \ln(1/\delta)\right)$ and n is sufficiently large, with probability at least $1 - \delta$, we have*

$$\mathcal{E}(f_t(x; \theta)) \leq C\left(\ln \frac{4}{\delta}\right) n^{-\frac{s\beta}{2(s\beta+1)}}$$

(12)

where C is a constant independent of n and δ.

This corollary shows that the generalization error of a fine-tuned, overparameterized neural network classifier converges at the rate of $n^{-\frac{s\beta}{2(s\beta+1)}}$. This result also highlights the need for additional efforts to understand the smoothness of real datasets with respect to the neural tangent kernel. A larger value of s corresponds to a faster convergence rate, indicating the possibility of better generalization performance. Determination of the smoothness parameter s will allow us to assess the performance of an overparameterized neural network classifier on a specific dataset.

5 Estimation of smoothness

In this section, we provide a simple example to illustrate how to determine the relative smoothness s of the ground-truth function with respect to the kernel. Then we introduce a simple method to estimate s with noise and apply the method to real datasets with respect to the NTK.

6

Determination of s. Suppose that $\mathcal{X} \in [0,1]$ and the marginal distribution $\mu_{\mathcal{X}}$ is a uniform distribution on $[0,1]$. We consider the min kernel $K_{min}(x,x') = \min(x,x')$ [43] and denote by $\mathcal{H}_{min}$ the corresponding RKHS. The eigenvalues and the eigenfunctions of $\mathcal{H}_{min}$ are

$$\lambda_j = \left(\frac{2j-1}{2}\pi\right)^{-2}, \quad e_j(x) = \sqrt{2}\sin(\frac{2j-1}{2}\pi x), \quad j \geq 1. \tag{13}$$

Thus, the EDR is $\beta = 2$. For illustration, we consider the ground true function $f^*(x) = \cos(2\pi x)$. Suppose $f^*(x) = \sum_j^\infty f_j e_j(x)$, then we have $f_j = \sqrt{2}\int_0^1 \cos(2\pi x)\sin(\frac{2j-1}{2}\pi x)dx \asymp j^{-1}$. Thus, $f_j \asymp j^{-r}$ where $r = 1$. By the definition of the interpolation space, we have $s = \frac{2r-1}{\beta} = 0.5$.
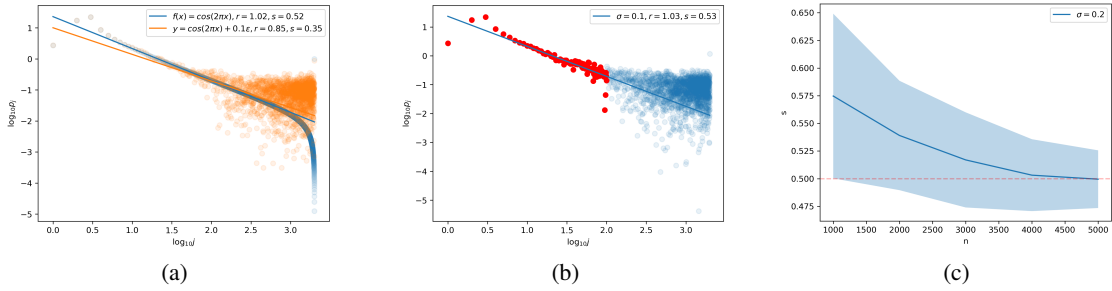

(a)   (b)   (c)

Figure 1: Experiments for estimating the smoothness parameter s in regression settings. (a) Naive estimation based on $2,000$ sample points for $\sigma = 0$ (blue) and $\sigma = 0.1$ (orange). (b) Truncation Estimation based on $2,000$ sample points with truncation point 100. In both plots (a) and (b), the x-axis is the logarithmic index j and the y-axis is the logarithmic p_j.(c) Truncation Estimation across various values of sample size n, each repeated 50 times. The blue line represents the average of estimates, the shaded area represents one standard deviation, and the true value is indicated by the orange dashed line.

Estimation of s in regression. To better understand the estimation process, we first consider regression settings where the noises have an explicit form and we then consider classification settings. Suppose that we have n i.i.d. samples of $X_{[n]} = [X_1, ..., X_n]^\top$ and $Y_{[n]} = [Y_1, ..., Y_n]^\top \in$ from $Y_i = f^*(X_i) + \sigma\epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0,1)$.

We start with a naive estimation method. Let $K_{min}(X_{[n]}, X_{[n]})$ be the kernel matrix. Suppose the eigendecomposition is given by $K_{min}(X_{[n]}, X_{[n]}) = V\Sigma V^\top$, where $V = [v_1, ..., v_n]$ is the eigenvector matrix, v_i's are the eigenvectors, and Σ is the diagonal matrix of the eigenvalues. We can estimate r by estimating the decay rate of p_j, where $p_j = Y_{[n]}^\top v_j$. To visualize the convergence rate r, we perform logarithmic least-squares to fit p_j with respect to the index j and display the values of the slope r and the smoothness parameter s.

For $\sigma = 0$, r can be accurately estimated by the above naive method since there is no noise in Y_i's. The blue line and dots in Figure 1 (a) present the estimation of s in this case, where the estimate is around the true value 0.5. However, for $\sigma = 0.1$, the naive estimation is not accurate, as shown by the orange line and dots in Figure 1 (a).

To improve the accuracy of the estimation, we introduce a simple modification called *Truncation Estimation*, described as follows. We select some fixed integer as a truncation point and estimate the decay rate of p_j up to the truncation point. For the example with $\sigma = 0.1$, we choose the truncation point 100 and the result is shown in Figure 1 (b). We observe that the estimation becomes much more accurate than the naive estimation, with an estimate of $s = 0.53$ not too far away from the true value 0.5. In general, noise in the data can worsen the estimation accuracy, while increasing the sample size can improve the accuracy and robustness of the estimation. In Figure 1 (c), we show the result for estimating s in repeated experiments with more noisy data ($\sigma = 0.2$), where we observe that as the sample size n increases, the estimation becomes accurate.

Estimation of s in classification. Now we consider the classification settings, where the population is given by $\mathbb{P}(Y = 1|X = x) = (f^*(x)+1)/2$. Unlike regression problems, the variance of the noise $\epsilon = y - f^*(x)$ is determined by $f^*(x)$ and may not be negligible. Nonetheless, in classification problems, we can still estimate the smoothness parameter s using Truncation Estimation, thanks to the fact that increasing the sample size can improve its performance. The results are shown in Figure 2, where we can indeed make similar observations to those in Figure 1 (b) and (c).
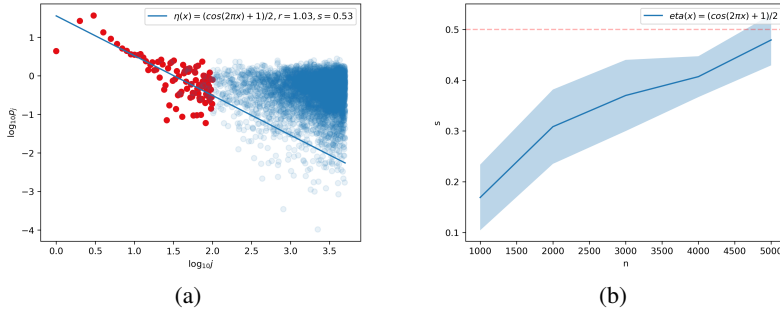
(a)   (b)

Figure 2: Experiments for estimating the smoothness parameter s in classification settings. (a) The experiment uses $5,000$ sample points and the truncation point is 100. (b) Truncation Estimation across various values of sample size n, each repeated 50 times. The blue line represents the average of estimates, the shaded area represents one standard deviation, and the true value is indicated by the orange dashed line.

Kernel	MNIST	Fashion-MNIST	CIFAR-10
NTK-1	0.4862 (0.0824)	0.4417 (0.0934)	0.1992 (0.0724)
NTK-2	0.4871 (0.0793)	0.4326 (0.0875)	0.2047 (0.0831)
NTK-3	0.4865 (0.0815)	0.4372 (0.0768)	0.1965 (0.0795)

Table 1: Truncation Estimation of the relative smoothness s of different real data sets with different NTKs. $NTK - L$ indicates the L-hidden-layer NTK. We only consider two classes of labels for each dataset: Label 1 and 7 for MNIST, trousers and sneakers for Fashion-MNIST, cars and horses for CIFAR-10. We randomly select 5,000 data points and choose the truncation point 100 to estimate s. For each dataset and each kernel, we repeat 50 times and the standard deviation is in parentheses.

As an application of Truncation Estimation, we estimate the relative smoothness of real data sets with respect to the NTK defined in (11). The results are shown in Table 1. We can see that with respect to the NTK, MNIST has the largest relative smoothness while CIFAR-10 has the smallest one. This result aligns with the common knowledge that MNIST is the easiest dataset while CIFAR-10 is the most difficult one of these three datasets.

Limitations The misspecified spectral algorithms (assuming $f_\rho^* \in [\mathcal{H}]^s$) are studied since 2009 (e.g., [40, 14, 34, 17, 46]). However, to the best of our knowledge, there is barely any result on the estimation of the smoothness s. This paper is the first to propose the s estimation method even though the method is more susceptible to noise when the sample size is not enough or f^* has more complex structures. For example, if $f^* = \sum_{j=1}^{\infty} f_j e_j(x)$, where $f_j^2 = j^{-s_1\beta-1}$ when j is odd and $f_j^2 = j^{-s_2\beta-1}$ when j is even ($s_1 > s_2$). For the kernel K with EDR β, $f_\rho^* \in [\mathcal{H}]^{s_2}$ instead of $[\mathcal{H}]^{s_1}$ or $[\mathcal{H}]^s$ for some $s \in (s_2, s_1)$. In this mixed smoothness case, our method tends to give an estimation $\hat{s} \in (s_2, s_1)$. A more detailed discussion of the limitations is presented in the appendix. We will try to find more accurate s estimation methods for general situations in the near future.

6 Discussion

In this paper, we study the generalization error of kernel classifiers in Sobolev space (the interpolation of the Sobolev RKHS). We show the optimality of kernel classifiers under the assumption that the ground true function is in the interpolation of RKHS with the kernel. The minimax optimal rate is $n^{-s\beta/2(s\beta+1)}$, where s is the smoothness parameter of the ground true function. Building upon the connection between kernel methods and neural networks, we obtain an upper bound on the generalization error of overparameterized neural network classifiers. To make our theoretical result more applicable to real problems, we propose a simple method called Truncation Estimation to estimate the relative smoothness s. Using this method, we examine the relative smoothness of three real datasets, including MNIST, Fashion-MNIST and CIFAR-10. Our results confirm that among these three datasets, MNIST is the simplest for classification using NTK classifiers while CIFAR-10 is the hardest.

8

References

[1] Robert A Adams and John JF Fournier. *Sobolev spaces*. Elsevier, 2003.

[2] Jean-Yves Audibert. Classification under polynomial entropy and margin assump-tions and randomized estimators. 2004. URL `https://api.semanticscholar.org/CorpusID:1986311`.

[3] Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *Annals of statistics*, 35 (2):608–633, 2007.

[4] Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8), 2008.

[5] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

[6] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To Understand Deep Learning We Need to Understand Kernel Learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 541–549. PMLR, July 2018.

[7] Gilles Blanchard, Olivier Bousquet, and Pascal Massart. Statistical performance of support vector machines. *The Annals of Statistics*, 36(2):489–531, 2008.

[8] Andrea Caponnetto. Optimal rates for regularization operators in learning theory. Technical report, MAS-SACHUSETTS INST OF TECH CAMBRIDGE COMPUTER SCIENCE AND ARTIFICIAL . . . , 2006.

[9] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[10] Ernesto De Vito, Andrea Caponnetto, and Lorenzo Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, 5:59–85, 2005.

[11] Ahmet Demirkaya, Jiasi Chen, and Samet Oymak. Exploring the role of loss functions in multiclass classification. In *2020 54th annual conference on information sciences and systems (ciss)*, pp. 1–5. IEEE, 2020.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.

[13] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[14] Lee H Dicker, Dean P Foster, and Daniel Hsu. Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. *Electronic Journal of Statistics*, 11(1):1022–1047, 2017.

[15] Liang Ding, Tianyang Hu, Jiahang Jiang, Donghao Li, Wenjia Wang, and Yuan Yao. Random smoothing regularization in kernel gradient descent learning. *arXiv preprint arXiv:2305.03531*, 2023.

[16] D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge Tracts in Mathematics. Cambridge University Press, 1996. doi: 10.1017/CBO9780511662201.

[17] Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *The Journal of Machine Learning Research*, 21(1):8464–8501, 2020.

[18] L Lo Gerfo, Lorenzo Rosasco, Francesca Odone, E De Vito, and Alessandro Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.

[19] Moritz Haas, David Holzmüller, Ulrike von Luxburg, and Ingo Steinwart. Mind the spikes: Benign overfitting of kernels and neural networks in fixed dimension. *arXiv preprint arXiv:2305.14077*, 2023.

[20] Thomas Hamm and Ingo Steinwart. Adaptive learning rates for support vector machines working on data with low intrinsic dimension. *The Annals of Statistics*, 49(6):3153–3180, 2021.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[22] Tianyang Hu, Jun Wang, Wenjia Wang, and Zhenguo Li. Understanding square loss in training overparametrized neural network classifiers. *arXiv preprint arXiv:2112.03657*, 2021.

[23] Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In *International Conference on Learning Representations*, 2020.

[24] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

[25] G Kerkyacharian and D Picard. Density estimation in besov spaces zyxwvutsrqponmlkjihgfedcbazyxwvut. *Statistics & probability letters*, 13:15–24, 1992.

[26] Simon Kornblith, Honglak Lee, Ting Chen, and Mohammad Norouzi. Demystifying loss functions for classification. 2020.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[28] Jianfa Lai, Manyun Xu, Rui Chen, and Qian Lin. Generalization ability of wide neural networks on $\mathbb{R}$. *arXiv preprint arXiv:2302.05933*, 2023.

[29] Yicheng Li, Zixiong Yu, Guhan Chen, and Qian Lin. Statistical optimality of deep wide neural networks. *arXiv preprint arXiv:2305.02657*, 2023.

[30] Yicheng Li, Haobo Zhang, and Qian Lin. On the saturation effect of kernel ridge regression. In *International Conference on Learning Representations*, February 2023.

[31] Sébastien Loustau. Aggregation of svm classifiers using sobolev spaces. *Journal of Machine Learning Research*, 9(7), 2008.

[32] Sébastien Loustau. Penalized empirical risk minimization over besov spaces. *Electronic Journal of Statistics*, 3: 824–850, 2009.

[33] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

[34] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis R. Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *ArXiv*, abs/1805.10074, 2018.

[35] Ryan Rifkin, Gene Yeo, Tomaso Poggio, et al. Regularized least-squares classification. *Nato Science Series Sub Series III Computer and Systems Sciences*, 190:131–154, 2003.

[36] Lorenzo Rosasco, Ernesto De Vito, and Alessandro Verri. Spectral methods for regularization in learning theory. *DISI, Universita degli Studi di Genova, Italy, Technical Report DISI-TR-05-18*, 2005.

[37] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

[38] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[39] Ingo Steinwart and Clint Scovel. Fast rates for support vector machines using gaussian kernels. *The Annals of Statistics*, 35(2):575–607, 2007.

[40] Ingo Steinwart, Don R Hush, Clint Scovel, et al. Optimal rates for regularized least squares regression. In *COLT*, pp. 79–93, 2009.

[41] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York ; London, 1st edition, 2009.

[42] Grace Wahba. Soft and hard classification by reproducing kernel hilbert space methods. *Proceedings of the National Academy of Sciences*, 99(26):16524–16530, 2002.

[43] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[44] Dao-Hong Xiang and Ding-Xuan Zhou. Classification with gaussians and convex loss. *Journal of Machine Learning Research*, 10(7), 2009.

[45] Yuhong Yang. Minimax nonparametric classification. i. rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, 1999.

[46] Haobo Zhang, Yicheng Li, Weihao Lu, and Qian Lin. On the optimality of misspecified kernel ridge regression. *arXiv preprint arXiv:2305.07241*, 2023.

7 Appendix

In this section, we first show the proof of the upper bound of the classification excess risk (A.1 and A.2) and then present the minimax lower bound (A.3). Before the proof, We list again the standard assumptions for general RKHS $\mathcal{H}$ in this section.

Assumption 4 (Source condition). *For $s > 0$, there is a constant $B > 0$ such that $f_\rho^* \in [\mathcal{H}]^s$ and*

$$\|f_\rho^*\|_{[\mathcal{H}]^s} \le B.$$

Assumption 5 (Eigenvalue Decay Rate (EDR)). *The EDR of the eigenvalues $\{\lambda_j\}$ associated to the kernel K is β, i.e.,*

$$cj^{-\beta} \le \lambda_j \le Cj^{-\beta} \tag{14}$$

for some positive constants c and C and $\beta > 1$.

Assumption 6 (Embedding index). *Suppose that there exists $\alpha_0 > 0$, such that*

$$\alpha_0 = \inf \left\{ \alpha \in [\frac{1}{\beta}, 1] : \|[\mathcal{H}]^\alpha \hookrightarrow L^\infty(\mathcal{X}, \mu)\| < \infty \right\},$$

and we refer to α_0 as the embedding index of an RKHS $\mathcal{H}$.

Define the sampling operator $K_x : \mathbb{R} \to \mathcal{H}, \ y \mapsto yK(x, \cdot)$ and its adjoint operator $K_x^* : \mathcal{H} \to \mathbb{R}, \ f \mapsto f(x)$. Further, we define the sample covariance operator $T_X : \mathcal{H} \to \mathcal{H}$ as

$$T_X := \frac{1}{n} \sum_{i=1}^n K_{X_i} K_{X_i}^*.$$

Then we know that $\|T_X\| \le \|T_X\|_1 \le \kappa^2$, where $\|\cdot\|$ denotes the operator norm and $\|\cdot\|_1$ denotes the trace norm. Further, define the sample basis function

$$g_Z := \frac{1}{n} \sum_{i=1}^n K_{X_i} Y_i \in \mathcal{H}.$$

We also introduce a more general framework known as spectra algorithm [36, 8, 5]. We define the filter function and the spectral algorithms as follows:

Definition 3 (Filter function). *Let $\left\{\varphi_\nu : \left[0, \kappa^2\right] \to R^+ \mid \nu \in \Gamma \subseteq R^+\right\}$ be a class of functions and $\psi_\nu(z) = 1 - z\varphi_\nu(z)$. If φ_ν and ψ_ν satisfy:*

- *$\forall \alpha \in [0, 1]$, we have*

$$\sup_{z \in [0, \kappa^2]} z^\alpha \varphi_\nu(z) \le E\nu^{1-\alpha}, \quad \forall \nu \in \Gamma; \tag{15}$$

- *$\exists \tau \ge 1$ s.t. $\forall \alpha \in [0, \tau]$, we have*

$$\sup_{z \in [0, \kappa^2]} |\psi_\nu(z)| z^\alpha \le F_\tau \nu^{-\alpha}, \quad \forall \nu \in \Gamma, \tag{16}$$

where E, F_τ are absolute constants, then we call φ_ν a filter function. We refer to ν as the regularization parameter and τ as the qualification.

Definition 4 (spectral algorithm). *Let φ_ν be a filter function index with $\nu > 0$. Given the samples Z, the spectral algorithm produces an estimator of f_ρ^* given by*

$$\hat{f}_\nu = \varphi_\nu (T_X) g_Z. \tag{17}$$

7.1 Some bounds

Throughout the proof, we denote

$$T_\nu = T + \nu^{-1}; \quad T_{X\nu} = T_X + \nu^{-1}$$

where ν is the regularization parameter. In addition, we denote $L^2(X, \mu)$ as L^2, $L^\infty(X, \mu)$ as L^∞ for brevity throughout the proof. We use $a_n \asymp b_n$ to denote that there exist constants c and C such that $ca_n \le b_n \le Ca_n, \forall n = 1, 2, \cdots$; use $a_n \lesssim b_n$ to denote that there exists an constant C such that $a_n \le Cb_n, \forall n = 1, 2, \cdots$ In addition, denote the effective dimension as

$$N(\nu) = \operatorname{tr}\left(T\left(T + \nu^{-1}\right)^{-1}\right) = \sum_{i \in N} \frac{\lambda_i}{\lambda_i + \nu^{-1}}$$

Lemma 7.1. *Suppose $\nu > 1$. If $\lambda_i \asymp i^{-\beta}$, we have*

$$N(\nu) \asymp \nu^{\frac{1}{\beta}}.$$

Proof. Since $ci^{-\beta} \le \lambda_i \le Ci^{-\beta}$, we have

$$N(\nu) = \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \nu^{-1}} \le \sum_{i=1}^{\infty} \frac{Ci^{-\beta}}{Ci^{-\beta} + \nu^{-1}} = \sum_{i=1}^{\infty} \frac{C}{C + \nu^{-1}i^\beta}$$

$$\le \int_0^\infty \frac{C}{\nu^{-1}x^\beta + C} dx = \nu^{\frac{1}{\beta}} \int_0^\infty \frac{C}{y^\beta + C} dy \le C_1 \nu^{\frac{1}{\beta}}$$

for some constant C_1. Since $\nu > 1$, the proof for the lower bound can be obtained similarly. $\qquad\square$

7.1.1 Approximation error

Recall that we have defined the sample basis function g_Z and the spectral algorithm $\hat{f}_\nu$. We also need the following notations: define the expectation of g_Z as

$$g = Eg_Z = \int_{\mathcal{X}} K(x, \cdot) f_\rho^*(x) d\mu(x) = S_k^* f_\rho^* \in \mathcal{H},$$

and

$$f_\nu = \varphi_\nu(T)g = \varphi_\nu(T) S_k^* f_\rho^*$$

The following conclusion based on [46] bounds the L^2-norm of $f_\nu - f_\rho^*$ for spectral algorithm:

Lemma 7.2. *Suppose that Assumption 4 holds for $0 < s \le 2\tau$. Then for any $\nu > 0$, we have*

$$\left\| f_\nu - f_\rho^* \right\|_{L^2} \le F_\tau R \nu^{-\frac{s}{2}}.$$

Proof. Because $f_\rho^* \in [\mathcal{H}]^s$, we assume $f_\rho^* = L^{\frac{s}{2}} g_0$ for some $g_0 \in L^2(P)$, so that $\|g_0\|_{L^2} \le B$ by Assumption 4. By the definition of f_ν, we have:

$$\begin{aligned}
\|f_\nu - f_\rho^*\|_{L^2} &= \|\varphi_\nu(T) S_k^* f_\rho^* - f_\rho^*\|_{L^2} \\
&= \|(\varphi_\nu(L)L - Id) L^{\frac{s}{2}} g_0\|_{L^2} \\
&\le \|(\psi_\nu(L) L^{\frac{s}{2}} g_0\|_{L^2} \\
&\le F_r B \nu^{-\frac{s}{2}}.
\end{aligned}$$

Where the second equality holds by the definition of natural embedding inclusion operator S_k, and Id denotes identity mapping. The first inequality holds because of the definition of ψ_ν and the second inequality holds for (16). $\qquad\square$

7.1.2 Estimation error

We rewrite the estimation error as follows

$$\begin{aligned}
\left\| \hat{f}_\nu - f_\nu \right\|_{L^2} &= \left\| T^{\frac{1}{2}} \left(\hat{f}_\nu - f_\nu \right) \right\|_{\mathcal{H}} \\
&= \left\| T^{\frac{1}{2}} T_\nu^{-\frac{1}{2}} \cdot T_\nu^{\frac{1}{2}} T_{X\nu}^{-\frac{1}{2}} \cdot T_{X\nu}^{\frac{1}{2}} \left(\hat{f}_\nu - f_\nu \right) \right\|_{\mathcal{H}} \\
&\le \left\| T^{\frac{1}{2}} T_\nu^{-\frac{1}{2}} \right\|_{\mathcal{B}(\mathcal{H})} \cdot \left\| T_\nu^{\frac{1}{2}} T_{X\nu}^{-\frac{1}{2}} \right\|_{\mathcal{B}(\mathcal{H})} \cdot \left\| T_{X\nu}^{\frac{1}{2}} \left(\hat{f}_\nu - f_\nu \right) \right\|_{\mathcal{H}}.
\end{aligned} \tag{18}$$

Step 1. The first part can be bounded by the following lemma, whose proof is simple and omitted.

Lemma 7.3.
$$\left\| T^{\frac{1}{2}}\left(T_\nu\right)^{-1/2}\right\|^2 = \sup_{i\geq 1}\frac{\lambda_i}{\lambda_i + \nu^{-1}} \leq 1.$$

For the second part, we recall a result [17, Lemma 11].

Lemma 7.4 ([17]). *Suppose that the RKHS $\mathcal{H}$ has the embedding index α_0. Then for any $\alpha_0 < \alpha \leq 1$ and all $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\left\| T_\nu^{-\frac{1}{2}}\left(T - T_X\right)T_\nu^{-\frac{1}{2}}\right\| \leq \frac{4M_\alpha^2\nu^\alpha}{3n}B + \sqrt{\frac{2M_\alpha^2\nu^\alpha}{n}B},$$

where

$$B = \ln\frac{4N(\nu)\left(\|T\| + \nu^{-1}\right)}{\delta\|T\|}.$$

Lemma 7.5. *If the sample size $n \geq 8M_\alpha^2\nu^\alpha B$, we have:*

$$\left\| T_\nu^{\frac{1}{2}}T_{X\nu}^{-\frac{1}{2}}\right\|_{\mathscr{B}(\mathcal{H})} \leq 3$$

holds with probability at least $1-\delta$.

Proof. We calculate $T_\nu^{\frac{1}{2}}T_{X\nu}^{-\frac{1}{2}}$ directly, and combined with Lemma 7.4

$$
\begin{aligned}
T_{X\nu} &= T_X + \nu^{-1} = T_X - T + T + \nu^{-1}\\
&= (T_\nu)^{\frac{1}{2}}\left[Id - (T_\nu)^{-\frac{1}{2}}(T - T_X)(T_\nu)^{-\frac{1}{2}}\right](T_\nu)^{\frac{1}{2}},
\end{aligned}
$$

By Lemma 7.4 when $n \geq 8M_\alpha^2\nu^\alpha B$, we have:

$$\left\| T_\nu^{-\frac{1}{2}}\left(T - T_X\right)T_\nu^{-\frac{1}{2}}\right\| \leq \frac{4}{3}\cdot\frac{1}{8} + \sqrt{\frac{1}{4}} \leq \frac{2}{3}$$

holds with probability at least $1-\delta$. So the Neumann series gives us the following bound

$$
\begin{aligned}
\left\| T_\nu^{\frac{1}{2}}T_{X\nu}^{-\frac{1}{2}}\right\|_{\mathscr{B}(\mathcal{H})} &= \left\|\left(Id - (T_\nu)^{-\frac{1}{2}}(T - T_X)(T_\nu)^{-\frac{1}{2}}\right)^{-1}\right\|_{\mathscr{B}(\mathcal{H})}\\
&\leq \sum_{k=0}^\infty\left\| T_\nu^{-\frac{1}{2}}\left(T - T_X\right)T_\nu^{-\frac{1}{2}}\right\|^k\\
&\leq \sum_{k=0}^\infty\left(\frac{2}{3}\right)^k = 3.
\end{aligned}
$$

$\square$

Step 2. For the third part in the last line of (18), we have

$$\left\| T_{X\nu}^{\frac{1}{2}}\left(\hat{f}_\nu - f_\nu\right)\right\|_{\mathcal{H}} \leq \left\| T_{X\nu}^{\frac{1}{2}}\varphi_\nu\left(T_X\right)\left(g_Z - T_X f_\nu\right)\right\|_{\mathcal{H}} + \left\| T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)f_\nu\right\|_{\mathcal{H}}. \tag{19}$$

The second part of RHS in (19) is complicated to calculate, but its proof follows the same argument as in Step 3 of Theorem 16 in [46]. Therefore, we provide the following result without proof:

Lemma 7.6 (Theorem 16 in [46]). *If $s < 2\tau$, we have:*

$$\left\| T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)f_\nu\right\|_{\mathcal{H}} \leq 6F_\tau ER\nu^{-\frac{s}{2}} + \Delta_1 I_{s>2}, \tag{20}$$

where Δ_1 denotes

$$\Delta_1 = 32\max\left(\frac{s-1}{2}, 1\right)EF_\tau R\kappa^{s-1}\nu^{-\frac{1}{2}}n^{-\frac{\min(s,3)-1}{4}}\ln\frac{6}{\delta}.$$

We need $s < 2\tau$, because we use (16) to upper bound $\|T_X^{\frac{s}{2}}\psi_\nu(T_X)\|$ for every s.

To bound the first term of RHS in (19), we begin with a lemma, whose proof is postponed to Section 7.1.3.

Lemma 7.7. *Suppose that Assumption 4, 5 and 6 hold for $\frac{1}{\beta} \leq \alpha_0 < 1$, and conditional probability of Y is given by*

$$\mathbb{P}(Y = 1 | X = x) = \frac{1 + f_\rho^*(x)}{2}.$$

Then given $\nu > 0$, and $n \geq 1$, for any $\delta > 0$ and $\alpha > \alpha_0$, the following bound is satisfied with probability not less than $1 - \delta$

$$\left\| (T_\nu)^{-1/2} \left((g_Z - T_X f_\nu) - (g - T f_\nu) \right) \right\|_{\mathcal{H}}^2 \leq \frac{128 \left(\log \frac{2}{\delta} \right)^2}{n} \left(N(\nu) + \frac{4 M_\alpha^2 \nu^\alpha}{n} \right).$$

The next lemma provides a bound on the first term of RHS in (19).

Lemma 7.8. *Suppose that Assumption 4, 5 and 6 hold for $\frac{1}{\beta} \leq \alpha_0 < 1$. If the sample size $n \geq 8 M_\alpha^2 \nu^\alpha B$, where B is defined in Lemma 7.4 and $\nu > 0$, then for any $\delta > 0$ and $\alpha > \alpha_0$, the following bound is satisfied with probability at least $1 - \delta$:*

$$\left\| T_{X\nu}^{\frac{1}{2}} \varphi_\nu \left(T_X \right) \left(g_Z - T_X f_\nu \right) \right\|_{\mathcal{H}} \leq C \left[\left(\log \frac{2}{\delta} \right) \left(\frac{N(\nu)^{\frac{1}{2}}}{\sqrt{n}} + \frac{2 M_\alpha \nu^{\frac{\alpha}{2}}}{n} \right) + \left\| f_\rho^* - f_\nu \right\|_{L^2} \right],$$

where C is an absolute constant.

Proof.

$$
\begin{aligned}
\left\| T_{X\nu}^{\frac{1}{2}} \varphi_\nu \left(T_X \right) \left(g_Z - T_X f_\nu \right) \right\|_{\mathcal{H}} &= \left\| T_{X\nu}^{\frac{1}{2}} \varphi_\nu \left(T_X \right) T_{X\nu}^{\frac{1}{2}} \cdot T_{X\nu}^{-\frac{1}{2}} T_\nu^{\frac{1}{2}} \cdot T_\nu^{-\frac{1}{2}} \left(g_Z - T_X f_\nu \right) \right\|_{\mathcal{H}} \\
&\leq \left\| T_{X\nu}^{\frac{1}{2}} \varphi_\nu \left(T_X \right) T_{X\nu}^{\frac{1}{2}} \right\|_{\mathscr{B}(\mathcal{H})} \cdot \left\| T_{X\nu}^{-\frac{1}{2}} T_\nu^{\frac{1}{2}} \right\|_{\mathscr{B}(\mathcal{H})} \\
&\quad \cdot \left\| T_\nu^{-\frac{1}{2}} \left(g_Z - T_X f_\nu \right) \right\|_{\mathcal{H}}.
\end{aligned}
\tag{21}
$$

The first term can be upper bounded as

$$
\begin{aligned}
\left\| T_{X\nu}^{\frac{1}{2}} \varphi_\nu \left(T_X \right) T_{X\nu}^{\frac{1}{2}} \right\|_{\mathscr{B}(\mathcal{H})} &= \left\| (T_X + \nu^{-1}) \varphi_\nu \left(T_X \right) \right\|_{\mathscr{B}(\mathcal{H})} \\
&\leq \left\| T_X \varphi_\nu \left(T_X \right) \right\|_{\mathscr{B}(\mathcal{H})} + \nu^{-1} \left\| \varphi_\nu \left(T_X \right) \right\|_{\mathscr{B}(\mathcal{H})} \\
&\leq 2E
\end{aligned}
$$

because we have $z \varphi_\nu(z) \leq E$ by (15) and $\varphi_\nu(z) \leq E\nu$ by (16). Using Lemma 7.5, the second term can be bounded as

$$\left\| T_{X\nu}^{-\frac{1}{2}} T_\nu^{\frac{1}{2}} \right\|_{\mathscr{B}(\mathcal{H})} \leq 3.$$

It remains to bound the third term. Lemma 7.7 shows that

$$\left\| (T_\nu)^{-1/2} \left((g_Z - T_X f_\nu) - (g - T f_\nu) \right) \right\|_{\mathcal{H}}^2 \leq \frac{128 \left(\log \frac{2}{\delta} \right)^2}{n} \left(N(\nu) + \frac{4 M_\alpha^2 \nu^\alpha}{n} \right). \tag{22}$$

Thus, we have

$$
\begin{aligned}
\left\| T_\nu^{-\frac{1}{2}} \left(g_Z - T_X f_\nu \right) \right\|_{\mathcal{H}} &\leq \left\| T_\nu^{-\frac{1}{2}} \left[(g_Z - T_X f_\nu) - (g - T f_\nu) \right] \right\|_{\mathcal{H}} + \left\| T_\nu^{-\frac{1}{2}} \left(g - T f_\nu \right) \right\|_{\mathcal{H}} \\
&\leq 8\sqrt{2} \left(\log \frac{2}{\delta} \right) \left(\frac{N(\nu)^{\frac{1}{2}}}{\sqrt{n}} + \frac{2 M_\alpha \nu^{\frac{\alpha}{2}}}{n} \right) + \left\| T_\nu^{-\frac{1}{2}} S_k^* \right\|_{\mathscr{B}(L^2, \mathcal{H})} \left\| f_\rho^* - f_\nu \right\|_{L^2} \\
&\leq 8\sqrt{2} \left(\log \frac{2}{\delta} \right) \left(\frac{N(\nu)^{\frac{1}{2}}}{\sqrt{n}} + \frac{2 M_\alpha \nu^{\frac{\alpha}{2}}}{n} \right) + \left\| f_\rho^* - f_\nu \right\|_{L^2},
\end{aligned}
$$

where the second term is approximation error that has been referred to in Lemma 7.2. $\qquad \square$

14

Step 3. Now we combine the bounds for the three parts of estimation error in (18): the first two parts are corresponding to Lemma 7.3 and 7.5 respectively, and the third part is corresponding to (19), (20), and Lemma 7.8. Then based on Assumptions the same as Lemma 7.8, combined with $s < 2\tau$, we conclude that for any $\delta > 0$, it holds with probability at least $1 - \delta$

$$\left\|\hat{f}_\nu - f_\nu\right\|_{L^2} \le C\left(\log\frac{2}{\delta}\right)\left(\frac{N(\nu)^{\frac{1}{2}}}{\sqrt{n}} + \frac{2M_\alpha\nu^{\frac{\alpha}{2}}}{n}\right) \tag{23}$$
$$+ \left\|f_\rho^* - f_\nu\right\|_{L^2} + F_\tau ER\nu^{-\frac{s}{2}} + \Delta_1 I_{s>2},$$

where C is an absolute constant.

7.1.3 Proof of Lemma 7.7

Lemma 7.9 (Lemma 13 in [17]). *Let $(\mathcal{X}, B)$ be a measurable space, $\mathcal{H}$ be a separable RKHS on $\mathcal{X}$ w.r.t. a bounded and measurable kernel k, and μ be a probability distribution on $\mathcal{X}$. Then the following equality is satisfied, for $\nu > 0$,*

$$\int_{\mathcal{X}} \left\|(T_\nu)^{-1/2} K(x, \cdot)\right\|_{\mathcal{H}}^2 d\mu(x) = N(\nu).$$

If, in addition, $M_\alpha < \infty$ is satisfied, then the following inequality is satisfied, for $\nu > 0$ and μ-almost all $x \in \mathcal{X}$,

$$\left\|(T_\nu)^{-1/2} K(x, \cdot)\right\|_{\mathcal{H}}^2 \le M_\alpha^2 \nu^{-\alpha}.$$

we also consider $C_x : \mathcal{H} \to \mathcal{H}$ the integral operator w.r.t. the point measure at $x \in \mathcal{X}$,

$$C_x f := f(x)K(x, \cdot) = \langle f, K(x, \cdot)\rangle_{\mathcal{H}} K(x, \cdot),$$

And we have the operator norm:

$$\left\|(T_\nu)^{-1/2} C_x (T_\nu)^{-1/2}\right\|_{\mathscr{B}(\mathcal{H})} = \left\|(T_\nu)^{-1/2} K(x, \cdot)\right\|_{\mathcal{H}}^2 \le M_\alpha^2 \nu^\alpha.$$

Lemma 7.10 (Bernstein's Inequality in [9]). *Let (Ω, B, P) be a probability space, H be a separable Hilbert space, and $\xi : \Omega \to H$ be a random variable with*

$$E_P[\|\xi\|_{\mathcal{H}}^2] \le \sigma^2,$$
$$\|\xi(\omega)\|_{\mathcal{H}} \le \frac{L}{2}, \quad a.s.$$

Then, for $\tau \ge 1$ and $n \ge 1$, the following concentration inequality is satisfied

$$\mathbb{P}^n\left((\omega_1, \ldots, \omega_n) \in \Omega^n : \left\|\frac{1}{n}\sum_{i=1}^n \xi(\omega_i) - E_P\xi\right\|_{\mathcal{H}}^2 \ge 32\frac{\tau^2}{n}\left(\sigma^2 + \frac{L^2}{n}\right)\right) \le 2e^{-\tau}.$$

Proof of Lemma 7.7. Consider the random variable $\xi : X \times \{-1, 1\} \to \mathcal{H}$:

$$\xi(x, y) = (T_\nu)^{-1/2} (y - f_\nu(x))K(x, \cdot).$$

Denoted by D the empirical measure corresponding to the sample $\{X_i, Y_i\}_{i=1}^n$. It holds that

$$\frac{1}{n}\sum_{i=1}^n (\xi(X_i, U_i) - E_P\xi) = E_D\xi - E\xi$$

$$= (T_\nu)^{-1/2} ((g_Z - T_X f_\nu) - (g - T f_\nu)).$$

By Lemma 7.9, it holds that

$$\|\xi(x, y)\|_{\mathcal{H}} \le 2\left\|(T_\nu)^{-1/2} K(x, \cdot)\right\|_{\mathcal{H}} \le 2M_\alpha\nu^{\frac{\alpha}{2}}$$

for μ-almost all $x \in \mathcal{X}$, and it holds that

$$E(\|\xi(x, y)\|_{\mathcal{H}}^2) = \int_{\mathcal{X}} \left\|(T_\nu)^{-1/2} K(x, \cdot)\right\|_{\mathcal{H}}^2 \left[\sum_{i=1,-1} |i - f_\nu(x)|^2 P(Y = i \mid X = x)\right] d\mu(x)$$

$$\le 4N(\nu).$$

We let $\sigma^2 = 4N(\nu)$ and $L = 4M_\alpha\nu^{\frac{\alpha}{2}}$ and apply the Bernstein inequality in Lemma 7.10 to complete the proof. □

7.2 Upper bound on excess risk

Theorem 7.11 (L^2-risk upper bound). *Suppose that Assumption 4, 5 and 6 holds, and that $0 < s \leq 2\tau$. Let $\hat{f}_\nu$ be the estimator defined by (17). Then by choosing $\nu \asymp n^{\frac{\beta}{s\beta+1}}$, for any fixed $\delta \in (0,1)$ and any $1 \geq \alpha > \alpha_0$, when n is sufficiently large, with probability at least $1 - \delta$, we have*

$$\left\| \hat{f}_\nu - f_\rho^* \right\|_{L^2}^2 \leq \left(\ln \frac{6}{\delta} \right)^2 C n^{-\frac{s\beta}{s\beta+1}},$$

where C is a constant independent of n and δ.

Proof. We decomposed L^2-risk into the sum of the approximation error and the estimation error as as

$$\|\hat{f}_\nu - f_\rho^*\|_{L^2} \leq \|f_\nu - f_\rho^*\|_{L^2} + \|\hat{f}_\nu - f_\nu\|_{L^2}.$$

Using Lemma 7.2 for the appoximation error and (23) for the estimation error, we have

$$\|\hat{f}_\nu - f_\rho^*\|_{L^2} \leq C \left(\log \frac{2}{\delta} \right) \left(\frac{N(\nu)^{\frac{1}{2}}}{\sqrt{n}} + \frac{2M_\alpha \nu^{\frac{\alpha}{2}}}{n} \right) + F_\tau(E+1)R\nu^{-\frac{s}{2}} + \Delta_1 I_{s>2}.$$

Choosing $\nu \asymp n^{\frac{\beta}{s\beta+1}}$, we can obtain the following rates:

- By Lemma 7.1,

$$\frac{N(\nu)^{\frac{1}{2}}}{\sqrt{n}} \asymp \frac{\nu^{\frac{1}{2\beta}}}{\sqrt{n}} = n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} \tag{24}$$

-

$$F_\tau(E+1)R\nu^{-\frac{s}{2}} \asymp n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} \tag{25}$$

- Following Theorem 16 in [46], for any $s > 2$, we have

$$\Delta_1 \lesssim \left(\ln \frac{6}{\delta} \right) n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} \tag{26}$$

- Since $\alpha_0 = \frac{1}{\beta}$ for Sobolev RKHS, we can choose some α in $(\alpha_0, s + \frac{2}{\beta})$. It follows that

$$\frac{\nu^{\frac{\alpha}{2}}}{n} \lesssim n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} \tag{27}$$

Combining the above (24), (25),(26) and (27), we complete the proof. $\square$

Proof of Theorem 2 . The classification excess risk can be rewritten as

$$\begin{aligned}
\mathcal{E}(\hat{f}) &= \mathcal{L}(\hat{f}) - \mathcal{L}^* \\
&= \boldsymbol{E}_X[(1-\eta(X))(\mathbb{I}\{\hat{f}(X) \geq 0\} - \mathbb{I}\{2\eta(X)-1 \geq 0\}) + \eta(X)(\mathbb{I}\{\hat{f}(X) < 0\} - \mathbb{I}\{2\eta(X)-1 < 0\}))] \\
&= \boldsymbol{E}_X(|2\eta(X)-1|\mathbb{I}\{\hat{f}(X)(2\eta(X)-1) < 0\}) \\
&= \boldsymbol{E}_X(|f_\rho^*(X)|\mathbb{I}\{\hat{f}(X)f_\rho^*(X) < 0\}).
\end{aligned} \tag{28}$$

Based on (28) and the classic upper classification upper bound by [13], we have:

$$\begin{aligned}
\mathcal{E}(\hat{f}_\nu) &= \boldsymbol{E}_X(|f_\rho^*(X)|\mathbb{I}\{\hat{f}_\nu(X)f_\rho^*(X) < 0\}) \\
&\leq \int_{\mathcal{X}} \left| f_\rho^*(X) - \hat{f}_\nu(X) \right| d\nu(X) \\
&\leq \|\hat{f}_\nu - f_\rho^*\|_{L^2}
\end{aligned}$$

The remaining proof directly follows from Theorem 7.11. $\square$

7.3 Minimax lower bound

Proposition 7.12 (Theorem 2.5 in [41]). *Assume that $M \geq 2$ and suppose that Θ contains elements $\theta_0, \theta_1, \ldots, \theta_M$ and $P_{\theta_0}, P_{\theta_1}, \ldots, P_{\theta_M}$ are the probability measures such that*

(i) $d(\theta_i, \theta_j) \geq 2s > 0, \forall 0 \leq i \leq j \leq M$;

(ii) $P_{\theta_j} \ll P_{\theta_0}, \forall j = 1, \ldots, M$ and

$$\frac{1}{M} \sum_{j=1}^{M} KL(P_{\theta_j}, P_{\theta_0}) \leq a \log M \tag{29}$$

with $0 < \alpha < 1/8$.

Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta(d(\hat{\theta}, \theta) > s) \geq \frac{\sqrt{M}}{1 + \sqrt{M}}(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}) > 0 \tag{30}$$

Lemma 7.13 (Varshamov-Gilbert Bound). *Given $m \geq 8$, there exist $M \geq 2^{m/8}$ different elements on $\omega^{(0)}, \ldots, \omega^{(M)}$ on $\{-1, 1\}^m$ and $\omega^{(0)} = (0, \ldots, 0)$ such that*

$$\sum_{k=1}^{m} |\omega_k^{(i)} - \omega^{(j)}| \geq \frac{m}{4}, \quad 0 \leq i < j \leq M. \tag{31}$$

Lemma 7.14. *For $r > \frac{d}{2}$, $H^r(\mathcal{X})$ is a separable RKHS with respect to a bound kernel and the corresponding EDR is*

$$\beta = \frac{2r}{d}.$$

Let $u : \mathbb{R}_+ \to \mathbb{R}_+$ be a nonincreasing infinitely differentiable function such that $u = 1$ on $[0, 1/4]$ and $u = 0$ on $[1/2, \infty)$. We can take $u(x) = \left(\int_{1/4}^{1/2} u_1(z)dz\right)^{-1} \int_x^\infty u_1(z)dz$ where

$$u_1(x) = \begin{cases} \exp\left\{-\frac{1}{(1/2-x)(x-1/4)}\right\}, & \text{for } x \in (1/4, 1/2), \\ 0, & \text{otherwise.} \end{cases} \tag{32}$$

Given an integral $q = q(n) \geq 1$, we define the regular grid on $\mathbb{R}^d$ as

$$G_q = \left\{\left(\frac{2k_1 + 1}{2q}, \ldots, \frac{2k_d + 1}{2q}\right) : k_i \in \{0, \ldots, q-1\}, i = 1, \ldots, d\right\}. \tag{33}$$

We consider the partition $\mathcal{X}_1, \ldots, \mathcal{X}_{q^d}$ of $[0, 1]^d$ canonically defined using the grid G_q. $x \in \mathcal{X}_k$ if $z_k \in G_q$ is the closest point to $x \in [0, 1]^d$. If there exist several points in G_q closest to x we define $x \in \mathcal{X}_k$ if z_k is closest to 0.

Lemma 7.15. $\psi(x) = C_\psi q^{-sr} \sum_{k=1}^{q^d} \phi(q[x - z_k]) \in H^{sr}(\mathcal{X})$, where $\phi(x) = u(\|x\|_2)$.

Proof of Lemma 7.15. By the definition of ϕ, we have $\phi \in H^{sr}$ for any fixed $sr > 0$. Thus, $\|\phi\|_{H^{sr}}^2$ is bounded.

Denote $\psi_k(x) = \phi(q[x - z_k])$ and $\varphi(x) = \psi_k(x + z_k) = \phi(qx)$. It is easy to find that: i) $\|\psi_k\|_{[H^{sr}]} = \|\varphi\|_{H^{sr}}$; ii) $\langle \psi_i, \psi_j \rangle_{H^{sr}} = 0$ for $i \neq j$.

$$\|\psi\|_{H^{sr}}^2 = C_\psi^2 q^{-2sr} \|\sum_{k=1}^{q^d} \psi_k\|_{H^{sr}}^2 \tag{34}$$

$$= C_\psi^2 q^{-2sr} \left(\sum_{k=1}^{q^d} \|\psi_k\|_{H^{sr}}^2 + 2\sum_{i \neq j} \langle \psi_i, \psi_j \rangle_{H^{sr}}^2\right) \tag{35}$$

$$= C_\psi^2 q^{d-2sr} \|\varphi\|_{H^{sr}}^2 \tag{36}$$

17

Denote the Fourier transform of ϕ and ψ as $\hat{\phi}$ and $\hat{\psi}$

$$\hat{\varphi}(\xi) = \int_{B(0,\frac{1}{2q})} \phi(qx)e^{-2\pi i\xi x}\mathrm{d}x \tag{37}$$

$$= q^{-d}\int_{B(0,\frac{1}{2})} \phi(y)e^{-2\pi i\frac{\xi}{q}y}\mathrm{d}y \tag{38}$$

$$= q^{-d}\hat{\phi}(\frac{\xi}{q}) \tag{39}$$

Since $u(x)$ is infinitely differentiable function on $[0,1]^d$, then $\|\phi\|_{H^s}$ is bounded for any fixed $s > 0$. Then

$$\|\varphi\|_{H^{sr}}^2 = q^{-2d}\int_{R^d} |\hat{\phi}(\frac{\xi}{q})|^2(1+\|\xi\|_2^2)^s\mathrm{d}\xi \tag{40}$$

$$= q^{-2d}\int_{R^d} |\hat{\phi}(z)|^2(1+q^2\|z\|_2^2)^{sr}q^d\mathrm{d}z \tag{41}$$

$$\leq q^{2sr-d}\int_{R^d} |\hat{\phi}(z)|^2(1+\|z\|_2^2)^{sr}\mathrm{d}z \tag{42}$$

$$= q^{2sr-d}\|\phi\|_{H^{sr}}^2. \tag{43}$$

Thus, $\|\psi\|_{H^{sr}}^2 \leq C_\psi^2\|\phi\|_{H^{sr}}^2$. $\qquad\square$

Proof of Theorem 1. Denote $m = q^d$, we define $\mathcal{X}_0 = \mathbb{R}^d \setminus \bigcup_{i=1}^m \mathcal{X}_i$. Thus, $\mathcal{X}_0, ..., \mathcal{X}_m$ form a partition of $\mathbb{R}^d$.

Define the hypercube $\mathcal{C} = \{P_\omega : \omega = (\omega_1, ..., \omega_m) \in \{-1,1\}^m\}$ of probability distribution P_ω of (X,Y) on $Z = \mathbb{R}^d \times \{0,1\}$ as follows.

For any P_ω, the marginal distribution of X does not depend on ω and has a density μ w.r.t. the Lebesgue measure on $\mathbb{R}^d$ defined in the following way. Denote $v = m^{-1}$. Let $\mu(x) = v/\lambda[B(0,(4q)^{-1})]$ for $x \in B(z,(4q)^{-1})$, $z \in G_q$ and $\mu(x) = 0$ otherwise.

By Lemma 7.13, there exist $M \geq 2^{m/8}$ different elements on $\omega^{(0)}, ..., \omega^{(M)}$ on $\{-1,1\}^m$ and $\omega^{(0)} = (0,...,0)$. We take

$$f_i(x) = \omega_k^{(i)}\psi(x), \quad x \in \mathcal{X}_k, \quad i \in [M] \tag{44}$$

and $f_i(x) = 0$ for $x \in \mathcal{X}_0$. $f_0 = 0$ for $x \in \mathcal{X}$. We will assume that $C_\psi \leq 1$ to ensure that $\eta_i(x) = (1 + f_i(x))/2$ take values in $[0,1]$.

By Lemma 7.15, we have $f_i \in H^{sr} = [\mathcal{H}]^s$, $i = 0, ..., M$. Since $\eta_i(x) = (1 + f_i(x))/2$, we have

$$KL(\rho_i^n, \rho_0^n) = nKL(\rho_i, \rho_0) \tag{45}$$

$$= n\sum_{k=1}^m \int_{\mathcal{X}_k} \left(\frac{1+\omega_k^{(i)}\psi(x)}{2}\ln(1+\omega_k^{(i)}\psi(x)) + \frac{1-\omega_k^{(i)}\psi(x)}{2}\ln(1-\omega_k^{(i)}\psi(x))\right)\mu(x)\mathrm{d}x \tag{46}$$

$$\leq \frac{1}{2}nm\int_{B(x_0,(4q)^{-1})} \frac{v}{\lambda[B(0,(4q)^{-1})]}\left(\ln(1-\psi^2(x)) + \psi(x)\ln\left(\frac{1+\psi(x)}{1-\psi(x)}\right)\right)\mathrm{d}x \tag{47}$$

$$\leq \frac{1}{2}nm\int_{B(x_0,(4q)^{-1})} \frac{v}{\lambda[B(0,(4q)^{-1})]}\left(\psi(x)\ln\left(\frac{1+\psi(x)}{1-\psi(x)}\right)\right)\mathrm{d}x \tag{48}$$

$$\leq Cnmvq^{-2sr} \tag{49}$$

The last inequality is because with sufficiently large n, $\psi(x) = C_\psi q^{-sr} < 1/2$ for $x \in B(x_0,(4q)^{-1})$. To satisfy the second condition of Proposition 7.12, we need $C_\psi nmvq^{-2sr} \leq a\frac{m}{8}\ln 2$. We can take $nvq^{-2sr} = \Theta(1)$. Thus, $q = n^{1/(2sr+d)}$.

For the first condition, we have

$$d(f_i, f_j) = \mathbb{E}_x[|f_i(x)|1_{f_i(x)f_j(x)<0}] \tag{50}$$

$$= \mathbb{E}_x\left[\sum_{k=1}^{m} \psi(x)1_{\sigma_k^{(i)}\sigma_k^{(j)}<0}1_{x\in\mathcal{X}_k}\right] \tag{51}$$

$$\geq \frac{C_\psi q^{-sr}mv}{4} \tag{52}$$

$$= Cn^{-sr/(2sr+d)} \tag{53}$$

for some constant $C > 0$. By Proposition 7.12, we have the minimax rate $n^{-sr/(2sr+d)}$. Since $r = \beta d/2$, we have the minimax rate $n^{-\frac{s\beta}{2(s\beta+1)}}$. $\qquad\square$

7.4 Embedding index of Sobolev and dot-product kernels

7.4.1 Sobolev kernel

The interpolation space of $H^r(\mathcal{X})$ under Lebesgue measure is given by

$$[H^r(\mathcal{X})]^s = H^{rs}(\mathcal{X}). \tag{54}$$

By the embedding theorem of (fractional) Sobolev space (Theorem 4.27 in [1]), letting $\theta = r - \frac{d}{2} > 0$, we have

$$H^r(\mathcal{X}) \hookrightarrow C^{0,\theta}(\mathcal{X}) \hookrightarrow L^\infty(\mathcal{X}), \quad \theta = r - \frac{d}{2}.$$

Combined with (54), for a Sobolev RKHS $\mathcal{H} = H^r(\mathcal{X}), r > \frac{d}{2}$ and any $\alpha > \frac{1}{\beta} = \frac{d}{2r}$, we have

$$[H^r(\mathcal{X})]^\alpha = H^{r\alpha}(\mathcal{X}) \hookrightarrow C^{0,\theta}(\mathcal{X}) \hookrightarrow L^\infty(\mathcal{X}).$$

Therefore $\alpha_0 = \frac{1}{\beta}$ for the embedding index of a Sobolev RKHS.

7.4.2 dot-product kernel

Let k be a dot-product kernel on $X = S^d$, the unit sphere in R^{d+1}, and $\mu = \sigma$ be the uniform measure on S^d. Then, it is well-known that k can be decomposed as

$$k(x,y) = \sum_{n=0}^{\infty} \mu_n \sum_{l=1}^{a_n} Y_{n,l}(x)Y_{n,l}(y), \tag{55}$$

where $\{Y_{n,l}\}$ is a set of orthonormal basis of $L^2\left(S^d, \sigma\right)$ called the spherical harmonics. a_n is multiplicity and satisfies

$$a_n := \binom{n+d}{n} - \binom{n-2+d}{n-2}$$

We let $\mu_n \asymp n^{-d\beta}$ for some $\beta > 1$ and also we have $a_n \asymp n^{d-1}$ and $\sum_{i=1}^{n} a_i \asymp n^d$, we prove embedding index $\alpha_0 = \frac{1}{\beta}$.

Proof. By the Theorem 9 in [17], we only need to prove that for any $\alpha > \frac{1}{\beta}$, $\sum_{n=0}^{\infty} \mu_n^\alpha \sum_{l=1}^{a_n} Y_{n,l}(x)^2 < \infty$.

$$\sum_{n=0}^{\infty} \mu_n^\alpha \sum_{l=1}^{a_n} Y_{n,l}(x)^2 \leq \sum_{n=0}^{\infty} \mu_n^\alpha a_n$$

$$\leq \sum_{n=0}^{\infty} Cn^{d-1}n^{-\alpha d\beta}$$

$$= C\sum_{n=0}^{\infty} n^{-1-d(\alpha\beta-1)} < \infty$$

$\qquad\square$

8 Detailed discussion for f with complex structures

In this section, we illustrate some f^* with the complex structure and analyze the feasibility of our theories and method on these cases:

Mixed smoothness: Suppose $f^* = \sum_{j=1}^{\infty} f_j e_j(x)$ and $\{f_j\}_{j \in N}$ has different different decay rates. A simple example is the two-smoothness case, where $\{f_j\}_{j \in N}$ has two different decay rates, where $f_j^2 = j^{-s_1 \beta - 1}$ when j is odd and $f_j^2 = j^{-s_2 \beta - 1}$ when j is even ($s_1 > s_2$). For the kernel K with EDR β, by the definition of the interpolation space, we have

$$\|f^*\|_{[\mathcal{H}]^s} = \sum_{j=1}^{\infty} \frac{f_j^2}{\lambda_j^s} = \sum_{\{j : j \text{ is odd}\}} j^{s\beta - s_1\beta - 1} + \sum_{\{j : j \text{ is even}\}} j^{s\beta - s_2\beta - 1} \tag{56}$$

Thus, for $s < s_2$, we have $\|f^*\|_{[\mathcal{H}]^s}$ is bounded, meaning that $f_\rho^* \in [\mathcal{H}]^s$ where s can be arbitrary close to s_2. In this case, our theory can still be applied to find the generalization ability of the kernel classifiers ($n^{-\frac{s_2\beta}{s_2\beta+1}}$). This can be also applied to multi-smoothness cases.

However, in this case, Truncation Estimation, introduced in Section 5, is inaccurate. With a sufficient sample size, Truncation Estimation will find the smoothness s between s_1 and s_2 for the two-smoothness case while we need to find $s = s_2$. In this case, the method can be improved by performing linear regression on top of $\tilde{p}_j = \sup_{k > j} p_k$ even though this improvement method tends to underestimate s. We will find new s estimation methods for general situations in the near future.

Sobolev space of low intrinsic dimensionality There is a popular assumption on the real data, called manifold assumption, assuming that f^* is supported on a submanifold. More specifically, for $x \in \Omega \subset \mathbb{R}^D$, they assume that f^* belongs to the space of the low intrinsic dimensionality d and $d < D$. In this case, [20, 15] have come up with some definitions of the low intrinsic dimension assumption:

Assumption 7 (Low intrinsic dimension). *There exist positive constants c_1 and $d \leq D$ such that for all $\delta \in (0, 1)$, we have*

$$\mathcal{N}_{l_\infty^D}(\delta, \Omega) \leq c\delta^{-d} \tag{57}$$

where l_∞^D is the $\mathbb{R}^D$ space equipped with l_∞ norm and $\mathcal{N}_{l_\infty^D}(\delta, \Omega)$ is the covering number.

On the Sobolev space with smoothness r With the assumption of low intrinsic dimension, [20, 15] improved their results from $n^{-\frac{2r}{2r+D}}$ to $n^{-\frac{2r}{2r+d}}$ (regression problems). Though our theories can not solve this case, we believe that the technology can be applied to classification problems and this will be our future work.

Well-separated data The well-separated assumption is another popular assumption on the real data (like MNIST, CIFAR-10, and so on) since the testing accuracy of some neural network models is near 100%. The well-separated assumption, in our settings, means that $f^*(x) \in \{1, -1\}$, violating the continuity of f^*. However, we can use a continuous function to approximate such a discontinuous function. For example, $x \in [0, 1]$ and $f^*(x) = 1$ if $x > 1/2$ and $f^*(x) = -1$ if $x \leq 1/2$ (two regions). Then we can use an infinitely differentiable function ($s = \infty$) to approximate f^* and thus the estimator finds out that the function is arbitrarily smooth. This idea can be extended to the cases with finite regions.

However, for the real data, normally with a super large dimension, the number of regions may depend on the dimension. In this situation, our theories need more effort to explain the generalization ability of the kernel classifier (like extending our theories to the high dimensional settings).