But How Does It Work in Theory? Linear SVM with Random Features

Yitong Sun
Department of Mathematics
University of Michigan
Ann Arbor, MI, 48109
syitong@umich.edu

Anna Gilbert
Department of Mathematics
University of Michigan
annacg@umich.edu

Ambuj Tewari
Department of Statistics
University of Michigan
tewaria@umich.edu

Abstract

We prove that, under low noise assumptions, the support vector machine with $N \ll m$ random features (RFSVM) can achieve the learning rate faster than $O(1/\sqrt{m})$ on a training set with m samples when an optimized feature map is used. Our work extends the previous fast rate analysis of random features method from least square loss to 0-1 loss. We also show that the reweighted feature selection method, which approximates the optimized feature map, helps improve the performance of RFSVM in experiments on a synthetic data set.

1 Introduction

Kernel methods such as kernel support vector machines (KSVMs) have been widely and successfully used in classification tasks (Steinwart and Christmann [2008]). The power of kernel methods comes from the fact that they implicitly map the data to a high dimensional, or even infinite dimensional, feature space, where points with different labels can be separated by a linear functional. It is, however, time-consuming to compute the kernel matrix and thus KSVMs do not scale well to extremely large datasets. To overcome this challenge, researchers have developed various ways to efficiently approximate the kernel matrix or the kernel function.

The random features method, proposed by Rahimi and Recht [2008], maps the data to a finite dimensional feature space as a random approximation to the feature space of RBF kernels. With explicit finite dimensional feature vectors available, the original KSVM is converted to a linear support vector machine (LSVM), that can be trained by faster algorithms (Shalev-Shwartz et al. [2011], Hsieh et al. [2008]) and tested in constant time with respect to the number of training samples. For example, Huang et al. [2014] and Dai et al. [2014] applied RFSVM or its variant to datasets containing millions of data points and achieved performance comparable to deep neural nets.

Despite solid practical performance, there is a lack of clear theoretical guarantees for the learning rate of RFSVM. Rahimi and Recht [2009] obtained a risk gap of order $O(1/\sqrt{N})$ between the best RFSVM and KSVM classifiers, where N is the number of features. Although the order of the error bound is correct for general cases, it is too pessimistic to justify or to explain the actual computational benefits of random features method in practice. And the model is formulated as a constrained optimization problem, which is rarely used in practice.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

Cortes et al. [2010] and Sutherland and Schneider [2015] considered the performance of RFSVM as a perturbed optimization problem, using the fact that the dual form of KSVM is a constrained quadratic optimization problem. Although the maximizer of a quadratic function depends continuously on the quadratic form, its dependence is weak and thus, both papers failed to obtain an informative bound for the excess risk of RFSVM in the classification problem. In particular, such an approach requires RFSVM and KSVM to be compared under the same hyper-parameters. This assumption is, in fact, problematic because the optimal configuration of hyper-parameters of RFSVM is not necessarily the same as those for the corresponding KSVM. In this sense, RFSVM is more like an independent learning model instead of just an approximation to KSVM.

In regression settings, the learning rate of random features method was studied by Rudi and Rosasco [2017] under the assumption that the regression function is in the RKHS, namely the *realizable* case. They show that the uniform feature sampling only requires $O(\sqrt{m}\log(m))$ features to achieve $O(1/\sqrt{m})$ risk of squared loss. They further show that a data-dependent sampling can achieve a rate of $O(1/m^{\alpha})$, where $1/2 \leq \alpha \leq 1$, with even fewer features, when the regression function is sufficiently smooth and the spectrum of the kernel integral operator decays sufficiently fast. However, the method leading to these results depends on the closed form of the least squares solution, and thus we cannot easily extend these results to non-smooth loss functions used in RFSVM. Bach [2017] recently shows that for any given approximation accuracy, the number of random features required is given by the degrees of freedom of the kernel operator under such an accuracy level, when optimized features are available. This result is crucial for sample complexity analysis of RFSVM, though not many details are provided on this topic in Bach's work.

In this paper, we investigate the performance of RFSVM formulated as a regularized optimization problem on classification tasks. In contrast to the slow learning rate in previous results by Rahimi and Recht [2009] and Bach [2017], we show, for the first time, that RFSVM can achieve fast learning rate with far fewer features than the number of samples when the optimized features (see Assumption 2) are available, and thus we justify the potential computational benefits of RFSVM on classification tasks. We mainly considered two learning scenarios: the realizable case, and then unrealizable case, where the Bayes classifier does not belong to the RKHS of the feature map. In particular, our contributions are threefold:

1. We prove that under Massart's low noise condition, with an optimized feature map, RFSVM can achieve a learning rate of $\tilde{O}(m^{-\frac{c_2}{1+c_2}})$ [1], with $\tilde{O}(m^{\frac{2c_2}{1+c_2}})$ number of features when the Bayes classifier belongs to the RKHS of a kernel whose spectrum decays polynomially ($\lambda_i = O(i^{-c_2})$). When the decay rate of the spectrum of kernel operator is sub-exponential, the learning rate can be improved to $\tilde{O}(1/m)$ with only $\tilde{O}(\ln^d(m))$ number of features.

2. When the Bayes classifier satisfies the separation condition; that is, when the two classes of points are apart by a positive distance, we prove that the RFSVM using an optimized feature map corresponding to Gaussian kernel can achieve a learning rate of $\tilde{O}(1/m)$ with $\tilde{O}(\ln^{2d}(m))$ number of features.

3. Our theoretical analysis suggests reweighting random features before training. We confirm its benefit in our experiments over synthetic data sets.

We begin in Section 2 with a brief introduction of RKHS, random features and the problem formulation, and set up the notations we use throughout the rest of the paper. In Section 3, we provide our main theoretical results (see the appendices for the proofs), and in Section 4, we verify the performance of RFSVM in experiments. In particular, we show the improvement brought by the reweighted feature selection algorithm. The conclusion and some open questions are summarized at the end. The proofs of our main theorems follow from a combination of the sample complexity analysis scheme used by Steinwart and Christmann [2008] and the approximation error result of Bach [2017]. The fast rate is achieved due to the fact that the Rademacher complexity of the RKHS of N random features and with regularization parameter λ is only $O(\sqrt{N\log(1/\lambda)})$, while N and $1/\lambda$ need not be too large to control the approximation error when optimized features are available. Detailed proofs and more experimental results are provided in the Appendices for interested readers.

[1] $\tilde{O}(n)$ represents a quantity less than $Cn\log^k(n)$ for some k.

2 Preliminaries and notations

Throughout this paper, a labeled data point is a point (x, y) in $\mathcal{X} \times \{-1, 1\}$, where $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$. $\mathcal{X} \times \{-1, 1\}$ is equipped with a probability distribution $\mathbb{P}$.

2.1 Kernels and Random Features

A positive definite kernel function $k(x, x')$ defined on $\mathcal{X} \times \mathcal{X}$ determines the unique corresponding reproducing kernel Hilbert space (RKHS), denoted by $\mathcal{F}_k$. A map ϕ from the data space $\mathcal{X}$ to a Hilbert space H such that $\langle \phi(x), \phi(x') \rangle_H = k(x, x')$ is called a feature map of k and H is called a feature space. For any $f \in \mathcal{F}$, there exists an $h \in H$ such that $\langle h, \phi(x) \rangle_H = f(x)$, and the infimum of the norms of all such hs is equal to $\|f\|_{\mathcal{F}}$. On the other hand, given any feature map ϕ into H, a kernel function is defined by the equation above, and we call $\mathcal{F}_k$ the RKHS corresponding to ϕ, denoted by $\mathcal{F}_\phi$.

A common choice of feature space is the L^2 space of a probability space (ω, Ω, ν). An important observation is that for any probability density function $q(\omega)$ defined on Ω, $\phi(\omega; x)/\sqrt{q(\omega)}$ with probability measure $q(\omega)\mathrm{d}\nu(\omega)$ defines the same kernel function with the feature map $\phi(\omega; x)$ under the distribution ν. One can sample the image of x under the feature map ϕ, an L^2 function $\phi(\omega; x)$, at points $\{\omega_1, \ldots, \omega_N\}$ according to the probability distribution ν to approximately represent x. Then the vector in $\mathbb{R}^N$ is called a random feature vector of x, denoted by $\phi_N(x)$. The corresponding kernel function determined by ϕ_N is denoted by k_N.

A well-known construction of random features is the random Fourier features proposed by Rahimi and Recht [2008]. The feature map is defined as follows,

$$\begin{aligned} \phi : \mathcal{X} &\to L^2(\mathbb{R}^d, \nu) \oplus L^2(\mathbb{R}^d, \nu) \\ x &\mapsto (\cos(\omega \cdot x), \sin(\omega \cdot x)) \ . \end{aligned}$$

And the corresponding random feature vector is

$$\phi_N(x) = \frac{1}{\sqrt{N}} (\cos(\omega \cdot x), \cdots, \cos(\omega \cdot x), \sin(\omega \cdot x), \cdots, \sin(\omega \cdot x))^\intercal \ ,$$

where ω_is are sampled according to ν. Different choices of ν define different translation invariant kernels (see Rahimi and Recht [2008]). When ν is the normal distribution with mean 0 and variance γ^{-2}, the kernel function defined by the feature map is Gaussian kernel with bandwidth parameter γ,

$$k_\gamma(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\gamma^2}\right) .$$

Equivalently, we may consider the feature map $\phi_\gamma(\omega; x) := \phi(\omega/\gamma; x)$ with ν being standard normal distribution.

A more general and more abstract feature map can be constructed using an orthonormal set of $L^2(\mathcal{X}, \mathbb{P}_\mathcal{X})$. Given the orthonormal set $\{e_i\}$ consisting of bounded functions, and a nonnegative sequence $(\lambda_i) \in \ell^1$, we can define a feature map

$$\phi(\omega; x) = \sum_{i=1}^{\infty} \sqrt{\lambda_i} e_i(x) e_i(\omega) \, ,$$

with feature space $L^2(\omega, \mathcal{X}, \mathbb{P}_\mathcal{X})$. The corresponding kernel is given by $k(x, x') = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x')$. The feature map and the kernel function are well defined because of the boundedness assumption on $\{e_i\}$. A similar representation can be obtained for a continuous kernel function on a compact set by Mercer's Theorem (Lax [2002]).

Every positive definite kernel function k satisfying that $\int k(x, x) \, \mathrm{d}\mathbb{P}_\mathcal{X}(x) < \infty$ defines an integral operator on $L^2(x, \mathcal{X}, \mathbb{P}_\mathcal{X})$ by

$$\begin{aligned} \Sigma : L^2(\mathcal{X}, \mathbb{P}_\mathcal{X}) &\to L^2(\mathcal{X}, \mathbb{P}_\mathcal{X}) \\ f &\mapsto \int_\mathcal{X} k(x, t) f(t) \, \mathrm{d}\mathbb{P}_\mathcal{X}(t) \, . \end{aligned}$$

Σ is of trace class with trace norm $\int k(x,x)\,\mathrm{d}\mathbb{P}_{\mathcal{X}}(x)$. When the integral operator is determined by a feature map ϕ, we denote it by Σ_ϕ, and the ith eigenvalue in a descending order by $\lambda_i(\Sigma_\phi)$. Note that the regularization paramter is also denoted by λ but without a subscript. The decay rate of the spectrum of Σ_ϕ plays an important role in the analysis of learning rate of random features method.

2.2 Formulation of Support Vector Machine

Given m samples $\{(x_i, y_i)\}_{i=1}^m$ generated i.i.d. by $\mathbb{P}$ and a function $f : \mathcal{X} \to \mathbb{R}$, usually called a hypothesis in the machine learning context, the empirical and expected risks with respect to the loss function ℓ are defined by

$$R_m^\ell(f) := \frac{1}{m}\sum_{i=1}^m \ell(y_i, f(x_i)) \quad R_{\mathbb{P}}^\ell(f) := \mathbb{E}_{(x,y)\sim\mathbb{P}}\ell(y, f(x)) \ ,$$

respectively.

The 0-1 loss is commonly used to measure the performance of classifiers:

$$\ell^{0-1}(y, f(x)) = \begin{cases} 1 & \text{if } f(x)y \le 0; \\ 0 & \text{if } f(x)y > 0. \end{cases}$$

The function that minimizes the expected risk under 0-1 loss is called the Bayes classifier, defined by

$$f_{\mathbb{P}}^*(x) := \operatorname{sgn}(\mathbb{E}[y \mid x]) \ .$$

The goal of the classification task is to find a good hypothesis f with small excess risk $R_{\mathbb{P}}^{0-1}(f) - R_{\mathbb{P}}^{0-1}(f_{\mathbb{P}}^*)$. And to find the good hypothesis based on the samples, one minimizes the empirical risk. However, using 0-1 loss, it is hard to find the global minimizer of the empirical risk because the loss function is discontinuous and non-convex. A popular surrogate loss function in practice is the hinge loss: $\ell^h(f) = \max(0, 1 - yf(x))$, which guarantees that

$$R_{\mathbb{P}}^h(f) - \inf_f R_{\mathbb{P}}^h(f) \ge R_{\mathbb{P}}^{0-1}(f) - R_{\mathbb{P}}^{0-1}(f_{\mathbb{P}}^*) \, ,$$

where R^h means R^{ℓ^h} and R^{0-1} means $R^{\ell^{0-1}}$. See Steinwart and Christmann [2008] for more details.

A regularizer can be added into the optimization objective with a scalar multiplier λ to avoid overfitting the random samples. Throughout this paper, we consider the most commonly used ℓ^2 regularization. Therefore, the solution of the binary classification problem is given by minimizing the following objective

$$R_{m,\lambda}(f) = R_m^h(f) + \frac{\lambda}{2}\|f\|_{\mathcal{F}}^2 \ ,$$

over a hypothesis class $\mathcal{F}$. When $\mathcal{F}$ is the RKHS of some kernel function, the algorithm described above is called kernel support vector machine. Note that for technical convenience, we do not include the bias term in the formulation of hypothesis so that all these functions are from the RKHS instead of the product space of RKHS and $\mathbb{R}$ (see Chapter 1 of Steinwart and Christmann [2008] for more explanation of such a convention). Note that $R_{m,\lambda}$ is strongly convex and thus the infimum will be attained by some function in $\mathcal{F}$. We denote it by $f_{m,\lambda}$.

When random features ϕ_N and the corresponding RKHS are considered, we add N into the subscripts of the notations defined above to indicate the number of random features. For example $\mathcal{F}_N$ for the RKHS, $f_{N,m,\lambda}$ for the solution of the optimization problem.

3 Main Results

In this section we state our main results on the fast learning rates of RFSVM in different scenarios.

First, we need the following assumption on the distribution of data, which is required for all the results in this paper.

Assumption 1. *There exists $V \ge 2$ such that*

$$|\mathbb{E}_{(x,y)\sim\mathbb{P}}[y \mid x]| \ge 2/V \, .$$

This assumption is called Massart's low noise condition in many references (see for example Koltchinskii et al. [2011]). When $V = 2$ then all the data points have deterministic labels almost surely. Therefore it is easier to learn the true classifier based on observations. In the proof, Massart's low noise condition guarantees the variance condition (Steinwart and Christmann [2008])

$$\mathbb{E}[(\ell^h(f(x)) - \ell^h(f^*_{\mathbb{P}}(x)))^2] \leq V(R^h(f) - R^h(f^*_{\mathbb{P}}))\,, \tag{1}$$

which is a common requirement for the fast rate results. Massart's condition is an extreme case of a more general low noise condition, called Tsybakov's condition. For the simplicity of the theorem, we only consider Massart's condition in our work, but our main results can be generalized to Tsybakov's condition.

The second assumption is about the quality of random features. It was first introduced in Bach [2017]'s approximation results.

Assumption 2. *A feature map $\phi : \mathcal{X} \to L^2(\omega, \Omega, \nu))$ is called optimized if there exists a small constant μ_0 such that for any $\mu \leq \mu_0$,*

$$\sup_{\omega\in\Omega} \|(\Sigma + \mu I)^{-1/2}\phi(\omega; x)\|^2_{L^2(\mathbb{P})} \leq \mathrm{tr}(\Sigma(\Sigma + \mu I)^{-1}) = \sum_{i=1}^{\infty} \frac{\lambda_i(\Sigma)}{\lambda_i(\Sigma) + \mu}\,.$$

For any given μ, the quantity on the left hand side of the inequality is called leverage score with respect to μ, which is directly related with the number of features required to approximate a function in the RKHS of ϕ. The quantity on the right hand side is called degrees of freedom by Bach [2017] and effective dimension by Rudi and Rosasco [2017], denoted by $d(\mu)$. Note that whatever the RKHS is, we can always construct optimized feature map for it. In the Appendix A we describe two examples of constructing optimized feature map. When a feature map is optimized, it is easy to control its leverage score by the decay rate of the spectrum of Σ, as described below.

Definition 1. *We say that the spectrum of $\Sigma : L^2(\mathcal{X}, \mathbb{P}) \to L^2(\mathcal{X}, \mathbb{P})$ decays at a polynomial rate if there exist $c_1 > 0$ and $c_2 > 1$ such that*

$$\lambda_i(\Sigma) \leq c_1 i^{-c_2}\,.$$

We say that it decays sub-exponentially if there exist $c_3, c_4 > 0$ such that

$$\lambda_i(\Sigma) \leq c_3 \exp(-c_4 i^{1/d})\,.$$

The decay rate of the spectrum of Σ characterizes the capacity of the hypothesis space to search for the solution, which further determines the number of random features required in the learning process. Indeed, when the feature map is optimized, the number of features required to approximate a function in the RKHS with accuracy $O(\sqrt{\mu})$ is upper bounded by $O(d(\mu)\ln(d(\mu)))$. When the spectrum decays polynomially, the degrees of freedom $d(\mu)$ is $O(\mu^{-1/c_2})$, and when it decays sub-exponentially, $d(\mu)$ is $O(\ln^d(c_3/\mu))$ (see Lemma 6 in Appendix C for details). Examples on the kernels with polynomial and sub-exponential spectrum decays can be found in Bach [2017]. Our proof of Lemma 8 also provides some useful discussion.

With these preparations, we can state our first theorem now.

Theorem 1. *Assume that $\mathbb{P}$ satisfies Assumption 1, and the feature map ϕ satisfies Assumption 2. If $f^*_{\mathbb{P}} \in \mathcal{F}_\phi$ with $\|f^*_{\mathbb{P}}\|_{\mathcal{F}_\phi} \leq R$. Then when the spectrum of Σ_ϕ decays polynomially, by choosing*

$$\lambda = m^{-\frac{c_2}{2+c_2}}$$
$$N = 10C_{c_1,c_2} m^{\frac{2}{2+c_2}} (\ln(32C_{c_1,c_2} m^{\frac{2}{2+c_2}}) + \ln(1/\delta))\,,$$

we have

$$R^{0-1}_{\mathbb{P}}(f_{N,m,\lambda}) - R^{0-1}_{\mathbb{P}}(f^*_{\mathbb{P}}) \leq C_{c_1,c_2,V,R} m^{-\frac{c_2}{2+c_2}} \left((\ln(1/\delta) + \ln(m))\right)\,,$$

with probability $1 - 4\delta$. When the spectrum of Σ_ϕ decays sub-exponentially, by choosing

$$\lambda = 1/m$$
$$N = 25C_{d,c_4} \ln^d(m)(\ln(80C_{d,c_4} \ln^d(m)) + \ln(1/\delta))\,,$$

we have

$$R^{0-1}_{\mathbb{P}}(f_{N,m,\lambda}) - R^{0-1}_{\mathbb{P}}(f^*_{\mathbb{P}}) \leq C_{c_3,c_4,d,R,V} \frac{1}{m} \left(\log^{d+2}(m) + \log(1/\delta)\right)\,,$$

with probability $1 - 4\delta$ when $m \geq \exp((c_4 \vee \frac{1}{c_4})d^2/2)$.

This theorem characterizes the learning rate of RFSVM in realizable cases; that is, when the Bayes classifier belongs to the RKHS of the feature map. For polynomially decaying spectrum, when $c_2 > 2$, we get a learning rate faster than $1/\sqrt{m}$. Rudi and Rosasco [2017] obtained a similar fast learning rate for kernel ridge regression with random features (RFKRR), assuming polynomial decay of the spectrum of Σ_ϕ and the existence of a minimizer of the risk in $\mathcal{F}_\phi$. Our theorem extends their result to classification problems and exponential decay spectrum. However, we have to use a stronger assumption that $f^*_{\mathbb{P}} \in \mathcal{F}_\phi$ so that the low noise condition can be applied to derive the variance condition. For RFKRR, the rate faster than $O(1/\sqrt{m})$ will be achieved whenever $c_2 > 1$, and the number of features required is only square root of our result. We think that this is mainly caused by the fact that their surrogate loss is squared. The result for the sub-exponentially decaying spectrum is not investigated for RFKRR, so we cannot make a comparison. We believe that this is the first result showing that RFSVM can achieve $\tilde{O}(1/m)$ with only $\tilde{O}(\ln^d(m))$ features. Note however that when d is large, the sub-exponential case requires a large number of samples, even possibly larger than the polynomial case. This is clearly an artifact of our analysis since we can always use the polynomial case to provide an upper bound! We therefore suspect that there is considerable room for improving our analysis of high dimensional data in the sub-exponential decay case. In particular, removing the exponential dependence on d under reasonable assumptions is an interesting direction for future work.

To remove the realizability assumption, we provide our second theorem, on the learning rate of RFSVM in unrealizable case. We focus on the random features corresponding to the Gaussian kernel as introduced in Section 2. When the Bayes classifier does not belong to the RKHS, we need an approximation theorem to estimate the gap of risks. The approximation property of RKHS of Gaussian kernel has been studied in Steinwart and Christmann [2008], where the margin noise exponent is defined to derive the risk gap. Here we introduce the simpler and stronger separation condition, which leads to a strong result.

The points in $\mathcal{X}$ can be collected in to two sets according to their labels as follows,

$$\mathcal{X}_1 := \{x \in \mathcal{X} \mid \mathbb{E}(y \mid x) > 0\}$$
$$\mathcal{X}_{-1} := \{x \in \mathcal{X} \mid \mathbb{E}(y \mid x) < 0\}.$$

The distance of a point $x \in \mathcal{X}_i$ to the set $\mathcal{X}_{-i}$ is denoted by $\Delta(x)$.

Assumption 3. *We say that the data distribution satisfies a separation condition if there exists $\tau > 0$ such that $\mathbb{P}_{\mathcal{X}}(\Delta(x) < \tau) = 0$.*

Intuitively, Assumption 3 requires the two classes to be far apart from each other almost surely. This separation assumption is an extreme case when the margin noise exponent goes to infinity.

The separation condition characterizes a different aspect of data distribution from Massart's low noise condition. Massart's low noise condition guarantees that the random samples represent the distribution behind them accurately, while the separation condition guarantees the existence of a smooth, in the sense of small derivatives, function achieving the same risk with the Bayes classifier.

With both assumptions imposed on $\mathbb{P}$, we can get a fast learning rate of $\ln^{2d+1} m/m$ with only $\ln^{2d}(m)$ random features, as stated in the following theorem.

Theorem 2. *Assume that $\mathcal{X}$ is bounded by radius ρ. The data distribution has density function upper bounded by a constant B, and satisfies Assumption 1 and 3. Then by choosing*

$$\lambda = 1/m \quad \gamma = \tau/\sqrt{\ln m} \quad N = C_{\tau,d,\rho} \ln^{2d} m(\ln\ln m + \ln(1/\delta)),$$

the RFSVM using an optimized feature map corresponding to the Gaussian kernel with bandwidth γ achieves the learning rate

$$R^{0-1}_{\mathbb{P}}(f_{N,m,\lambda}) - R^{0-1}_{\mathbb{P}}(f^*_{\mathbb{P}}) \leq C_{\tau,V,d,\rho,B} \frac{\ln^{2d+1}(m)(\ln\ln(m) + \ln(1/\delta))}{m},$$

with probability greater than $1 - 4\delta$ for $m \geq m_0$, where m_0 depends on τ, ρ, d.

To the best of our knowledge, this is the first theorem on the fast learning rate of random features method in the unrealizable case. It only assumes that the data distribution satisfies low noise and separation conditions, and shows that with an optimized feature distribution, the learning rate of

$\tilde{O}(1/m)$ can be achieved using only $\ln^{2d+1}(m) \ll m$ features. This justifies the benefit of using RFSVM in binary classification problems. The assumption of a bounded data set and a bounded distribution density function can be dropped if we assume that the probability density function is upper bounded by $C\exp(-\gamma^2\|x\|^2/2)$, which suffices to provide the sub-exponential decay of spectrum of Σ_ϕ. But we prefer the simpler form of the results under current conditions. We speculate that the conclusion of Theorem 2 can be generalized to all sub-Gaussian data.

The main drawback of our two theorems is the assumption of an optimized feature distribution, which is hard to obtain in practice. Developing a data-dependent feature selection method is therefore an important problem for future work on RFSVM. Bach [2017] proposed an algorithm to approximate the optimized feature map from any feature map. Adapted to our setup, the reweighted feature selection algorithm is described as follows.

1. Select M i.i.d. random vectors $\{\omega_i\}_{i=1}^M$ according to the distribution $d\nu_\gamma$.
2. Select L data points $\{x_i\}_{i=1}^L$ uniformly from the training set.
3. Generate the matrix Φ with columns $\phi_M(x_i)/\sqrt{L}$.
4. Compute $\{r_i\}_{i=1}^M$, the diagonal of $\Phi\Phi^\intercal(\Phi\Phi^\intercal + \mu I)^{-1}$.
5. Resample N features from $\{\omega_i\}_{i=1}^M$ according to the probability distribution $p_i = r_i/\sum r_i$.

The theoretical guarantees of this algorithm have not been discussed in the literature. A result in this direction will be extremely useful for guiding practioners. However, it is outside the scope of our work. Instead, here we implement it in our experiment and empirically compare the performance of RFSVM using this reweighted feature selection method to the performance of RFSVM without this preprocessing step; see Section 4.

For the realizable case, if we drop the assumption of optimized feature map, only weak results can be obtained for the learning rate and the number of features required (see E for more details). In particular, we can only show that $1/\epsilon^2$ random features are sufficient to guarantee the learning rate less than ϵ when $1/\epsilon^3$ samples are available. Though not helpful for justifying the computational benefit of random features method, this result matches the parallel result for RFKRR in Rudi and Rosasco [2017] and the approximation result in Sriperumbudur and Szabo [2015]. We conjecture that this upper bound is also optimal for RFSVM.

Rudi and Rosasco [2017] also compared the performance of RFKRR with Nystrom method, which is the other popular method to scale kernel ridge regression to large data sets.We do not find any theoretical guarantees on the fast learning rate of SVM with Nystrom method on classification problems in the literature, though there are several works on its approximation quality to the accurate model and its empirical performance (see Yang et al. [2012], Zhang et al. [2012]). The tools used in this paper should also work for learning rate analysis of SVM using Nystrom method. We leave this analysis to the future.

4 Experimental Results

In this section we evaluate the performance of RFSVM with the reweighted feature selection algorithm[2]. The sample points shown in Figure 3 are generated from either the inner circle or outer annulus uniformly with equal probability, where the radius of the inner circle is 0.9, and the radius of the outer annulus ranges from 1.1 to 2. The points from the inner circle are labeled by -1 with probability 0.9, while the points from the outer annulus are labeled by 1 with probability 0.9. In such a simple case, the unit circle describes the Bayes classifier.

First, we compared the performance of RFSVM with that of KSVM on the training set with 1000 samples, over a large range of regularization parameter ($-7 \le \log\lambda \le 1$). The bandwidth parameter γ is fixed to be an estimate of the average distance among the training samples. After training, models are tested on a large testing set ($> 10^5$). For RFSVM, we considered the effect of the number of features by setting N to be $1, 3, 5, 10$ and 20, respectively. Moreover, both feature selection methods, simple random feature selection (labeled by 'unif' in the figures), which does not apply any preprocess on drawing features, and reweighted feature selection (labeled by 'opt' in the figures) are

[2]The source code is available at `https://github.com/syitong/randfourier`.

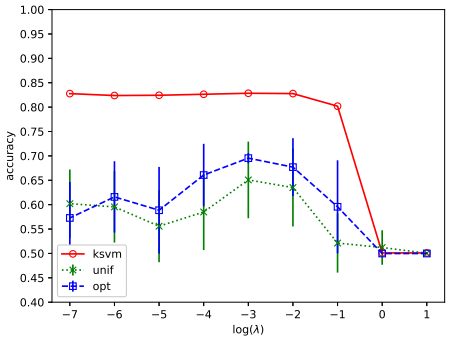

Figure 1: RFSVM with 5 features.

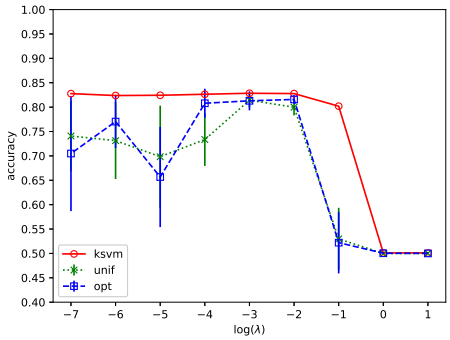

Figure 2: RFSVM with 20 features.

"ksvm" is for KSVM with Gaussian kernel, "unif" is for RFSVM with direct feature sampling, and "opt" is for RFSVM with reweighted feature sampling. Error bars represent standard deviation over 10 runs.

inspected. For the reweighted method, we set $M = 100N$ and $L = 0.3m$ to compute the weight of each feature. Every RFSVM is run 10 times, and the average accuracy and standard deviation are presented.

The results of KSVM, RFSVMs with 1 and 20 features are shown in Figure 1 and Figure 2 respectively (see the results of other levels of features in Appendix F in the supplementary material). The performance of RFSVM is slightly worse than the KSVM, but improves as the number of features increases. It also performs better when the reweighted method is applied to generate features.

To further compare the performance of simple feature selection and reweighted feature selection methods, we plot the learning rate of RFSVM with $O(\ln^2(m))$ features and the best λs for each sample size m. KSVM is not included here since it is too slow on training sets of size larger than 10^4 in our experiment compared to RFSVM. The error rate in Figure 4 is the excess risk between learned classifiers and the Bayes classifier. We can see that the excess risk decays as m increases, and the RFSVM using reweighted feature selection method outperforms the simple feature selection.

According to Theorem 2, the benefit brought by optimized feature map, that is, the fast learning rate, will show up when the sample size is greater than $O(\exp(d))$ (see Appendix D). The number

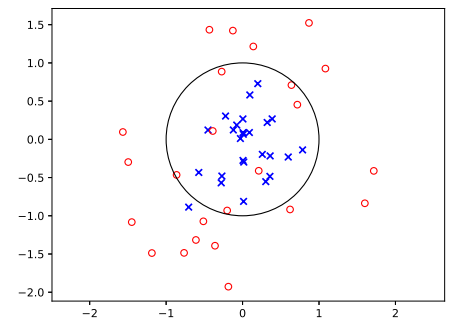

Figure 3: Distribution of Training Samples.

50 points are shown in the graph. Blue crosses represent the points labeled by -1, and red circles the points labeled by 1. The unit circle is one of the best classifier for these data with 90% accuracy.

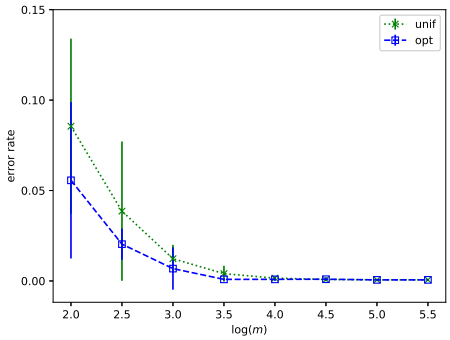

Figure 4: Learning Rate of RFSVMs.

The excess risks of RFSVMs with the simple random feature selection ("unif") and the reweighted feature selection ("opt") are shown for different sample sizes. The error rate is the excess risk. The error bars represent the standard deviation over 10 runs.

of random features required also depends on d, the dimension of data. For data of small dimension and large sample size, as in our experiment, it is not a problem. However, in applications of image recognition, the dimension of the data is usually very large and it is hard for our theorem to explain the performance of RFSVM. On the other hand, if we do not pursue the fast learning rate, the analysis for general feature maps, not necessarily optimized, gives a learning rate of $O(m^{-1/3})$ with $O(m^{2/3})$ random features, which does not depend on the dimension of data (see Appendix E). Actually, for high dimensional data, there is barely any improvement in the performance of RFSVM by using reweighted feature selection method (see Appendix F). It is important to understand the role of d to fully understand the power of random features method.

5 Conclusion

Our study proves that the fast learning rate is possible for RFSVM in both realizable and unrealizable scenarios when the optimized feature map is available. In particular, the number of features required is far less than the sample size, which implies considerably faster training and testing using the random features method. Moreover, we show in the experiments that even though we can only approximate the optimized feature distribution using the reweighted feature selection method, it, indeed, has better performance than the simple random feature selection. Considering that such a reweighted method does not rely on the label distribution at all, it will be useful in learning scenarios where multiple classification problems share the same features but differ in the class labels. We believe that a theoretical guarantee of the performance of the reweighted feature selection method and properly understanding the dependence on the dimensionality of data are interesting directions for future work.

Acknowledgements

AT acknowledges the support of a Sloan Research Fellowship.

ACG acknowledges the support of a Simons Foundation Fellowship.

References

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017.

Corinna Cortes, Mehryar Mohri, and Ameet Talwalkar. On the impact of kernel approximation on learning accuracy. *Journal of Machine Learning Research*, 9:113–120, 2010. ISSN 1532-4435.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.

Bo Dai, Bo Xie, Niao He, Yingyu Liang, Anant Raj, Maria-Florina F Balcan, and Le Song. Scalable kernel methods via doubly stochastic gradients. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3041–3049. Curran Associates, Inc., 2014.

Moulines Eric, Francis R Bach, and Zaïd Harchaoui. Testing for homogeneity with kernel Fisher discriminant analysis. In *Advances in Neural Information Processing Systems*, pages 609–616, 2008.

Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathiya Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear svm. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 408–415, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390208.

P. S. Huang, H. Avron, T. N. Sainath, V. Sindhwani, and B. Ramabhadran. Kernel methods match deep neural networks on timit. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 205–209, May 2014. doi: 10.1109/ICASSP.2014.6853587.

Vladimir. Koltchinskii, SpringerLink (Online service), and École d'Été de Probabilités de Saint-Flour. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems École d'Été*

de Probabilités de Saint-Flour XXXVIII-2008. Lecture Notes in Mathematics,0075-8434 ;2033. Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2011.

P.D. Lax. *Functional analysis*. Pure and applied mathematics. Wiley, 2002. ISBN 9780471556046. URL https://books.google.com/books?id=-jbvAAAAMAAJ.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1313–1320. Curran Associates, Inc., 2009.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3218–3228, 2017.

Clint Scovel, Don Hush, Ingo Steinwart, and James Theiler. Radial kernels and their reproducing kernel hilbert spaces. *Journal of Complexity*, 26(6):641–660, 2010.

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: primal estimated sub-gradient solver for svm. *Mathematical Programming*, 127(1):3–30, 2011. ISSN 1436-4646. doi: 10.1007/s10107-010-0420-4.

Bharath Sriperumbudur and Zoltan Szabo. Optimal rates for random fourier features. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1144–1152. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5740-optimal-rates-for-random-fourier-features.pdf.

I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer New York, 2008. ISBN 9780387772424.

Dougal J. Sutherland and Jeff G. Schneider. On the error of random fourier features. *CoRR*, abs/1506.02785, 2015.

Harold Widom. Asymptotic behavior of the eigenvalues of certain integral equations. *Transactions of the American Mathematical Society*, 109(2):278–295, 1963. ISSN 00029947. URL http://www.jstor.org/stable/1993907.

Tianbao Yang, Yu-feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 476–484. Curran Associates, Inc., 2012.

Kai Zhang, Liang Lan, Zhuang Wang, and Fabian Moerchen. Scaling up kernel svm on limited resources: A low-rank linearization approach. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1425–1434, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL http://proceedings.mlr.press/v22/zhang12d.html.

A Examples of Optimized Feature Maps

Assume that a feature map $\phi : (X) \to L^2(\omega, \Omega, \nu)$ satisfies that $\phi(\omega; x)$ is bounded for all ω and x. We can always convert it to an optimized feature map using the method proposed by Bach [2017]. We rephrase it using our notation as follows.

Define

$$p(\omega) = \frac{\|(\Sigma + \mu I)^{-1/2}\phi(\cdot;\omega)\|^2_{L^2(\mathcal{X},\mathbb{P})}}{\int_\Omega \|(\Sigma + \mu I)^{-1/2}\phi(\cdot;\omega)\|^2_{L^2(\mathcal{X},\mathbb{P})}\,\mathrm{d}\nu(\omega)}\,. \tag{2}$$

Since ϕ is bounded, its L^2 norm is finite. The function p defined above is a probability density function with respect to ν, and we denote the new probability measure by μ. Then the new feature map is given by $\widetilde{(\phi)}(\omega; x) = \phi(\omega; x)/\sqrt{p(\omega)}$ together with the measure μ. With $\tilde{\phi}$, we have

$$\sup_{\omega\in\Omega}\left\|(\Sigma + \mu I)^{-1/2}\tilde{\phi}(\cdot;\omega)\right\|^2 = \sup_{\omega\in\Omega}\frac{\left\|(\Sigma + \mu I)^{-1/2}\phi(\cdot;\omega)\right\|^2}{p(\omega)} \tag{3}$$

$$= \int_\Omega \|(\Sigma + \mu I)^{-1/2}\phi(\cdot;\omega)\|^2_{L^2(\mathcal{X},\mathbb{P})}\,\mathrm{d}\nu(\omega) \tag{4}$$

$$= \mathrm{tr}(\Sigma(\Sigma + \mu I)^{-1})\,. \tag{5}$$

When the feature map is constructed mapping into $L^2(\mathcal{X}, \mathbb{P})$ as described in Section 2, it is optimized. Indeed, we can compute

$$\sup_{\omega\in\mathcal{X}}\left\|(\Sigma + \mu I)^{-1/2}\phi(\cdot;\omega)\right\|^2 = \sup_{\omega\in\mathcal{X}}\left\|\sum_{i=1}^{\infty}\frac{\sqrt{\lambda_i}}{\sqrt{\lambda_i + \mu}}e_i(\cdot)\right\|^2 \tag{6}$$

$$= \sum_{i=1}^{\infty}\frac{\lambda_i}{\lambda_i + \mu}\,. \tag{7}$$

As an example for this type of feature map, we can consider $\{e_i\}$ to be the Walsh system, which is an orthonormal basis for $L^2([0, 1])$. Any Bayes classifier with finitely many discontinuities and discontinuous only at dyadic, namely points expressable by finite bits, points, will be a finite linear combination of Walsh basis. This guarantees that the assumptions in Theorem 1 can be satisfied. Our first experiment also make use of this construction.

The construction above is inspired by the use of spline kernel in Rudi and Rosasco [2017]. However, our situation is more complicated since the target function, Bayes classifier, is discontinuous. While the functions in the RKHS generated by the spline kernel must be continuous (Cucker and Smale [2002]). Though we can construct Bayes classifier using the Walsh basis, we have yet to understand the variety of possible Bayes classifiers in such a space.

B Local Rademacher Complexity of RFSVM

Before the proofs, we first briefly summarize the use of each lemmas and theorems. Theorem 3 and 4 are two fundamental external results for our proof. Lemma 7 and 8 refine results that appeared in previous works, so that we can apply them to our case. Lemma 3, 4 and 5 are the key results to establish fast rate for RFSVM, parallel to Steinwarts' work for KSVM. All other smaller and simpler lemmas included in the appendices are for the purposes of clarity and completeness. The proofs are not hard but quite technical.

First, both of our theorems are consequences of the following fundamental theorem.

Theorem 3. *(Theorem 7.20 in Steinwart and Christmann [2008]) For a RKHS $\mathcal{F}$, denote $\inf_{f\in\mathcal{F}} R^1_{\mathbb{P},\lambda}(f) - R^*$ by r^*. For $r > r^*$, consider the following function classes*

$$\mathcal{F}_r := \{f \in \mathcal{F} \mid R^1_{\mathbb{P},\lambda}(f) - R^* \le r\}$$

and

$$\mathcal{H}_r := \{\ell^1 \circ f - \ell^1 \circ f^*_{\mathbb{P}} \mid f \in \mathcal{F}_r\}\,.$$

Assume that there exists $V \geq 1$ such that for any $f \in \mathcal{F}$,

$$\mathbb{E}_{\mathbb{P}}(\ell^1 \circ f - \ell^1 \circ f_{\mathbb{P}}^*)^2 \leq V(R_{\mathbb{P}}^1(f) - R^*)\,.$$

If there is a function $\varphi_m : [0,\infty) \to [0,\infty)$ such that $\varphi_m(4r) \leq 2\varphi_m(r)$ and $\mathfrak{R}_m(\mathcal{H}_r) \leq \varphi_m(r)$ for all $r \geq r^*$, Then, for any $\delta \in (0,1]$, $f_0 \in \mathcal{F}$ with $\|\ell^{\text{hinge}} \circ f_0\|_\infty \leq B_0$, and

$$r > \max\left\{30\varphi_m(r), \frac{72V\ln(1/\delta)}{m}, \frac{5B_0\ln(1/\delta)}{m}, r^*\right\},$$

we have

$$R_{\mathbb{P},\lambda}^1(f_{m,N,\lambda}) - R^* \leq 6\left(R_{\mathbb{P},\lambda}^h(f_0) - R^*\right) + 3r$$

with probability greater than $1 - 3\delta$.

To establish the fast rate of RFSVM using the theorem above, we must understand the local Rademacher complexity of RFSVM: that is, find a formula for $\varphi_m(r)$. B_0, r^* and f_0 are only related with the approximation error, and we leave the discussion of them to next sections. The variance condition Equation 1 is satisfied under Assumption 1. With this variance condition, we can upper bound the Rademacher complexity of RFSVM in terms of number of features and regularization parameter. It is particularly important to have $1/\lambda$ inside the logarithm function.

First, we will need the summation version of Dudley's inequality using entropy number defined below, instead of covering number.

Definition 2. *For a semi-normed space $(E, \|\cdot\|)$, we define its (dyadic) entropy number by*

$$e_n(E, \|\cdot\|) := \inf\left\{\varepsilon > 0 : \exists s_1, \ldots, s_{2^{n-1}} \in B^1 \text{ s.t. } B^1 \subset \bigcup_{i=1}^{2^{n-1}} B(s_i, \varepsilon)\right\},$$

where B^1 is the unit ball in E and $B(a,r)$ is the ball with center at a and radius r.

To take off the loss function from the hypothesis class, we have the following lemma. $\|\cdot\|_{L_2(D)}$ is the semi-norm defined by $\|\cdot\|_{L_2(D)} := (\frac{1}{m}\sum_i f^2(x_i))^{1/2}$.

Lemma 1. $e_i(\mathcal{H}_r, \|\cdot\|_{L_2(D)}) \leq e_i(\mathcal{F}_r, \|\cdot\|_{L_2(D)})$

Proof. Assume that T is an ϵ-covering over $\mathcal{F}_r$ with $|T| = 2^i$. By definition $\epsilon \geq e_i(\mathcal{F}_r, \|\cdot\|_{L_2(D)})$. Then $T' = \ell^1 \circ T - \ell^1 \circ f_{\mathbb{P}}^*$ is a covering over $\mathcal{H}_r$. For any f and g in $\mathcal{F}_r$,

$$\left\|\ell^1 \circ f - \ell^1 \circ g\right\|_{L_2(D)} \leq 1 \cdot \|f - g\|_{L_2(D)}\,,$$

because ℓ^1 is 1-Lipschitz. And hence the radius of the image of an ϵ-ball under ℓ^1 is less than ϵ. Therefore $\ell^1 \circ T - \ell^1 \circ f_{\mathbb{P}}^*$ is an ϵ-covering over $\mathcal{H}_r$ with cardinatily 2^i and $\epsilon \leq e_i(\mathcal{F}_r, \|\cdot\|_{L_2(D)})$. By taking infimum over the radius of all such T and T', the statement is proved. □

Now we need to give an upper bound for the entropy number of $\mathcal{F}_r$ with semi-norm $\|\cdot\|_{L_2(D)}$ using a volumetric estimate.

Lemma 2. $e_i(\mathcal{F}_r, \|\cdot\|_{L_2(D)}) \leq 3(2r/\lambda)^{1/2}2^{-i/2N}$.

Proof. Since $\mathcal{F}$ consists of functions

$$f(x) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} w_{c_i}\cos\left(\frac{\omega_i \cdot x}{\gamma}\right) + w_{s_i}\sin\left(\frac{\omega_i \cdot x}{\gamma}\right),$$

under the semi-norm $\|\cdot\|_{L_2(D)}$ it is isometric with the $2N$-dimensional subspace U of $\mathbb{R}^m$ spanned by the vectors

$$\left\{\left[\cos\left(\frac{\omega_i \cdot x_1}{\gamma}\right), \ldots, \cos\left(\frac{\omega_i \cdot x_m}{\gamma}\right)\right]^\intercal, \left[\sin\left(\frac{\omega_i \cdot x_1}{\gamma}\right), \ldots, \sin\left(\frac{\omega_i \cdot x_m}{\gamma}\right)\right]^\intercal\right\}_{i=1}^{N}$$

for fixed m samples. For each $f \in \mathcal{F}_r$, we have $R^1_{\mathbb{P},\lambda}(f) - R^* \leq r$, which implies that $\|f\|_{\mathcal{F}} \leq (2r/\lambda)^{1/2}$. By the property of RKHS, we get

$$|f(x)| \leq \|f\|_{\mathcal{F}} \|k(x,\cdot)\|_{\mathcal{F}} \leq \left(\frac{2r}{\lambda}\right)^{1/2} \cdot 1 \,,$$

where we use the fact that $k(x,\cdot)$ is the evaluation functional in the RKHS.

Denote the isomorphism from $\mathcal{F}$ (modulo the equivalent class under the semi-norm) to U by I. Then we have

$$I(\mathcal{F}_r) \subset B^m_\infty \left(\left(\frac{2r}{m\lambda}\right)^{1/2} \right) \cap U \subset B^m_2 \left(\left(\frac{2r}{\lambda}\right)^{1/2} \right) \cap U \,.$$

The intersection region can be identified as a ball of radius $(2r/\lambda)^{1/2}$ in $\mathbb{R}^{2N}$. Its entropy number by volumetric estimate is given by

$$e_i \left(B^{2N}_2 \left(\left(\frac{2r}{\lambda}\right)^{1/2} \right), \|\cdot\|_2 \right) \leq 3 \left(\frac{2r}{\lambda}\right)^{1/2} 2^{-\frac{i}{2N}} \,.$$

□

With the lemmas above, we can get an upper bound on the entropy number of $\mathcal{H}_r$. However, we should note that such an upper bound is not the best when i is small. Because the ramp loss ℓ^1 is bounded by 2, the radius of $\mathcal{H}_r$ with respect to $\|\cdot\|_{L_2(D)}$ is bounded by 1, which is irrelevant with r/λ. This observation will give us finer control on the Rademacher complexity.

Lemma 3. *Assume that $\lambda < 1/2$. Then*

$$\mathfrak{R}_D(\mathcal{H}_r) \leq \sqrt{\frac{(\ln 16) N \log_2 1/\lambda}{m}} \left(3\sqrt{2}\rho + 18\sqrt{r} \right) \,,$$

where $\rho = \sup_{h \in \mathcal{H}_r} \|h\|_{L_2(D)}$.

Proof. By Theorem 7.13 in Steinwart and Christmann [2008], we have

$$\mathfrak{R}_D(\mathcal{H}_r) \leq \sqrt{\frac{\ln 16}{m}} \left(\sum_{i=1}^{\infty} 2^{i/2} e_{2^i}(\mathcal{H}_r \cup \{0\}, \|\cdot\|_{L_2(D)}) + \sup_{h \in \mathcal{H}_r} \|h\|_{L_2(D)} \right) \,.$$

It is easy to see that $e_i(\mathcal{H}_r \cup \{0\}) \leq e_{i-1}(\mathcal{H}_r)$ and $e_0(\mathcal{H}_r) \leq \sup_{h \in \mathcal{H}_r} \|h\|_{L_2(D)}$. Since $e_i(\mathcal{H}_r)$ is a decreasing sequence with respect to i, together with the lemma above, we know that

$$e_i(\mathcal{H}_r) \leq \min \left\{ \sup_{h \in \mathcal{H}_r} \|h\|_{L_2(D)}, 3 \left(\frac{2r}{\lambda}\right)^{1/2} 2^{-\frac{i}{2N}} \right\} \,.$$

Even though the second one decays exponentially, it may be much greater than the first term when $2r/\lambda$ is huge for small is. To achieve the balance between these two bounds, we use the first one for first T terms in the sum and the second one for the tail. So

$$\mathfrak{R}_D(\mathcal{H}_r) \leq \sqrt{\frac{\ln 16}{m}} \left(\sup_{h \in \mathcal{H}_r} \|h\|_{L_2(D)} \sum_{i=0}^{T-1} 2^{i/2} + 3 \left(\frac{2r}{\lambda}\right)^{1/2} \sum_{i=T}^{\infty} 2^{i/2} 2^{-\frac{2^i - 1}{2N}} \right) \,.$$

The first sum is $\frac{\sqrt{2}^T - 1}{\sqrt{2} - 1}$. When T is large enough, the second sum is upper bounded by the integral

$$\int_{T-1}^{\infty} 2^{x/2} 2^{-2^x - 1/2N} \, \mathrm{d}x \leq \frac{6N}{2^{T/2}} \cdot 2^{-\frac{2^T}{4N}} \,.$$

To make the form simpler, we bound $\frac{\sqrt{2}^T - 1}{\sqrt{2} - 1}$ by $3 \cdot 2^{T/2}$, and denote $\sup_{h \in \mathcal{H}_r} \|h\|_{L_2(D)}$ by ρ. Taking T to be

$$\log_2 \left(2N \log_2 \left(\frac{1}{\lambda} \right) \right) \,,$$

we get the upper bound of the form

$$\mathfrak{R}_D(\mathcal{H}_r) \le \sqrt{\frac{\ln 16}{m}} \left(3\rho \sqrt{2N \log_2 \frac{1}{\lambda}} + \frac{18\sqrt{Nr}}{\log_2(1/\lambda)} \right) ,$$

When $\lambda < 1/2$, $\log_2 1/\lambda > 1$, so we can further enlarge the upper bound to the form

$$\mathfrak{R}_D(\mathcal{H}_r) \le \sqrt{\frac{(\ln 16) N \log_2 1/\lambda}{m}} \left(3\sqrt{2}\rho + 18\sqrt{r} \right) ,$$

□

Next lemma analyzes the expected Rademacher complexity for $\mathcal{H}_r$.

Lemma 4. *Assume $\lambda < 1/2$ and $\mathbb{E}h^2(x,y) \le V\mathbb{E}h(x,y)$. Then*

$$\mathfrak{R}_m(\mathcal{H}_r) \le C_1 \sqrt{\frac{N(V+1)\log_2(1/\lambda)}{m}} \sqrt{r} + C_2 \frac{N \log_2(1/\lambda)}{m} .$$

Proof. With Lemma 3, we can directly compute the upper bound for $\mathfrak{R}_m(\mathcal{H}_r)$ by taking expectation over $D \sim \mathbb{P}^m$.

$$\begin{aligned} \mathfrak{R}_m(\mathcal{H}_r) &= \mathbb{E}_{D\sim\mathbb{P}^m} \mathfrak{R}_D(\mathcal{H}_r) \\ &\le \sqrt{\frac{(\ln 16) N \log_2 1/\lambda}{m}} \left(3\sqrt{2}\mathbb{E} \sup_{h\in\mathcal{H}_r} \|h\|_{L_2(D)} + 18\sqrt{r} \right) . \end{aligned}$$

By Jensen's inequality and A.8.5 in Steinwart and Christmann [2008], we have

$$\begin{aligned} \mathbb{E} \sup_{h\in\mathcal{H}_r} \|h\|_{L_2(D)} &\le \left(\mathbb{E} \sup_{h\in\mathcal{H}_r} \|h\|^2_{L_2(D)} \right)^{1/2} \\ &\le \left(\mathbb{E} \sup_{h\in\mathcal{H}_r} \frac{1}{m} \sum_{i=1}^m h^2(x_i,y_i) \right)^{1/2} \\ &\le \left(\sigma^2 + 8\mathfrak{R}_m(\mathcal{H}_r) \right)^{1/2} , \end{aligned}$$

where $\sigma^2 := \mathbb{E}h^2$. When $\sigma^2 > \mathfrak{R}_m(\mathcal{H}_r)$, we have

$$\begin{aligned} \mathfrak{R}_m(\mathcal{H}_r) &\le \sqrt{\frac{(\ln 16) N \log_2(1/\lambda)}{m}} \left(9\sqrt{2}\sigma + 18\sqrt{r} \right) \\ &\le \sqrt{\frac{(\ln 16) N \log_2(1/\lambda)}{m}} \left(9\sqrt{2}\sqrt{Vr} + 18\sqrt{r} \right) \\ &\le 36 \sqrt{\frac{2(\ln 16) N(V+1)\log_2(1/\lambda)}{m}} \sqrt{r} . \end{aligned}$$

The second inequality is because $\mathbb{E}h^2 \le V\mathbb{E}h$ and $\mathbb{E}h \le r$ for $h \in \mathcal{H}_r$.

When $\sigma^2 \le \mathfrak{R}_m(\mathcal{H}_r)$, we have

$$\begin{aligned} \mathfrak{R}_m(\mathcal{H}_r) &\le \sqrt{\frac{(\ln 16) N \log_2(1/\lambda)}{m}} \left(9\sqrt{2}\sqrt{\mathfrak{R}_m(\mathcal{H}_r)} + 18\sqrt{r} \right) \\ &\le 36 \sqrt{\frac{(\ln 16) N \log_2(1/\lambda)}{m}} \sqrt{r} + 36^2 \frac{(\ln 16) N \log_2(1/\lambda)}{m} . \end{aligned}$$

The last inequality can be obtained by dividing the formula into two cases, either $\mathfrak{R}_m(\mathcal{H}_r) < r$ or $\mathfrak{R}_m(\mathcal{H}_r) \ge r$ and then take the sum of the upper bounds of two cases.

Combining all these inequalities, we finally obtain an upper bound

$$\mathfrak{R}_m(\mathcal{H}_r) \le C_1 \sqrt{\frac{(V+1) N \log_2(1/\lambda)}{m}} \sqrt{r} + C_2 \frac{N \log_2(1/\lambda)}{m} ,$$

where C_1 and C_2 are two absolute constants. □

The last lemma gives the explicit formula of $\varphi_m(r)$. Now we can get the formula for r.

Lemma 5. *When*

$$r = (900C_1^2 + 120C_2)N(V+1)\frac{\ln(1/\lambda)}{m} + (5B_0 + 72V)\frac{\ln(1/\delta)}{m} \tag{8}$$

we have

$$r \geq \max\{30\varphi_m(r), \frac{72V\ln(1/\delta)}{m}, \frac{5B_0\ln(1/\delta)}{m}\}. \tag{9}$$

It can be check by simply plugging r into $\varphi_m(r)$.

C Proof of Theorem 1

With Theorem 3 and Lemma 5, we are almost done with the proof of Theorem 1. The only missing part is an upper bound of the approximation error $R^h_{\mathbb{P},\lambda}(f_0) - R^*$. This upper bound has been established in Proposition 1 in Bach [2017]. We rephrase it as below.

Theorem 4. *(Proposition 1 of Bach [2017]) Assume that ϕ is an optimized feature map and f belongs to the RKHS $\mathcal{F}$ of ϕ. For $\delta > 0$, when*

$$N \geq 5d(\mu)\log\left(\frac{16d(\mu)}{\delta}\right), \tag{10}$$

there exists $\beta \in \mathbb{R}^N$ with norm less than 2, such that

$$\sup_{\|f\|_{\mathcal{F}}\leq 1} \|f - \beta\cdot\phi_N(\cdot)\|_{L^2(\mathcal{X},\mathbb{P})} \leq 2\sqrt{\mu}\,, \tag{11}$$

with probability greater than $1-\delta$.

Now we prove two simple lemmas connecting the decay rate of Σ to the magnitude of $d(\mu)$.

Lemma 6. *If $\lambda_i(\Sigma) \leq c_1 i^{-c_2}$, where $c_2 > 1$, we have*

$$d(\mu) \leq \frac{2c_2}{c_2-1}\left(\frac{c_1}{\mu}\right)^{1/c_2}, \tag{12}$$

for $\mu < c_1$.

If $\lambda_i(\Sigma) \leq c_3\exp(-c_4 i^{1/d})$, we have

$$d(\mu) \leq 5c_4^{-d}\ln^d(c_3/\mu)\,, \tag{13}$$

for $\mu < c_3\exp\left(-\left(c_4 \vee \frac{1}{c_4}\right)d^2\right)$.

Proof. Both results make use the following observation:

$$d(\mu) = \sum_{i=1}^{\infty}\frac{\lambda_i}{\lambda_i+\mu} \leq m_\mu + \frac{1}{\mu}\sum_{m_\mu+1}^{\infty}\lambda_i\,, \tag{14}$$

where $m_\mu = \max\{i : \lambda_i \leq \mu\}|$.

When $\lambda_i \leq c_1 i^{-c_2}$, denote $t_\mu = (c_1/\mu)^{1/c_2}$ and then $m_\mu = \lfloor t_\mu \rfloor$. For the tail part,

$$\frac{1}{\mu}\sum_{m_\mu+1}^{\infty}\lambda_i \leq 1 + \frac{1}{\mu}\int_{t_\mu}^{\infty} c_1 x^{-c_2}\,\mathrm{d}x \tag{15}$$

$$\leq 1 + \frac{1}{c_2-1}\left(\frac{c_1}{\mu}\right)^{\frac{1}{c_2}}. \tag{16}$$

Combining them together, when $c_1/\mu > 1$, the constant 1 can be absorbed by the second term with a coefficient 2.

When $\lambda_i \le c_3 \exp(-c_4 i^{1/d})$, denote $t_\mu = \frac{1}{c_4^d} \ln^d\left(\frac{c_3}{\mu}\right)$, and then $m_\mu = \lfloor t_\mu \rfloor$. For the tail part, we need to discuss different situations.

First, if $d = 1$, then we directly have

$$\frac{1}{\mu} \sum_{m_\mu+1}^{\infty} \frac{\lambda_i}{\lambda_i + \mu} \le \frac{1}{\mu} \left(\mu + \int_{t_\mu}^{\infty} c_3 \exp(-c_4 x)\, \mathrm{d}x \right) \tag{17}$$

$$= 1 + \frac{1}{c_4}\,. \tag{18}$$

When $\mu < c_3 \exp(-(c_4 \vee \frac{1}{c_4}))$, we can combine these terms into $3t_\mu$.

Second, if $d \ge 2$, when $\mu \le c_3 \exp(-c_4 e)$, we have that

$$\exp(-c_4 x^{1/d}) \le \exp(-c_4 \frac{t_\mu^{1/d}}{\ln t_\mu} \ln x) = x^{-c_4 \frac{t_\mu^{1/d}}{\ln t_\mu}}\,. \tag{19}$$

Then,

$$\frac{1}{\mu} \sum_{m_\mu+1}^{\infty} \lambda_i \le 1 + \frac{1}{\mu} \int_{t_\mu}^{\infty} c_3 \exp(-c_4 x^{-1/d})\, \mathrm{d}x \tag{20}$$

$$\le 1 + \frac{c_3}{\mu} \int_{t_\mu}^{\infty} x^{-c_4 \frac{t_\mu^{1/d}}{\ln t_\mu}} \tag{21}$$

$$= 1 + \frac{t_\mu}{c_4 \frac{t_\mu^{1/d}}{\ln t_\mu} - 1}\,. \tag{22}$$

When $c_4 \ge 1$, we may assume that $\mu \le c_3 \exp(-c_4 d^2)$, and then

$$c_4 \frac{t_\mu^{1/d}}{\ln t_\mu} - 1 \ge \frac{c_4 d^2}{2d \ln d} \ge \frac{4}{3}\,. \tag{23}$$

So the upper bound has the form $5t_\mu$.

When $c_4 < 1$, we may assume that $\mu \le c_3 \exp(-d^2/c_4)$, and then

$$c_4 \frac{t_\mu^{1/d}}{\ln t_\mu} - 1 \ge \frac{d^2/c_4}{2d \ln(d/c_4)} \ge \frac{4}{3}\,. \tag{24}$$

So the upper bound also has the form $5t_\mu$. □

Now with all these preparation, we can complete our proof of Theorem 1

Proof. Under the assumption of Theorem 1, $B_0 = 1$ and $r^* = 0$ in Theorem 3. By Lemma 5, we have

$$r = (900C_1^2 + 120C_2)N(V+1)\frac{\ln(1/\lambda)}{m} + (5 + 72V)\frac{\ln(1/\delta)}{m}\,. \tag{25}$$

By Theorem 4, we have

$$R^h_{\mathbb{P},\lambda}(f_0) - R^* \le 2\sqrt{\mu}R + 4R^2\frac{\lambda}{2}\,, \tag{26}$$

with probability $1 - \delta$ when $N \ge 5d(\mu)\log\left(\frac{16d(\mu)}{\delta}\right)$.

When the spectrum of Σ decays polynomially,

$$d(\mu) \le \frac{2c_2}{c_2 - 1}\left(\frac{c_1}{\mu}\right)^{1/c_2}\,. \tag{27}$$

Assume $m > c_1^{-(2+c_2)/(2c_2)}$. By choosing $\mu = c_1 m^{-\frac{2c_2}{2+c_2}} < c_1$ and $\lambda = m^{-c_2/(2+c_2)}$, we have

$$N = 10c_{1,2}m^{\frac{2}{2+c_2}}\left(\ln(32c_{1,2}m^{\frac{2}{2+c_2}}) + \ln(1/\delta)\right), \tag{28}$$

and

$$R^h_{\mathbb{P},\lambda}(f_{m,N,\lambda}) - R^* \leq \frac{12R}{m^{\frac{c_2}{2+c_2}}} + \frac{12R^2}{m^{\frac{c_2}{2+c_2}}} \tag{29}$$

$$+ 30C_{1,2}c_{1,2}(\ln 32c_{1,2} + \frac{2}{2+c_2}\ln m + \ln(1/\delta))(V+1)\frac{c_2}{2+c_2}\frac{\ln m}{m^{\frac{c_2}{2+c_2}}} \tag{30}$$

$$+ \frac{15+216V}{m}\ln(1/\delta)\,, \tag{31}$$

with probability $1-4\delta$, where

$$C_{1,2} = 900C_1^2 + 120C_2, \quad c_{1,2} = \frac{c_2 c_1^{1/c_2}}{c_2 - 1}\,. \tag{32}$$

When the spectrum of Σ decays sub-exponentially,

$$d(\mu) \leq 5c_4^{-d}\ln^d(c_3/\mu)\,. \tag{33}$$

Assume that $m > \exp(-(c_4 \vee \frac{1}{c_4})d^2/2)$. By choosing $\mu = c_3/m^2$ and $\lambda = 1/m$, we have

$$N = 25c_{d,4}\ln^d(m)(\ln(80c_{d,4}\ln^d(m)) + \ln(1/\delta))\,, \tag{34}$$

and

$$R^h_{\mathbb{P},\lambda}(f_{m,N,\lambda}) - R^* \leq \frac{12R\sqrt{c_3}}{m} \tag{35}$$

$$+ \frac{12R^2}{m} + 150C_{1,2}c_{d,4}(\ln 160c_{d,4} + d\ln\ln m + \ln(1/\delta))(V+1)\frac{\ln^{d+1} m}{m} \tag{36}$$

$$+ \frac{15+216V}{m}\ln(1/\delta)\,, \tag{37}$$

with probability $1-4\delta$, where

$$C_{1,2} = 900C_1^2 + 120C_2, \quad c_{d,4} = \left(\frac{2}{c_4}\right)^d\,. \tag{38}$$

□

D Proof of Theorem 2

Theorem 2 requires a further analysis of the approximation error of RKHS to the Bayes classifier. This part adopts Steinwart and Christmann [2008]'s idea of margin noise exponent. We say that the data distribution $\mathbb{P}$ has margin noise exponent $\beta > 0$ if there exists a positive constant c such that

$$\int_{\{} x : \Delta(x) < t\}|y|\mathrm{d}\mathbb{P}(x,y) \leq ct^{-\beta} \quad \forall t \in (0,1)\,. \tag{39}$$

Therefore, infinite β corresponds to our separation condition with $\tau = 1$. However, the original proof of the approximation error that works with the margin noise exponent cannot be generalized to the case of infinite β, because the coefficient $\Gamma(d+\beta)/2^d$ will blow up (see Theorem 8.18 in Steinwart and Christmann [2008]). This issue can be resolved by modifying the original proof, as shown below.

Lemma 7. *Assume that there exists $\tau > 0$ such that*

$$\int_{\{x:\Delta(x)<t\}} |2\eta(x) - 1|\,\mathrm{d}\mathbb{P}_{\mathcal{X}}(x) = 0\,, \forall t < \tau\,, \tag{40}$$

where $\mathcal{X} \subset B^d(\rho)$ and $\eta(x)$ is a version of $\mathbb{P}(y=1|x)$. Then there exists a function f in the RKHS generated by the kernel

$$k_\gamma(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\gamma^2}\right) \tag{41}$$

where $\gamma < \tau/\sqrt{d-1}$ *such that*

$$R^h(f) - R^* < \frac{4\tau^{d-2}}{\Gamma(d/2)} \exp\left(-\frac{\tau^2}{\gamma^2}\right) \gamma^{d-2}\,,$$
$$\|f\|_{\mathcal{F}} \leq \frac{(\sqrt{\pi/2}\rho^2)^{d/2}}{\Gamma(d/2+1)} \gamma^{-d/2}$$

and

$$|f(x)| \leq 1\,. \tag{42}$$

Proof. First we define

$$\mathcal{X}_y := \{x : (2\eta(x)-1)y > 0\} \text{ for } y = \pm 1\,, \tag{43}$$

and $g(x) := (\sqrt{2\pi}\gamma)^{-d/2}\mathrm{sign}(2\eta(x)-1)$. It is square integrable since $\eta(x) = 1/2$ for all $x \notin \mathcal{X}$. Then we map g onto the RKHS by the integral operator determined by k_γ,

$$f(x) := \int_{\mathbb{R}^d} \phi_\gamma(t;x) g(t)\, \mathrm{d}t\,, \tag{44}$$

where

$$\phi_\gamma(t;x) = \left(\frac{2}{\pi\gamma^2}\right)^{d/4} \exp\left(-\frac{\|x-t\|^2}{\gamma^2}\right)\,. \tag{45}$$

Note that it is a special property of Gaussian kernel that the feature map onto $L^2(\mathbb{R}^d)$ also has a Gaussian form. For other type of kernels, we may not have such a convenient characterization.

We know that

$$\|f\|_{\mathcal{H}} = \|g\|_{L^2} \leq \frac{\sqrt{\mathrm{Vol}(B^d(\rho))}}{(\sqrt{2\pi}\gamma)^{d/2}} = \frac{(\sqrt{\pi/2}\rho^2)^{d/2}}{\Gamma(d/2+1)} \gamma^{-d/2}\,. \tag{46}$$

Moreoever,

$$\begin{aligned}
|f(x)| &\leq \int_{\mathbb{R}^d} \phi_\gamma(t;x)(\sqrt{2\pi}\gamma)^{-d/2}\, \mathrm{d}t \\
&= (\pi\gamma^2)^{-d/2} \int_{\mathbb{R}^d} \exp\left(-\frac{\|x-t\|^2}{\gamma^2}\right)\, \mathrm{d}t \\
&= 1\,.
\end{aligned}$$

Since f is uniformly bounded by 1, by Zhang's inequality, we have

$$R^h(f) - R^* = \mathbb{E}_{\mathbb{P}_{\mathcal{X}}}(|f(x) - \mathrm{sign}(2\eta(x)-1)||2\eta(x)-1|)\,. \tag{47}$$

Now we give an upper bound on $|f(x) - \mathrm{sign}(2\eta(x)-1)|$. Assume $x \in \mathcal{X}_1$. Then we know that $f(x) \leq \mathrm{sign}(2\eta(x)-1) = 1$,

$$\begin{aligned}
1 - f(x) &= 1 - \left(\frac{1}{\pi\gamma^2}\right)^{d/2} \int_{\mathbb{R}^d} \exp\left(-\frac{\|x-t\|^2}{\gamma^2}\right) \mathrm{sign}(2\eta(t)-1)\, \mathrm{d}t \\
&= 1 - \left(\frac{1}{\pi\gamma^2}\right)^{d/2} \int_{\mathcal{X}_1} \exp\left(-\frac{\|x-t\|^2}{\gamma^2}\right)\, \mathrm{d}t \\
&\quad + \left(\frac{1}{\pi\gamma^2}\right)^{d/2} \int_{\mathcal{X}_{-1}} \exp\left(-\frac{\|x-t\|^2}{\gamma^2}\right)\, \mathrm{d}t \\
&\leq 2 - 2\left(\frac{1}{\pi\gamma^2}\right)^{d/2} \int_{B(x,\Delta(x))} \exp\left(-\frac{\|x-t\|^2}{\gamma^2}\right)\, \mathrm{d}t \\
&\leq 2 - 2\left(\frac{1}{\pi\gamma^2}\right)^{d/2} \int_{B(0,\Delta(x))} \exp\left(-\frac{\|t\|^2}{\gamma^2}\right)\, \mathrm{d}t \\
&= 2\left(\frac{1}{\pi\gamma^2}\right)^{d/2} \int_{\mathbb{R}^d \setminus B(0,\Delta(x))} \exp\left(-\frac{\|t\|^2}{\gamma^2}\right)\, \mathrm{d}t \\
&= \frac{4}{\Gamma(d/2)\gamma^d} \int_{\Delta(x)}^{\infty} \exp\left(-\frac{r^2}{\gamma^2}\right) r^{d-1}\, \mathrm{d}r\,.
\end{aligned}$$

Here the key is that $B(x, \Delta(x)) \subset \mathcal{X}_1$ when $x \in \mathcal{X}_1$. For $x \in \mathcal{X}_{-1}$, we have the same upper bound for $1 + f(x)$. Therefore, we have

$$\begin{aligned} R^h(f) - R^* &\leq \frac{4}{\Gamma(d/2)\gamma^d} \int_{\mathcal{X}} \int_0^\infty \mathbf{1}_{(\Delta(x),\infty)}(r) \exp\left(-\frac{r^2}{\gamma^2}\right) r^{d-1} |2\eta(x) - 1| \, \mathrm{d}r \mathrm{d}\mathbb{P}_{\mathcal{X}}(x) \\ &= \frac{4}{\Gamma(d/2)\gamma^d} \int_0^\infty \int_{\mathcal{X}} \mathbf{1}_{(0,r)}(\Delta(x)) \exp\left(-\frac{r^2}{\gamma^2}\right) r^{d-1} |2\eta(x) - 1| \, \mathrm{d}\mathbb{P}_{\mathcal{X}}(x) \mathrm{d}r \\ &\leq \frac{4}{\Gamma(d/2)\gamma^d} \int_\tau^\infty \exp\left(-\frac{r^2}{\gamma^2}\right) r^{d-1} \, \mathrm{d}r \end{aligned}$$

To get the last line, we apply the assumption on the expected label clarity. Now we only need to give an estimate of the integral.

$$\int_\tau^\infty \exp\left(-\frac{r^2}{\gamma^2}\right) r^{d-1} \, \mathrm{d}r \leq \int_\tau^\infty C \exp\left(-\alpha \frac{r^2}{\gamma^2}\right) \, \mathrm{d}r \tag{48}$$

where

$$C = \tau^{d-1} \exp(-(d-1)/2) \quad \alpha = 1 - 2\gamma^2 \tau^{-2}(d-1) \, . \tag{49}$$

It is required that $\gamma < \sqrt{2}\tau/\sqrt{d-1}$ so that $\alpha > 0$. And then we can give an upper bound to the excess risk

$$R^h(f) - R^* \leq \frac{4\tau^d}{\Gamma(d/2)(2\tau^2 - (d-1)\gamma^2)} \exp\left(-\frac{\tau^2}{\gamma^2}\right) \gamma^{d-2} \, . \tag{50}$$

If we further require that $\gamma < \tau/\sqrt{d-1}$, then we have a simpler upper bound,

$$\frac{4\tau^{d-2}}{\Gamma(d/2)} \exp\left(-\frac{\tau^2}{\gamma^2}\right) \gamma^{d-2} \, . \tag{51}$$

□

Some remarks on this result:

1. The proof follows almost step by step the proof of Steinwart and Christmann [2008]. The only difference occurs at where we apply our assumption.
2. The approximation error is basically dominated by $\exp(-c/\gamma^2)$, and thus leaves us large room for balancing with the norm of the approximator.
3. The proof here only works for Gaussian kernel. A similar conclusion may hold for General RBF kernels using the fact that any RBF kernel can be expressed as an average of Gaussian kernel over different values of γ. A relevant reference is Scovel et al. [2010].

The last component for the proof of Theorem 2 is the sub-exponential decay rate of the spectrum of Σ determined by the Gaussian kernel. The distribution of the spectrum of the convolution operator with respect to a distribution density function p has been studied by Widom [1963]. It shows that the number of eigenvalues of Σ greater than μ is asymptotic to $(2\pi)^{-d}$ times the volume of

$$\left\{(x, \xi) : p(x)\hat{k}(\xi) > \mu\right\} \, ,$$

where $\hat{k}$ is the Fourier transform of the kernel function k. By applying Widom [1963]'s work in our case, we have the following lemma. It is essentially Corollary 27 in Eric et al. [2008], but our version explicitly shows the dependence on the band width β.

Lemma 8. *Assume $\hat{k}(\xi) \leq \alpha \exp(-\beta\|\xi\|^2)$. If the density function $p(x)$ of probability distribution $\mathbb{P}_{\mathcal{X}}$ is bounded by B and $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$ with radius ρ, then*

$$\lambda_i(\Sigma) \leq C\alpha B \exp\left(-\beta\left(\frac{4\Gamma^{4/d}(d/2+1)}{\pi^{4/d}\rho^2}\right) i^{2/d}\right) \, ,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots$ are eigenvalues of Σ in descending order.

Proof. Denote by E_t the set

$$\left\{(x,\xi) : \hat{k}(\xi)p(x) > t\right\}.$$

The volume, that is, the Lebesgue measure of E_t is denoted by $\mathrm{Vol}(E_t)$. By Theorem II of Widom [1963], the non-increasing function $\phi(\alpha)$ defined on $\mathbb{R}^+$ which is equi-measurable with $p(x)\hat{k}(\xi)$ describes the behaviour of λ_is. Indeed, $\lambda_i \leq C\phi((2\pi)^d i)$. By the volume formula of $2d$-dimensional ball we have the following estimate,

$$\begin{aligned}\sup\{s \in \mathbb{R}^+ : \phi(s) > t\} &= \mathrm{Vol}(E_t) \\ &\leq C_{d,\rho}\left(\frac{\ln(\alpha B/t)}{\beta}\right)^{d/2},\end{aligned}$$

where

$$C_{d,\rho} = \frac{\rho^d \pi^{d+2}}{\Gamma^2(d/2+1)}. \tag{52}$$

Solving for t, we know that

$$\phi(s) \leq \alpha B \exp\left(-\beta\left(\frac{s}{A}\right)^{2/d}\right).$$

Therefore, we have

$$\lambda_i(\Sigma) \leq C\alpha B \exp\left(-\beta\left(\frac{(2\pi)^d i}{A}\right)^{2/d}\right) \tag{53}$$

$$= C\alpha B \exp\left(-\beta\left(\frac{4\Gamma^{4/d}(d/2+1)}{\pi^{4/d}\rho^2}\right) i^{2/d}\right). \tag{54}$$

□

Now we can prove Theorem 2.

Proof. Note that, by Lemma 7, we can construct $g \in \mathcal{F}$ such that $R^h_{\mathbb{P},\lambda} - R^*$ is controlled. And by Theorem 4, we can find an $f_0 \in \mathcal{F}_N$ with similar risk to g. And this will be our f_0 as required by Theorem 3. So we have

$$R^h_{\mathbb{P},\lambda}(f_0) - R^* \leq \frac{2(\sqrt{\pi/2}\rho^2)^d}{\Gamma^2(d/2+1)}\frac{\lambda}{\gamma^d} + \frac{2(\sqrt{\pi/2}\rho^2)^{d/2}}{\Gamma(d/2+1)}\sqrt{\mu} \tag{55}$$

$$+ \frac{4\tau^{d-2}}{\Gamma(d/2)}\exp\left(\frac{\tau^2}{\gamma^2}\right)\gamma^{d-2}, \tag{56}$$

and $\|f_0\|_{\mathcal{F}_N} \leq 2$, with probability $1-\delta$, when $N = 5d(\mu)\ln(16d(\mu)/\delta)$. We choose $\gamma = \tau/\sqrt{\ln m}$ and $\lambda = 1/m$. Under the boundedness assumption on the density function and the property of Gaussian kernel, we know that by Lemma 8,

$$\lambda_i(\Sigma) \leq C\gamma B \exp\left(-\gamma^2\frac{4\Gamma^{4/d}(d/2+1)}{\pi^{4/d}\rho^2} i^{2/d}\right). \tag{57}$$

And similar to the second part of Theorem 1, by identifying

$$c_3 = C\gamma B = CB\tau/\sqrt{\ln m} \quad c_4 = \frac{4\tau^2\Gamma^{4/d}(d/2+1)}{\pi^{4/d}\rho^2 \ln m} := \frac{A}{\ln m}, \tag{58}$$

and choosing $\mu = c_3/(m^{2d^2} \vee \exp(\frac{d^2}{c_4} \vee c_4 d^2))$, we have

$$d(\mu) \leq 5d^{2d}(c_4^{-2d} \vee 1 \vee c_4^{-d}2^d \ln^d m). \tag{59}$$

Then when $m \geq \exp(A)$, we have $d(\mu) \leq 5(A^2 \wedge A/2)^{-d}\ln^{2d} m$, and

$$N = 5d(\mu)(\ln(16d(\mu)) + \ln(1/\delta)) \tag{60}$$

$$\leq 25(A^2 \wedge A/2)^{-d}\ln^{2d} m(\ln(80(A^2 \wedge A/2)^{-d}) + 2d\ln\ln m + \ln(1/\delta)). \tag{61}$$

Plug N and λ into Equation 8.

$$3r = 75C_{1,2}c_{d,\tau,\rho}(\ln(80c_{d,\tau,\rho}) + 2d\ln\ln m + \ln(1/\delta))(V+1)\frac{\ln^{2d+1} m}{m} \tag{62}$$

$$+ \frac{15+216V}{m}\ln(1/\delta) + 3r^*\,, \tag{63}$$

where

$$C_{1,2} = 900C_1^2 + 120C_2, \quad c_d = (A^2 \wedge A/2)^{-d}\,. \tag{64}$$

We can bound r^* by $R^h_{\mathbb{P},\lambda}(f_0) - R^*$. Therefore, the overall upper bound on the excess error is

$$R^1_{\mathbb{P},\lambda}(f_{m,N,\lambda}) - R^* \le \frac{18(\sqrt{\pi/2}\rho^2)^d}{\Gamma^2(d/2+1)}\frac{\ln^{d/2} m}{\tau m} + \frac{18(\sqrt{\pi/2}\rho^2)^{d/2}}{\Gamma(d/2+1)}\frac{\sqrt{CB\tau}\ln^{1/4} m}{m^{d^2}} \tag{65}$$

$$+ \frac{36\tau^{d-2}}{\Gamma(d/2)}\frac{\tau^{d-2}}{m\ln^{d/2-1} m} \tag{66}$$

$$+ 75C_{1,2}c_{d,\tau,\rho}(\ln(80c_{d,\tau,\rho}) + 2d\ln\ln m + \ln(1/\delta))(V+1)\frac{\ln^{2d+1} m}{m} \tag{67}$$

$$+ \frac{15+216V}{m}\ln(1/\delta)\,. \tag{68}$$

□

E Learning Rate without Optimized Feature Maps

In this section, we discuss the learning rate of RFSVM without an optimized feature map. As shown by Rudi and Rosasco [2017], RFKRR can achieve excess risk of $O(1/\sqrt{m})$ using $O(\sqrt{m}\log(m))$ features. However, it is inappropriate to directly compare this result with the learning rate in classification scenario. Because as surrogate loss functions, least square loss has a different calibration function with for example hinge loss. Basically, $O(\epsilon)$ risk under square loss only implies $O(\sqrt{\epsilon})$ risk under $0-1$ loss, while $O(\epsilon)$ risk under hinge loss implies $O(\epsilon)$ risk under $0-1$ loss. Therefore, Rudi and Rosasco [2017]'s analysis only implies an excess risk of $O(m^{-1/4})$ in classification problems with $\tilde{O}(\sqrt{m})$ features.

For RFSVM, we expect a similar result. Without assuming an optimized feature map, the leverage score can only be upper bounded by κ^2/μ, where κ is the upper bound on the function $\phi(\omega; x)$ for all ω, x. Substituting κ^2/μ for $d(\mu)$ in the proofs of learning rates, we need to balance $\sqrt{\mu}$ with $1/(\mu m)$ to achieve the optimal rate. This balance is not affected by the spectrum of Σ or whether $f^*_{\mathbb{P}}$ belongs to $\mathcal{F}$. Obviously, setting $\mu = m^{-2/3}$, we get a learning rate of $m^{-1/3}$, with $\tilde{O}(m^{2/3})$ random features. Even though this result is also new for RFSVM in regularized formulation, the gap to previous analysis like Rahimi and Recht [2008] is too large. Considering that the random features used in practice that are not optimized also have quite good performance, we need further analysis on RFSVM without optimized feature map.

F Supplementary Figures

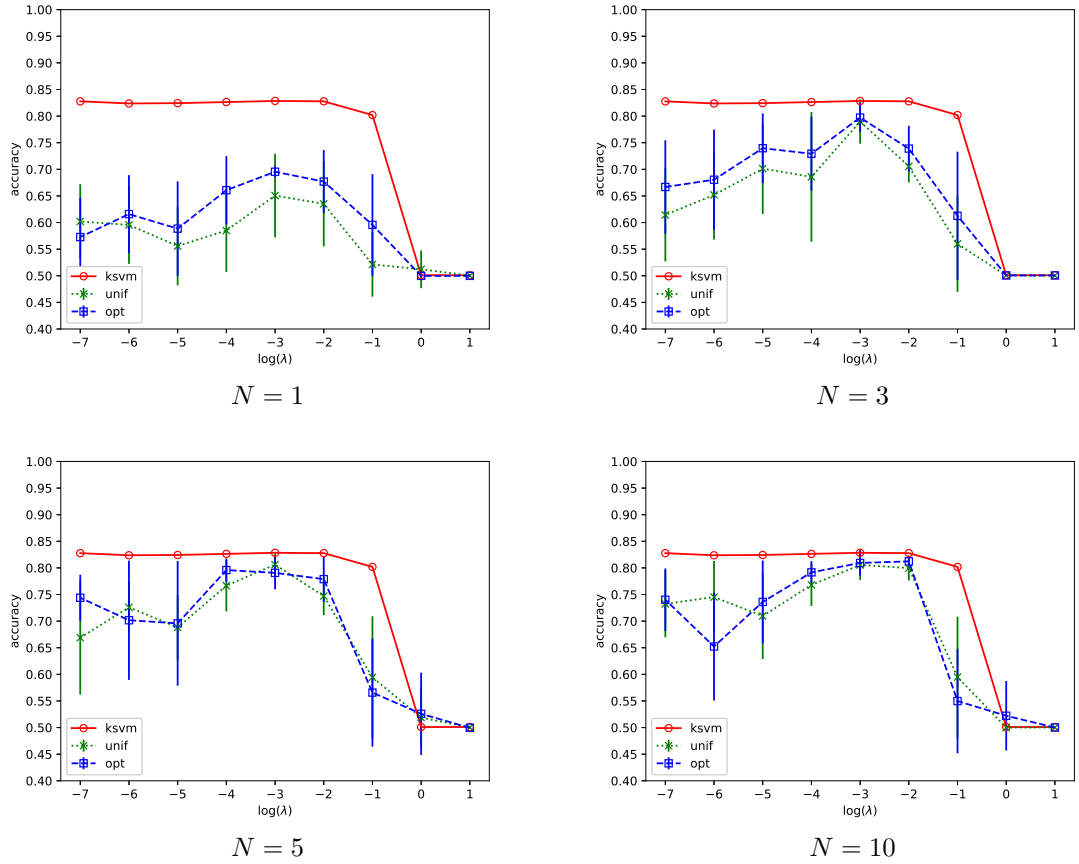

Figure 5: Comparison between RFSVMs with KSVM Using Gaussian Kernel.

"ksvm" is for KSVM with Gaussian kernel, "unif" is for RFSVM with direct feature sampling, and "opt" is for RFSVM with reweighted feature sampling. Error bars represent standard deviation over 10 runs. Each sub-figure shows the performance of RFSVM with different number of features N.

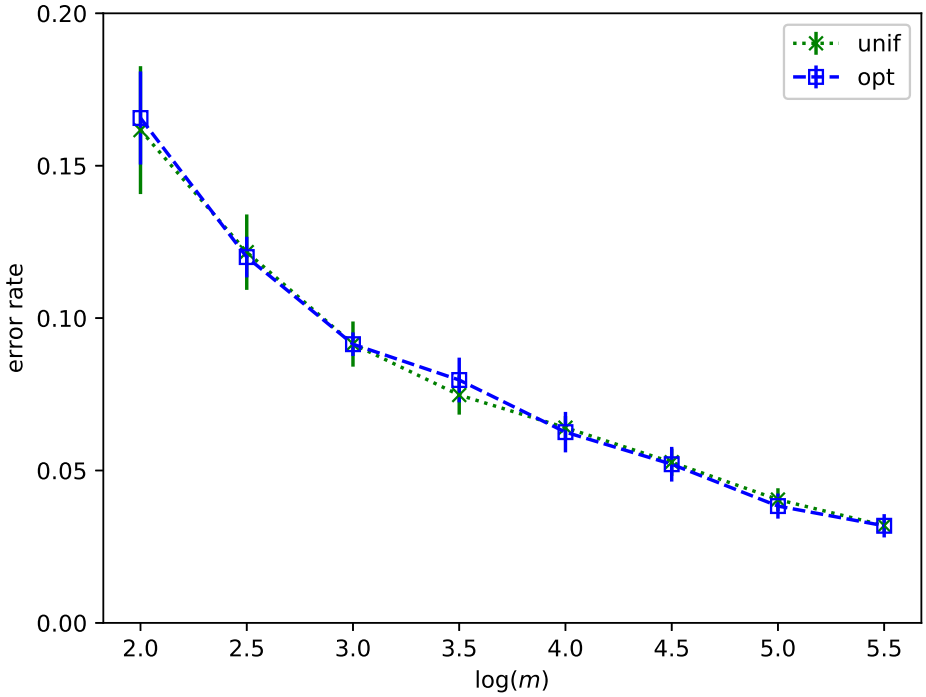

Figure 6: The excess risks of RFSVMs with the simple random feature selection ("unif") and the reweighted feature selection ("opt") are shown for different sample sizes in the binary classification task over 10 dimensional data. The data with probability 0.9 to be -1 are within the 10 dimensional ball centered at the origin and radius 0.9, and the data with probability 0.9 to be 1 are within the shell of radius 1.1 to 2. The error rate is the excess risk. The error bars represent the standard deviation over 10 runs.

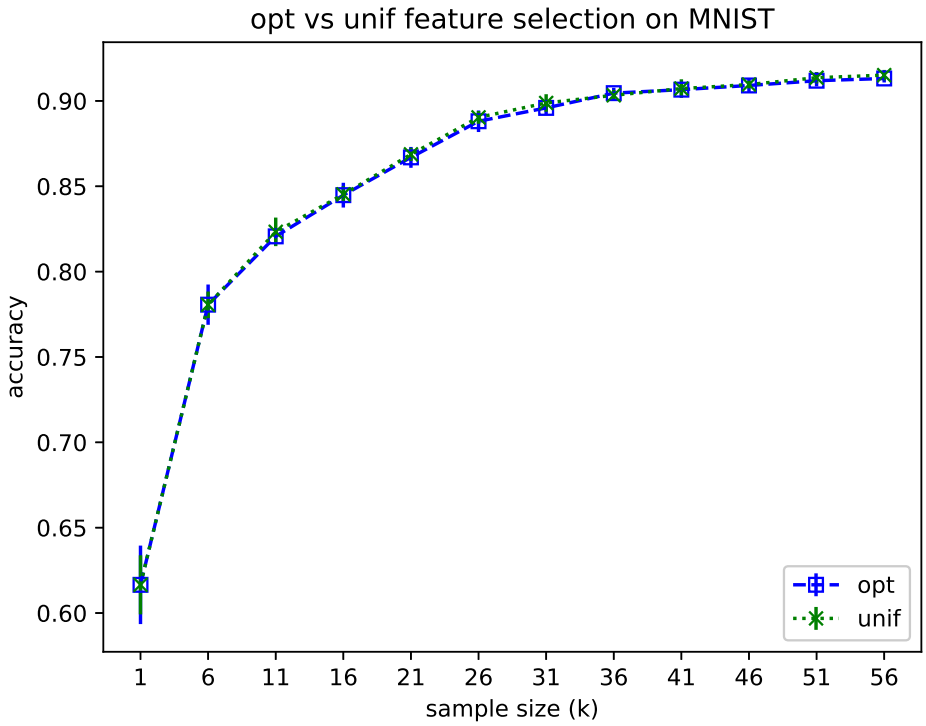

Figure 7: The classification accuracy of RFSVM with the simple random feature selection ("unif") and the reweighted feature selection ("opt") are shown for different sample sizes in the hand-written digit recognition (MNIST)