Master's Thesis

Generalization Error Analysis of Stochastic Gradient Descent on Classification Problems under Low Noise Condition (低ノイズ条件下の識別問題における確率的勾配降下法の汎化誤差解析)

48-186219 Shingo Yashima （八嶋 晋吾）

Supervisor Associate Professor Taiji Suzuki （鈴木 大慈 准教授）

January, 2020

Department of Mathematical Informatics
Graduate School of Information Science and Technology
The University of Tokyo

Abstract

The trade-off between statistical and computational performances is key for modern machine learning algorithms. On the one hand, the ultimate goal is to achieve the best possible prediction error. On the other hand, time and memory requirements are unavoidable constraints in large scale problems. In this view, sketching and stochastic gradient methods are among the most commonly used techniques to derive efficient large-scale learning algorithms.

In this thesis, we consider learning binary classification problems in reproducing kernel Hilbert space with the use of such techniques. Recent studies [43, 42] have shown that the expected classification error converges exponentially fast with stochastic gradient descent under the specific condition on label distribution, called strong low noise condition. Based on these analyses, we give the following two contributions in this thesis.

Firstly, we extend these analyses to the general low noise condition and show that the convergence of the expected classification error faster than the optimal rate of the expected risk is achievable with SGD under such a condition.

Secondly, we consider solving the problem with the combination of sketching and stochastic gradient descent, which yields much better computational efficiency. Analyzing the error induced by the approximation of random features, which is the most popular kernel sketching method, we show the exponential convergence of the expected classification error under the strong low noise condition is achieved even if random features approximation is applied. Additionally, we demonstrate that the convergence rate does not depend on the number of features and there is a significant computational benefit in using random features.

These results suggest the theoretical validity of these commonly used approximation methods in classification problems.

Keywords Binary classification, Stochastic gradient descent, Kernel method, Random features, Low noise condition

Contents

Chapter 1

Introduction

1.1 Motivation and Purpose

Classification is one of the main tasks in machine learning and has been frequently appeared in practical problems such as image recognition. We consider an ordinal supervised setting, where a set of input data and corresponding discrete labels is given as training data and we aim to learn a classifier that can produce correct labels from unseen data. The performance of a classifier is usually measured by the *expected classification error* (also known as accuracy), which indicates the probability of misclassification by a learned classifier. However, directly minimizing the expected classification error is known to intractable due to its non-convexity [4]. Thus, convex surrogate losses (e.g., logistic loss or hinge loss), which upper bounds classification errors, are usually used as an objective function and we try to minimize the expected classification error through minimizing the *expected risk*, which is the expectation of such a tractable convex loss. This convex approximation is theoretically justified by consistency property of loss functions [10, 66] and these minimization problems are solved through general criteria of statistical learning theory, such as empirical risk minimization (ERM).

On the other hand, the size of training datasets is becoming larger and larger in modern machine learning and it is crucial to investigate computationally efficient methods to optimize learning models with such large datasets. Stochastic gradient descent (SGD) [49], that can learn from a single or a few passes over the data, is a workhorse of such a large scale problems and is widely used in optimizations of practical models, such as deep neural networks (DNN) and kernel methods, owing to its scalability, wide applicability, simplicity of implementation, and superior performance. In the optimization literature, a great deal of works has proposed more sophisticated variants of SGD using, for example, variance reduction [30, 20, 41] or acceleration technique [2, 37].

From theoretical perspectives, the convergence rates of objective functions (risk) are extensively studied. For example, optimal convergence rates is known as $O(1/\sqrt{n})$ and $O(1/n)$, where n denotes the sample size, for convex and strongly convex loss functions, respectively [39, 1] and these rates are actually attained [38].

Going back to the origins of supervised classification problems, we are interested in not the convergence rates of risk, but the one of classification error. It is known that the excess classification error (equal to the expected classification error minus the minimal expected classification error over all measurable functions) is upper bounded by a function of the excess risk [10], but always such an upper bound results in a loose bound (e.g., no difference or taking square roots from the excess risk). As a result, the convergence rates of the expected classification errors are always equal or slower than $O(1/\sqrt{n})$ or $O(1/n)$. However, there are a variety of extra assumptions that could allow for faster rates.

In particular, such general relationships between excess loss and excess error has been

refined by the use of the *low noise condition* [63, 10], which indicates how hard the prediction problems are. Intuitively, if the label probability of an input point is equal over all labels, it is difficult to decrease the classification error on that point and such cases are considered as the problems with noisy label distribution. In contrast to this, if labels are deterministic (i.e., labels have no noise), those are the easiest problems where we only have to fit the model to the label distribution. The low noise condition quantifies the mass of input points that have noisy label distribution and this condition can accelerate the convergence rate of the excess classification errors faster than $1/n$ and even exponentially fast, under the specific setting with empirical risk minimization (ERM) [6].

In this thesis, we investigate the effectiveness of SGD in binary classification problems, where the label set is $\{-1, 1\}$, in terms of the convergence of the classification error. Recent studies [43, 42] have shown that the exponential convergence property of SGD in reproducing kernel Hilbert space (RKHS) under the *strong low noise condition*, which is the strongest version of the low noise condition. Inspired by these analyses, we demonstrate the following two contributions in this thesis.

First, we show that even under the general (not strong) low noise condition, the faster convergence rate than $O(1/n)$ can be achievable with SGD in Chapter 3. Since the strong low noise condition is somewhat restrictive when considering real-world datasets, our results give a better explanation of the practical success of SGD in classification problems.

After that, in Chapter 4 we consider applying the approximation method of RKHS called random features [47], which allow reducing data-dimensionality, hence memory requirements, by random projections. Combining this with SGD, it yields much better computational efficiency and this combination is widely used in practice. We analyze the required approximated dimension to achieve the exponential convergence of the expected classification errors under the strong low noise condition and show that there is the theoretically computational benefit to apply these approximation methods.

1.2 Organization of the Thesis

In Chapter 2, we introduce a problem setting of binary classification problems and details of algorithm which is treated in this thesis and show some related known results on these topics. Then we present the results on the convergence analysis of SGD under the general low noise condition in Chapter 3. Next, we demonstrate the result about the validity of random features under the strong low noise condition in Chapter 4. Finally, we conclude the thesis and discuss future directions in Chapter 5. Missing proofs in Section 2-4 are found in Appendix.

1.3 Notation

In general, for a probability measure μ on a topological space $\mathcal{X}$, $L^2(d\mu)$ denotes a space of square-integrable functions with respect to μ, that is,

$$L^2(d\mu) \stackrel{\text{def}}{=} \left\{ f : \mathcal{X} \to \mathbb{R} \;\middle|\; \|f\|_{L^2(d\mu)} < \infty \right\} / \mathcal{N},$$
$$\|f\|_{L^2(d\mu)} \stackrel{\text{def}}{=} \left(\int_{\mathcal{X}} |f(x)|^2 d\mu(x) \right)^{1/2},$$

where $\mathcal{N}$ denotes the kernel space of the norm:

$$\mathcal{N} \stackrel{\text{def}}{=} \{ f : \mathcal{X} \to \mathbb{R} \mid f(x) = 0 \text{ for } \mu\text{-almost sure } x \in \mathcal{X} \}.$$

We denote one with respect to the Lebesgue measure by $L^2(\mathcal{X})$. In addition, $L^\infty(d\mu)$ denotes a space of functions for which the essential supremum with respect to μ is bounded:

$$L^\infty(d\mu) \stackrel{\text{def}}{=} \left\{ f : \mathcal{X} \to \mathbb{R} \;\middle|\; \|f\|_{L^\infty(d\mu)} < \infty \right\} / \mathcal{N},$$
$$\|f\|_{L^\infty(d\mu)} \stackrel{\text{def}}{=} \inf\{C \geq 0 \mid |f(x)| \leq C \text{ for } \mu\text{-almost sure } x \in \mathcal{X}\}.$$

Similarly, we denote one with respect to the Lebesgue measure by $L^\infty(\mathcal{X})$.
Let V be a Hilbert space. We denote its inner product and the induced norm by $\langle \cdot, \cdot \rangle_V$ and $\| \cdot \|_V$, respectively. For $A : V \to V$, we denote an operator norm of A by $\|A\|_{\text{op}}$, that is,

$$\|A\|_{\text{op}} = \sup_{v \in V} \frac{\|Av\|_V}{\|v\|_V}.$$

For $a, b \in V$, we define an outer product $a \otimes_V b : V \to V$ as follows:

$$(a \otimes_V b)v = \langle b, v \rangle_V a, \quad \forall v \in V.$$

Let W be a closed subspace of V, then a projection onto W is well-defined and we denote its operator by $\mathcal{P}_W$. Then we have

$$v = \mathcal{P}_W v + \mathcal{P}_{W^\perp} v, \quad \forall v \in V.$$

Furthermore, we define a partial order $\preceq$ between linear, positive semi-definite and self-adjoint operators $A, B : V \to V$ as follows:

$$A \preceq B \quad \stackrel{\text{def}}{\iff} \quad \langle Av, v \rangle_V \leq \langle Bv, v \rangle_V, \ \forall v \in V.$$

Chapter 2

Preliminary

2.1 Binary Classification Problem

2.1.1 Classification error and surrogate loss

In this section we introduce settings and basic properties of binary classification problem. Let $\mathcal{X}$ and $\mathcal{Y} = \{-1, 1\}$ be a feature space and the set of binary labels, respectively; ρ denotes a probability measure on $\mathcal{X} \times \mathcal{Y}$, by $\rho_{\mathcal{X}}$ the marginal distribution on X, and by $\rho(\cdot|X)$ the conditional distribution on Y, where $(X, Y) \sim \rho$. In the classification problem, our final objective is to choose a discriminant function $g : \mathcal{X} \to \mathbb{R}$ such that the sign of $g(X)$ is an accurate prediction of Y. Therefore, we intend to minimize the *expected classification error* $\mathcal{R}$ defined below:

$$\mathcal{R}(g) = \mathbb{E}_{(X,Y)\sim\rho}\left[I(\mathrm{sgn}(g(X)), Y)\right], \tag{2.1}$$

where $\mathrm{sgn}(x) = 1$ if $x > 0$ and -1 otherwise, and I represents 0-1 loss:

$$I(y, y') = \begin{cases} 1 & (y \neq y') \\ 0 & (y = y'). \end{cases}$$

Infimum value of $\mathcal{R}$ amongst all measurable functions is denoted by $\mathcal{R}_*$ and it is called the *Bayes classification error*. By definition, any function g satisfying $\mathrm{sgn}(g(X)) =$

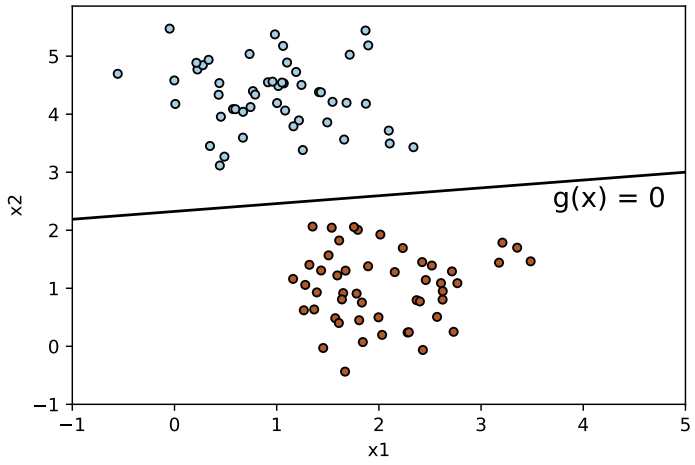


Fig. 2.1: Example of binary classification problems.

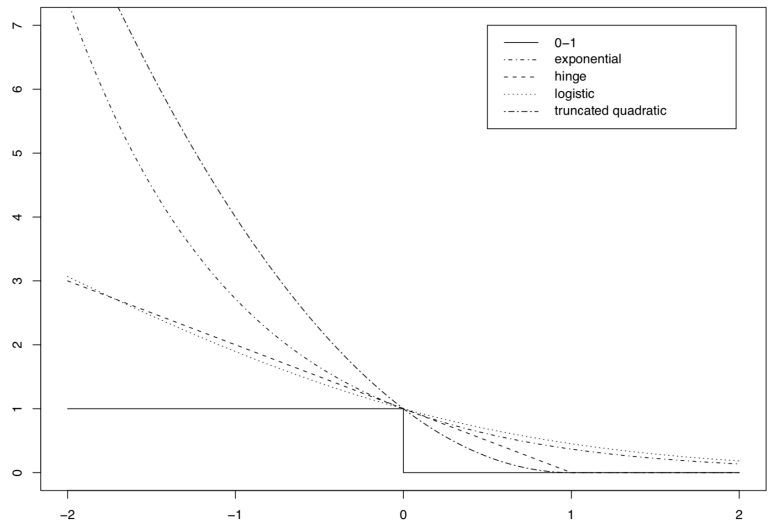


Fig. 2.2: 0-1 loss and surrogate loss functions [10].

$\mathrm{sgn}(2\rho(1|X) - 1)$ almost surely on the set $\{X \mid \rho(1|X) \neq 1/2\}$ has $\mathcal{R}(g) = \mathcal{R}_*$, and such a function is called the *Bayes classifier*. However, directly minimizing (2.1) to obtain the Bayes classifier is generally intractable because of its non-convexity [4]. Thus, we generally use the convex surrogate loss $l(\zeta, y)$, which is an upper bound of 0-1 loss, and minimize the *expected risk* $\mathcal{L}$, which is the expectation of the loss function value with respect to ρ:

$$\mathcal{L}(g) = \mathbb{E}_{(X,Y)\sim\rho}\left[l(g(X), Y)\right]. \tag{2.2}$$

Its infimum value amongst all measurable functions is denoted by $\mathcal{L}_*$ and it is called the *Bayes risk*. In general, the loss function l has a form $l(\zeta, y) = \phi(\zeta y)$ where $\phi : \mathbb{R} \to \mathbb{R}$ is a non-negative convex function. The typical examples are logistic loss $\phi(v) = \log(1 + \exp(-v))$ for logistic regression [25], hinge loss $\phi(v) = \max\{0, 1 - v\}$ for support vector machine [18], and exponential loss $\phi(v) = \exp(-v)$ for AdaBoost [24]. These functions are illustrated in Figure 2.2.

Next, we show the relationship between the expected classification error $\mathcal{R}$ and the expected risk $\mathcal{L}$. Firstly, the conditional expected risk on data point x is given as follows:

$$\mathbb{E}[\phi(Yg(X))|X = x] = \rho(1|x)\phi(g(x)) + (1 - \rho(1|x))\phi(-g(x)).$$

It is useful to think of this in terms of a generic probability $\mu \in [0, 1]$ and a generic classifier value $\alpha \in \mathbb{R}$, and we denote it by $C_\mu(\alpha)$:

$$C_\mu(\alpha) \stackrel{\text{def}}{=} \mu\phi(\alpha) + (1 - \mu)\phi(-\alpha).$$

For $\mu \in [0, 1]$, define its optimal value with respect to α as follows:

$$l_*(\mu) \stackrel{\text{def}}{=} \inf_{\alpha\in\mathbb{R}} C_\mu(\alpha) = \inf_{\alpha\in\mathbb{R}} \left\{\mu\phi(\alpha) + (1 - \mu)\phi(-\alpha)\right\}.$$

If its infimum is uniquely attained for some α, we can define the *link function* $h_* : [0, 1] \to \mathbb{R}$, which connects the hypothesis space and the probability measure:

$$h_*(\mu) \stackrel{\text{def}}{=} \arg\min_{\alpha\in\mathbb{R}} C_\mu(\alpha) = \arg\min_{\alpha\in\mathbb{R}} \left\{\mu\phi(\alpha) + (1 - \mu)\phi(-\alpha)\right\}.$$

Intuitively, for given data point x, $h_*(\rho(1|x))$ is the optimal prediction value which minimizes the expected risk on x. It can be checked that the following equality holds:

$$\mathcal{L}_* = \inf_{g:\mathcal{X}\to\mathbb{R}} \mathcal{L}(g) = \mathbb{E}\left[l_*(\rho(1|X))\right] = \mathbb{E}[l(h_*(\rho(1|X)), Y)].$$

In addition, define the optimal value of $C_\mu(\alpha)$ under the constraint that the sign of α disagrees with that of $2\mu - 1$:

$$l_*^-(\mu) \stackrel{\text{def}}{=} \inf_{\alpha:\alpha(2\mu-1)<0} C_\mu(\alpha) = \inf_{\alpha:\alpha(2\mu-1)<0} \left\{\mu\phi(\alpha) + (1-\mu)\phi(-\alpha)\right\}.$$

Then we define transform function $\psi : [0,1] \to [0,\infty)$ by $\psi \stackrel{\text{def}}{=} \tilde{\psi}^{**}$, where

$$\tilde{\psi}(\theta) = l_*^-\left(\frac{1+\theta}{2}\right) - l_*\left(\frac{1+\theta}{2}\right).$$

and $g^{**} : [0,1] \to \mathbb{R}$ is the Fenchel-Legendre biconjugate[*1] of $g : [0,1] \to \mathbb{R}$, which satisfies[*2]

$$\text{epi } g^{**} = \overline{\text{co}}\ \text{epi } g.$$

Finally, we can relate the *excess classification error* (equivalent to the expected classification error minus the Bayes classification error) to the *excess risk* (equivalent to the expected risk minus the Bayes risk) through transform function ψ [10, Theorem 3]:

$$\psi(\mathcal{R}(g) - \mathcal{R}_*) \le \mathcal{L}(g) - \mathcal{L}_*. \tag{2.3}$$

Since our final objective is minimizing the expected classification error, it is natural to expect that minimizing (2.2) ensures minimizing (2.1) (the Bayes risk consistency). In other words, a function that attains $\mathcal{L}_*$ should also attain $\mathcal{R}_*$. The following proposition shows conditions on ϕ to satisfy this consistency.

Proposition 2.1 (Theorem 3 in [10])**.** *The following conditions are equivalent:*

- *For any sequence of measurable functions $g_i : \mathcal{X} \to \mathbb{R}$ and any distribution ρ on $\mathcal{X} \times \mathcal{Y}$,*
$$\lim_{i\to\infty} \mathcal{L}(g_i) = \mathcal{L}_* \text{ implies } \lim_{i\to\infty} \mathcal{R}(g_i) = \mathcal{R}_*.$$
- *For any sequence θ_i in $[0,1]$,*
$$\lim_{i\to\infty} \psi(\theta_i) = 0 \text{ if and only if } \lim_{i\to\infty} \theta_i = 0.$$

When these conditions hold, such ϕ is called *classification-calibrated*.

Example (Logistic loss)**.** Here $\phi(\alpha) = \log(1 + \exp(-\alpha))$. If $\mu = 0$, then $C_\mu(\alpha) \to 0$ as $\alpha \to -\infty$; if $\mu = 1$, then $C_\mu(\alpha) \to 0$ as $\alpha \to \infty$. Thus we have $l_*(0) = l_*(1) = 0$ and h_* cannot be defined for $\mu = 0, 1$. For $\mu \in (0,1)$, we have

$$\begin{aligned} h_*(\mu) &= \log\frac{\mu}{1-\mu}, \\ l_*(\mu) &= -\mu\log\mu - (1-\mu)\log(1-\mu), \\ l_*^-(\mu) &= C_\mu(0) = \log 2, \end{aligned}$$

[*1] For $g : [0,1] \to \mathbb{R}$, its Fenchel-Legendre conjugate $g^* : [0,1] \to \mathbb{R}$ is given by $g^*(y) = \sup_{x\in[0,1]}\{\langle x, y\rangle - g(x)\}$, and its biconjugate g^{**} is further conjugate of g^*.

[*2] Here $\overline{\text{co}}\ S$ is the closure of the convex hull of the set S, and epi g is the epigraph of the function g, that is, the set $\{(x,t) : x \in [0,1], g(x) \le t\}$.

and so

$$\psi(\theta) = \tilde{\psi}(\theta) = \frac{1+\theta}{2}\log(1+\theta) + \frac{1-\theta}{2}\log(1-\theta),$$

where the first equality follows from that $\tilde{\psi}$ is convex. Since $\psi(0) = 0$ and ψ is strictly increasing and continuous, it satisfies the second condition in Proposition 2.1 and is thus classification-calibrated. In addition, we have $\log(1+\theta) \approx \theta$ around $\theta = 0$ and thus $\psi(\theta) \approx \theta^2$, so we can see from (2.3) that the convergence rate of the excess classification error is bounded by square root of that of the excess risk in the case of logistic loss.

For other loss functions, similar calculation yields that $\psi(\theta) = 1 - \sqrt{1-\theta^2}$ for exponential loss and $\psi(\theta) = \theta$ for hinge loss. So they also satisfy the second condition of the proposition and are thus classification-calibrated.

Remark. For convex ϕ, a simpler condition which is equivalent to being classification-calibrated is known (Theorem 6 in [10]); ϕ is classification-calibrated if and only if ϕ is differentiable at 0 and $\phi'(0) < 0$.

If there is a one-to-one correspondence between a probability and a classifier value thorough the link function h_*, one can relate the excess risk to the divergence between corresponding probability measures. Since l_* is a concave function (Lemma 2.1 in [66]), if additionally l_* is differentiable, we can define the Bregman divergence associates with l_* as follows:

$$d_{l_*}(\eta_1, \eta_2) \stackrel{\text{def}}{=} -l_*(\eta_2) + l_*(\eta_1) + l_*'(\eta_1)(\eta_2 - \eta_1).$$

We have the following proposition.

Proposition 2.2 (Theorem 2.2 in [66])**.** *Suppose ϕ is differentiable and h_* is well-defined, differentiable and invertible. Then l_* is also differentiable and the following equality holds:*

$$\mathcal{L}(g) - \mathcal{L}_* = \mathbb{E}\left[d_{l_*}\left(h_*^{-1}(g(X)), \rho(1|X)\right)\right].$$

If the divergence d_{l_*} is positive, that is, $d_{l_*}(\eta_1, \eta_2) = 0$ if and only if $\eta_1 = \eta_2$, we can see that the Bayes classifier g_*, which satisfies $\mathcal{L}(g_*) = \mathcal{L}_*$, is equal to $h_*(\rho(1|\cdot))$ almost surely.

Example (Logistic loss)**.** Suppose $\rho(1|X) \in (0,1)$ almost surely. Indeed, $\phi(\alpha) = \log(1 + \exp(-\alpha))$ is differentiable and $h_*(\mu) = \log(\mu/(1-\mu))$ is well-defined, differentiable and invertible on $(0,1)$. Furthermore, we have

$$d_{l_*}(\eta_1, \eta_2) = \eta_2 \log\left(\frac{\eta_1}{\eta_2}\right) + (1-\eta_2)\log\left(\frac{1-\eta_1}{1-\eta_2}\right).$$

This coincides with the Kullback-Leibler divergence between two Bernoulli distributions with parameter η_1 and η_2, and is thus positive.

2.1.2 Low noise condition

Although properties introduced above hold for any distribution ρ on samples, it is useful and reasonable to consider the case where there is small noise on labels. For example, in image datasets such as MNIST, a label is almost deterministically produced for almost all images and there are few ambiguous images that can be classified into multiple labels. From this perspective, it is reasonable to assume the conditional label probability $\rho(1|X)$ tends to be bounded away from 1/2. Here we formally introduce this low noise condition.

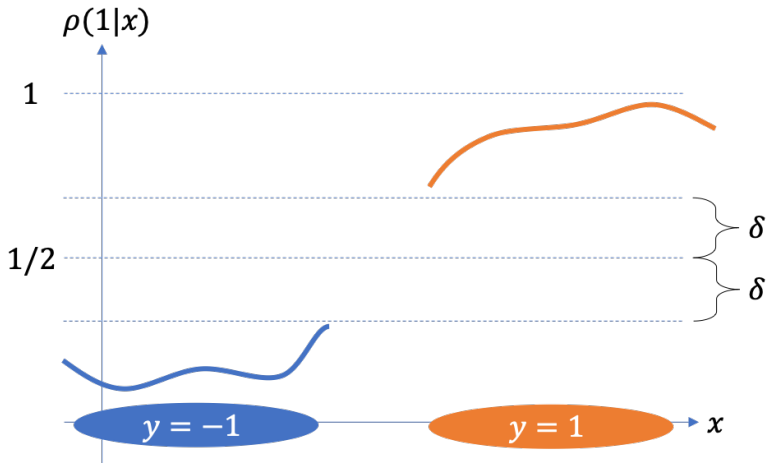


Fig. 2.3: The example of a conditonal distribution satisfying strong low noise condition.

Definition 2.1 (Low noise condition [63, 10, 6])**.** *A conditional label probability* $\rho(1|X)$ *satisfies low noise condition if there exists* $C, \alpha > 0$ *such that the following inequality holds for all* $\delta \in [0, 1/2]$*:*

$$\mathbb{P}\left[|\rho(Y = 1|X) - 1/2| \leq \delta\right] \leq C\delta^{\alpha}.$$

The strongest version of the low noise condition, which corresponds to the case of $\alpha = \infty$, is described as follows:

Definition 2.2 (Strong low noise condition [63, 6])**.** *A conditional label probability* $\rho(1|X)$ *satisfies the strong low noise condition if there exists* $\delta \in [0, 1/2]$ *such that the following inequality holds for* $\rho_{\mathcal{X}}$*-almost sure* $x \in \mathcal{X}$*:*

$$|\rho(Y = 1|x) - 1/2| > \delta.$$

In this situation, the conditional probability is uniformly bounded away from $1/2$ as shown in Figure 2.3. Under the low noise condition, there is a tighter version of bound (2.3) (Theorem 10 in [10]):

$$C\left(\mathcal{R}(g) - \mathcal{R}_*\right)^{1-\frac{1}{\alpha}} \psi\left(\frac{\left(\mathcal{R}(g) - \mathcal{R}_*\right)^{\frac{1}{\alpha}}}{2C}\right) \leq \mathcal{L}(g) - \mathcal{L}_*. \tag{2.4}$$

We note that since ψ is convex, it never gives a worse rate than (2.3). However, even if we set α to ∞, the resulting convergence of $\mathcal{R}$ cannot be faster than that of $\mathcal{L}$.

Remark (Relation to the complete separability)**.** In the theoretical analysis of classification problems, it is also common to consider labels to be produced deterministically and sometimes impose the complete separability, where the Bayes classifier can predict labels with 100% accuracy [56, 29]. This is much stronger condition than the low noise condition because it corresponds to the extreme case of the strong low noise condition, where δ is close to $1/2$.

In this section, we have introduced some basic properties of binary classification problems, mainly about the relationship between the classification error and the loss function. However, in order to minimize the expected risk $\mathcal{L}$ in practice, we have to specify a hypothesis space and an optimization method. Throughout this thesis, we consider a reproducing kernel Hilbert space (RKHS) as the former and stochastic gradient descent (SGD) as the latter. We introduce them in the following section.

2.2 Stochastic Gradient Descent on RKHS

2.2.1 Reproducing kernel Hilbert space

Firstly, we introduce a reproducing kernel Hilbert space (RKHS), which is widely adapted in a non-parametric statistical learning. Here we define a positive definite kernel.

Definition 2.3 (Positive definite kernel)**.** *$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive definite kernel if for any finite set $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, the $n \times n$ matrix whose (i, j) entry is $k(x_i, x_j)$ is positive semi-definite.*

Popular examples are Gaussian kernel with $k(x_1, x_2) = \exp(-\|x_1 - x_2\|_2^2/\sigma^2)$ for a band width parameter $\sigma > 0$, and polynomial kernel with $k(x_1, x_2) = (1 + x_1 \cdot x_2)^j$ for a degree parameter j where $\mathcal{X} = \mathbb{R}^d$. For every positive definite kernel k, a Hilbert space spanned by $\{k(\cdot, x) \mid x \in \mathcal{X}\}$ is defined:

Definition 2.4 (Reproducing kernel Hilbert space (RKHS))**.** *A reproducing kernel Hilbert space $\mathcal{H}$ associates with a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ satisfying the following properties:*

1. *$k(\cdot, x) \in \mathcal{H}$ for any $x \in \mathcal{X}$.*
2. *$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$ for any $x \in \mathcal{X}$ and $f \in \mathcal{H}$.*

It is known that every positive definite kernel has a unique RKHS [3]. We note that if $\mathcal{X} = \mathbb{R}^d$ and $k(x_1, x_2) = x_1 \cdot x_2$, its RKHS $\mathcal{H}$ is equal to $\mathbb{R}^d$ so the above definition includes the ordinal finite-dimensional Euclid space. For a more detailed characterization of RKHS, see Appendix A. Although in general RKHS can be infinite dimensional and thus have plenty expressive power, the above property enables us to treat the optimization problem in same algebraic framework as parametric models, and several studies extend linear parametric models to non-parametric ones on RKHS, such as kernel SVM [12], kernel PCA [53] and kernel k-means [21].

In this thesis, we consider solving the minimization problem of the expected risk introduced in the last section with RKHS $\mathcal{H}$ for some kernel k as the hypothesis space:

$$\min_{g \in \mathcal{H}} \mathcal{L}(g).$$

In general, non-parametric models have strong expressive power, so it is common to add regularization term to the loss function in order to avoid over-fitting. Particularly, we consider the following Tikhonov regularization with regularization parameter λ:

$$\min_{g \in \mathcal{H}} \left\{ \mathcal{L}_\lambda(g) \stackrel{\text{def}}{=} \mathcal{L}(g) + \frac{\lambda}{2} \|g\|_{\mathcal{H}}^2 \right\}. \tag{2.5}$$

We also use the following notation:

$$g_\lambda \stackrel{\text{def}}{=} \arg\min_{g \in \mathcal{H}} \mathcal{L}_\lambda(g).$$

The uniqueness of the minimizer is ensured by the regularization term. We note that the purpose of the regularization in (2.5) is to accelerate and stabilize the optimization, rather than to avoid over-fitting since we directly optimize the expected (not empirical) loss $\mathcal{L}$. See also remark at the end of this section.

2.2.2 Stochastic gradient descent

Now we introduce stochastic gradient descent (SGD) [49] as an optimization method of (2.5). It is a workhorse for large-scale machine learning problems due to its computational efficiency. The main idea of SGD is to replace a gradient of the objective with an unbiased (sometimes biased) estimate of that which can be calculated from one or small batch of samples. First, recall the definition of a gradient of a function F on $\mathcal{H}$ at $g \in \mathcal{H}$; it is an element $\nabla F(g) \in \mathcal{H}$ satisfying the following equation:

$$F(g+h) = F(g) + \langle \nabla F(g), h \rangle_{\mathcal{H}} + o(\|h\|_{\mathcal{H}}).$$

For the expected risk $\mathcal{L}$, when $k(x,x)$ is uniformly bounded on $\mathcal{X}$, its gradient at g takes the form $\mathbb{E}[\partial_\zeta l(g(X), Y) k(\cdot, X)]$, where ∂_ζ is a partial derivative with respect to the first variable of l. To see this, we have

$$\begin{aligned}
\mathcal{L}(g+h) &= \mathbb{E}[l((g+h)(X), Y)] \\
&= \mathbb{E}[l(g(X), Y) + \partial_\zeta l(g(X), Y) h(X) + o(|h(X)|)] \\
&= \mathcal{L}(g) + \langle \mathbb{E}[\partial_\zeta l(g(X), Y) k(\cdot, X)], h \rangle_{\mathcal{H}} + o(\|h\|_{\mathcal{H}}),
\end{aligned}$$

where the last equality follows from the fact that

$$\begin{aligned}
h(X) &= \langle h, k(\cdot, X) \rangle_{\mathcal{H}}, \\
|h(X)| &\leq \|h\|_{\mathcal{H}} \sqrt{k(X,X)}.
\end{aligned}$$

Thus, the stochastic gradient of $\mathcal{L}_\lambda$ at $g \in \mathcal{H}$ for a sample $Z = (X, Y) \in \mathcal{X} \times \mathcal{Y}$ is given by

$$G_\lambda(g, Z) \stackrel{\text{def}}{=} \partial_\zeta l(g(X), Y) k(\cdot, X) + \lambda g.$$

The algorithm of stochastic gradient descent is described in Algorithm 2.1.

Algorithm 2.1 Stochastic Gradient Descent (SGD)

Input: regularization parameter λ, number of iterations T, learning rates $\{\eta_t\}_{t=1}^T$
Output: classifier g_{T+1}

```
Initialize g_1 ∈ H
for t = 1, ..., T do
    Randomly draw a sample z_t = (x_t, y_t) ~ ρ
    g_{t+1} ← g_t − η_t G_λ(g_t, z_t)
end for
return g_{T+1}
```

Although this simple scheme is widely used in practice, it is also common to average all intermediate iterates with specified weights in order to decrease the effect of variance of stochastic gradients and stabilize the optimization. From theoretical perspective, the averaging scheme is used to derive the optimal rate of the stochastic optimization [46, 51, 36], which is known to $O(1/\sqrt{T})$ for convex and $O(1/T)$ for strongly convex objectives [39, 1]. The algorithm of this averaged stochastic gradient descent is described in Algorithm 2.2. The popular choices of α_t are uniform weights (called Polyak-Ruppert averaging) [46, 51] and polynomial decaying weights [32, 42]. For strongly convex objectives, SGD

Algorithm 2.2 Averaged Stochastic Gradient Descent (ASGD)

Input: regularization parameter λ, number of iterations T, learning rates $\{\eta_t\}_{t=1}^{T}$, averaging weights $\{\alpha_t\}_{t=1}^{T+1}$.
Output: classifier $\overline{g}_{T+1}$
  Initialize $g_1 \in \mathcal{H}$
  for $t = 1, \ldots, T$ **do**
    Randomly draw a sample $z_t = (x_t, y_t) \sim \rho$
    $g_{t+1} \leftarrow g_t - \eta_t G_\lambda(g_t, z_t)$
  end for
  return $\overline{g}_{T+1} = \sum_{t=1}^{T+1} \alpha_t g_t$

with the polynomial decaying averaging achieves the optimal $O(1/T)$ convergence rate, whereas that with the uniform averaging only achieve a sub-optimal $O(\log T/T)$ rate [32]. From this reason, we deal with polynomial decaying averaging and a learning rate adopted in [42], which is described as

$$\alpha_t = \frac{2(\gamma + t - 1)}{(2\gamma + T)(T + 1)},$$
$$\eta_t = \frac{2}{\lambda(\gamma + t)},$$

where $\gamma > 0$ is an offset parameter for the time index. This learning rate is also used in [14]. We note that an averaged iterate $\overline{g}_t$ can be updated iteratively as follows:

$$\overline{g}_1 = g_1,$$
$$\overline{g}_{t+1} = (1 - \theta_t)\overline{g}_t + \theta_t g_{t+1}, \quad \theta_t = \frac{2(\gamma + t)}{(t + 1)(2\gamma + t)}.$$

Using this formula, we can compute the averaged output without storing all internal iterate $(g_t)_{t=1}^{T+1}$. Throughout this thesis, we consider Algorithm 2.2 as an optimization method and simply refer to it as SGD.

If we assume that a cost of kernel evaluation is $O(1)$, the running time complexity is $O(t)$ for iteration t, thus overall complexity is $O(T^2)$ after T steps. Since we have to store T basis $\{k(\cdot, x_1), \ldots, k(\cdot, x_T)\}$ and their coefficients, the space complexity is $O(T)$.

Remark (Regularization and SGD)**.** It should be noted that there are two different paradigms in analyses of stochastic optimization: finite sum and online setting. In the former setting, we have all samples before actual training starts and we draw a sample from them at each iteration. Thus, the goal of optimization is to minimize a (regularized) empirical risk of given finite samples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$:

$$\min_{g \in \mathcal{H}} \left\{ \hat{\mathcal{L}}_\lambda(g) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} l(g(x_i), y_i) + \frac{\lambda}{2} \|g\|_{\mathcal{H}}^2 \right\}.$$

In this setting, the generalization error of generated hypothesis $\overline{g}_{T+1}$ is decomposed as follows:

$$\mathcal{L}(\overline{g}_{T+1}) - \mathcal{L}_* = \underbrace{\mathcal{L}(\overline{g}_{T+1}) - \mathcal{L}(\hat{g}_\lambda)}_{\text{optimization error of SGD}} + \underbrace{\mathcal{L}(\hat{g}_\lambda) - \mathcal{L}_*}_{\text{generalization error of ERM}},$$

where $\hat{g}_\lambda$ is an empirical risk minimizer:

$$\hat{g}_\lambda \stackrel{\text{def}}{=} \arg\min_{g \in \mathcal{H}} \hat{\mathcal{L}}_\lambda(g).$$

Thus, the performance of SGD is measured only from an optimization perspective, that is, the number of iterations required to get close to the empirical risk minimizer. In addition, the regularization term is included in the objective since that is essential for infinite-dimensional models to learn with empirical risk minimization (ERM) in general [15, 35, 59]. The trade-offs of optimization and generalization error in this setting are discussed in [13]. Besides, variance reduction technique [30, 20] gives faster convergence rates in finite sum setting (e.g., linear convergence with strong convexity), which utilizes the fact that exact gradients can be computed with $O(n)$ time in this setting.

In contrast to this, in the online setting, we obtain a new sample in each iteration and thus directly optimize towards a population risk minimizer. In this sense, regularization terms are not essential in this setting if we set an appropriate step size to balance optimization and generalization error. Indeed, some recent studies have shown that the optimal rates of, in particular, non-parametric least squares regression can be achieved by SGD without regularization terms [22, 44, 34, 26].

Our analysis lies in the online setting rather than the finite sum setting. However, we add a regularization term in the objective (2.5) in order to apply stability arguments to derive a convergence of generalization error as explained in the following section. It is an important future work to investigate whether the generalization analysis of SGD without a regularization term for general convex loss function is achievable or not.

Remark. Since we consider the one-pass SGD where we obtain a new sample in each iteration, sample size n and iteration number T have the same meaning throughout this thesis.

2.3 Convergence Properties of SGD

In this section, we introduce some results about the convergence rate of SGD, which are from [42]. Since a returned hypothesis $\overline{g}_{T+1}$ is a random variable depending on samples $\{(x_1, y_1), \ldots, (x_T, y_T)\}$, we evaluate the convergence of $\overline{g}_{T+1}$ separately; the convergence of the expected point $\mathbb{E}[\overline{g}_{T+1}]$ to the population risk minimizer g_λ and the concentration of $\overline{g}_{T+1}$ to its expectation $\mathbb{E}[\overline{g}_{T+1}]$ (see Figure 2.4).

2.3.1 Basic matters in convex optimization

Firstly, we provide some definitions that are widely adapted in the literature of convex optimization [40]. Let V be a Hilbert space throughout this section. We note that for a differantiable function $f : V \to \mathbb{R}$, its gradient ∇f is a linear operator from V to V.

Definition 2.5. *We say that a function $f : V \to \mathbb{R}$ is G-Lipschitz continuous ($G > 0$) if f satisfies*

$$|f(v_1) - f(v_2)| \leq G\|v_1 - v_2\|_V$$

for any $v_1, v_2 \in V$.

Definition 2.6. *$f : V \to \mathbb{R}$ is called L-Lipschitz smooth ($L > 0$) if f is differentiable and satisfies*

$$\|\nabla f(v_1) - \nabla f(v_2)\|_V \leq L\|v_1 - v_2\|_V \tag{2.6}$$

for any $v_1, v_2 \in V$.

If f is convex, (2.6) is equivalent to the following (Theorem 2.1.5 in [40]):

$$\langle \nabla f(v_1) - \nabla f(v_2), v_1 - v_2 \rangle_V \leq L\|v_1 - v_2\|_V^2. \tag{2.7}$$

Definition 2.7. *A convex and differentiable function* $f : V \to \mathbb{R}$ *is called* μ*-strongly convex* ($\mu > 0$) *if it satisfies*

$$f(v_2) \geq f(v_1) + \langle \nabla f(v_1), v_2 - v_1 \rangle_V + \frac{\mu}{2}\|v_2 - v_1\|_V^2$$

for any $v_1, v_2 \in V$.

As f is convex, it can be easily checked that a function $f(\cdot) + \frac{\lambda}{2}\|\cdot\|_V^2$ is λ-strongly convex.

2.3.2 Convergence of the expected point

Next, we show the convergence rate of the expected point to the population minimizer. Before entering to the result, we have to impose some ordinal assumptions about the loss function and the kernel function.

Assumption 2.1. $l(\cdot, y) : \mathbb{R} \to \mathbb{R}$ *is convex, differentiable, G-Lipschitz continuous and L-Lipschitz smooth. That is, for any* $\zeta, \zeta' \in \mathbb{R}$ *and* $y \in \mathcal{Y}$,

$$|\partial_\zeta l(\zeta, y)| \leq G,$$
$$|\partial_\zeta l(\zeta, y) - \partial_\zeta l(\zeta', y)| \leq L|\zeta - \zeta'|.$$

To control the boundedness of functions in $\mathcal{H}$, we need the following:

Assumption 2.2. *There exists* $R > 0$ *such that* $k(x, x) \leq R^2$ *for any* $x \in \mathcal{X}$.

This assumption yields an important relationship between different norms of $f \in \mathcal{H}$:

$$\|f\|_{L^\infty(d\rho_\mathcal{X})} = \sup_{x \in \mathrm{supp}(\rho_\mathcal{X})} \langle f, k(\cdot, x) \rangle_\mathcal{H} \leq \sup_{x \in \mathrm{supp}(\rho_\mathcal{X})} \sqrt{k(x,x)}\|f\|_\mathcal{H} \leq R\|f\|_\mathcal{H}.$$

Let $l(g, z)$ denote $l(g(x), y)$ for $z = (x, y)$ and $\partial_g l(g, z)$ denote the gradient of $l(g, z)$ with respect to $g \in \mathcal{H}$. Combining Assumption 2.1 and 2.2 yields LR^2-smoothness in $\mathcal{H}$, since

$$\begin{aligned}\langle \partial_g l(g, z) - \partial_g l(g', z), g - g' \rangle_\mathcal{H} &= \langle (\partial_\zeta l(g(x), y) - \partial_\zeta l(g'(x), y))k(\cdot, x), g - g' \rangle_\mathcal{H} \\ &\leq L(g(x) - g'(x))^2 \\ &\leq LR^2\|g - g'\|_\mathcal{H}^2\end{aligned}$$

holds for any $z \in \mathcal{X} \times \mathcal{Y}$ and it is known as an equivalent condition of smoothness from (2.7). Under these assumptions, the convergence rate of the expected point in Algorithm 2.2 is derived.

Proposition 2.3 (Proposition C in [42])**.** *Suppose Assumption 2.1, 2.2 holds. Consider Algorithm 2.2 with* $\eta_t = \frac{2}{\lambda(\gamma+t)}$ *and* $\alpha_t = \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)}$ *and assume* $\|g_1\|_\mathcal{H} \leq (2\gamma_1 + 1/\lambda)GR$ *and* $\eta_1 \leq \min\{1/LR^2, 1/2\lambda\}$. *Then, it follows that*

$$\|\mathbb{E}[\overline{g}_{T+1}] - g_\lambda\|_\mathcal{H}^2 \leq \frac{2}{\lambda}\left(\frac{18G^2R^2}{\lambda(2\gamma + T)} + \frac{\lambda\gamma(\gamma - 1)}{2(2\gamma + T)(T + 1)}\|g_1 - g_\lambda\|_\mathcal{H}^2\right).$$

This result shows that the expected point $\mathbb{E}[\overline{g}_{T+1}]$ converges to the population risk minimizer g_λ at sub-linear rate. This is shown by standard arguments in analyses of SGD and the proof is found in Appendix C.

2.3.3 Concentration to the expected point

Next proposition shows that the random variable $\overline{g}_{T+1}$ concentrates to its mean with an exponentially decaying probability.

Proposition 2.4 (Proposition 2 and D in [42]). *Suppose Assumption 2.1, 2.2 holds. Consider Algorithm 2.2 with* $\eta_t = \frac{2}{\lambda(\gamma+t)}$ *and* $\alpha_t = \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)}$ *and assume* $\|g_1\|_{\mathcal{H}} \leq (2\gamma_1 + 1/\lambda)GR$ *and* $\eta_1 \leq \min\{1/LR^2, 1/2\lambda\}$. *Then, it follows that*

$$\mathbb{P}\left[\left\|\overline{g}_{T+1} - \mathbb{E}[\overline{g}_{T+1}]\right\|_{\mathcal{H}} \geq \epsilon\right] \leq 2\exp\left(-\frac{\lambda^2(2\gamma+T)}{2^6 \cdot 3^2 G^2 R^2}\epsilon^2\right).$$

Here we briefly show the idea of the proof. The proof relies on the stability of SGD and a concentration inequality for martingale differences. Let $Z_1, \ldots, Z_T$ be i.i.d. random variables following ρ and $\mathcal{F}_t$ be a σ-field generated by $Z_1, \ldots, Z_t$. It consists a filtration since $\mathcal{F}_{t-1} \subset \mathcal{F}_t$ holds. Let $\overline{g}_T$ be an output of Algorithm 2.2 trained on samples $Z_1, \ldots, Z_T$ and define

$$D_t \stackrel{\text{def}}{=} \mathbb{E}[\overline{g}_{T+1}|\mathcal{F}_t] - \mathbb{E}[\overline{g}_{T+1}|\mathcal{F}_{t-1}]$$

for $1 \leq t \leq T$. Then we have

$$\overline{g}_{T+1} - \mathbb{E}[\overline{g}_{T+1}] = \sum_{t=1}^{T} D_t \tag{2.8}$$

and $D_1, \ldots, D_T$ consists a martingale difference sequence, that is,

- D_t is $\mathcal{F}_t$-measurable,
- $\mathbb{E}[D_t|\mathcal{F}_{t-1}] = 0$.

We utilize the following concentration inequality to obtain a probabilistic bound of (2.8):

Lemma 2.1 (Theorem 3.4 in [45]). *Let* $D_1, \ldots, D_T$ *be a martingale difference sequence taking values in* $\mathcal{H}$. *Assume that there exists* $c_T > 0$ *such that* $\sum_{t=1}^{T} \|D_t\|_\infty^2 \leq c_T^2$, *where* $\|D_t\|_\infty$ *is a essential supremum of* $\|D_t\|_{\mathcal{H}}$. *Then for any* $\epsilon > 0$, *we have*

$$\mathbb{P}\left[\sup_{1\leq s\leq T}\left\|\sum_{t=1}^{s} D_t\right\|_{\mathcal{H}} \geq \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2}{2c_T^2}\right).$$

The bound on c_T is obtained in the following manner. Let Z_t' be a random variable following ρ which is independent from $Z_1, \ldots, Z_T$ and $\overline{g}_{T+1}^t$ be an output of Algorithm 2.2 depending on $(Z_1, \ldots, Z_{t-1}, Z_t', Z_{t+1}, \ldots, Z_T)$. Then we have

$$\mathbb{E}[\overline{g}_{T+1}^t|\mathcal{F}_t] = \mathbb{E}[\overline{g}_{T+1}|\mathcal{F}_{t-1}]$$

and

$$\begin{aligned}
\|D_t\|_\infty &= \|\mathbb{E}[\overline{g}_{T+1}|\mathcal{F}_t] - \mathbb{E}[\overline{g}_{T+1}|\mathcal{F}_{t-1}]\|_\infty \\
&= \|\mathbb{E}[\overline{g}_{T+1} - \overline{g}_{T+1}^t|\mathcal{F}_t]\|_\infty \\
&\leq \mathbb{E}[\|\overline{g}_{T+1} - \overline{g}_{T+1}^t\|_\infty|\mathcal{F}_t] \\
&= \|\overline{g}_{T+1} - \overline{g}_{T+1}^t\|_\infty.
\end{aligned}$$

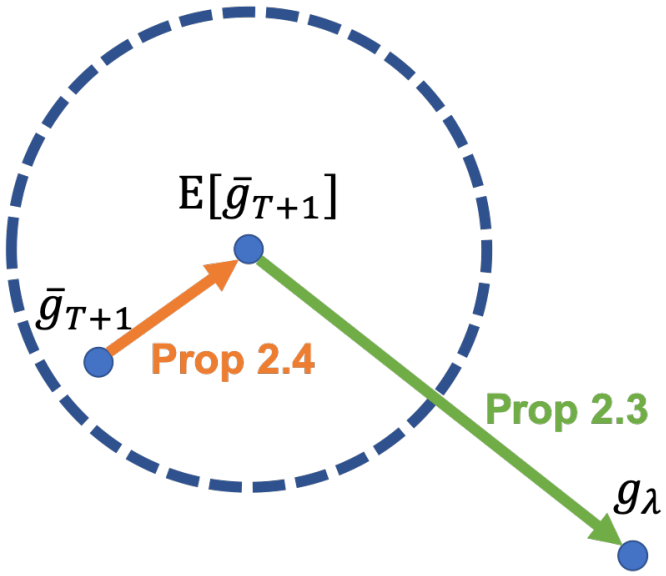


Fig. 2.4: Convergence of $\overline{g}_{T+1}$.

Thus, if $\|\overline{g}_{T+1} - \overline{g}^t_{T+1}\|_{\mathcal{H}}$ is uniformly bounded over $Z_1, \ldots, Z_T, Z'_t$ for all $1 \leq t \leq T$, the existence of c_T can be shown. To show the concentration with respect to T, it is essential for c_T to be an decreasing function of T. Indeed, we can derive c_T on the order of $1/\sqrt{T}$ by utilizing a stability of SGD on strongly convex objectives. The full proof is found in Appendix C.

We note that similar arguments about generalization analysis using a stability of a learning algorithm is discussed in [55] and that of SGD is also derived in [27] for another purpose. Intuitively, if an output of an algorithm has small deviation when one of training samples are replaced with another one, it means that the output generalizes well. The main characteristic of these arguments is that it does not depend on a complexity of a hypothesis class, which is necessary for analyses based on the uniform convergence (e.g., [54]).

Remark. We could derive c_T in Lemma 2.1 in a similar manner even if the objective function is not strongly convex. However the resulting c_T is increasing in T in the absence of strong convexity [27], so the meaningful convergence cannot be obtained. This is the reason why we deal with the regularization term in the objective (2.5).

2.4 Positioning of the Thesis

One of the main purposes of the thesis is to show the effectiveness of SGD in terms of the convergence speed of the excess classification error. Since the convergence of the expected risk derived from Proposition 2.3 is $O(1/\sqrt{T})$, straightforward analysis introduced in Section 2.1 yields equal or slower rate than $O(1/\sqrt{T})$ of the excess classification error. Despite this fact, recent studies [43, 42] have shown that the exponential convergence of the excess classification error is achieved by SGD under the strong low noise condition. The main idea behind their proof is analyzing the convergence of the excess classification error not via the convergence of the excess risk (as introduced in Section 2.1), but via the convergence of the hypothesis, which is described in Proposition 2.3 and Proposition 2.4. This idea is originally used in the classical analysis of the classification problem

[6], which showed that the ERM classifier achieves the fast convergence of the excess classification error under the low noise condition. Utilizing these ideas, we demonstrate a fast convergence of the excess classification under the low noise condition in Section 3. The relation to the previous results is shown in Table 2.1.

	Low noise	Strong low noise
ERM	[63, 6]	[31, 6]
SGD	Chapter 3	[43, 42]

Table 2.1: Previous research about classification problems.

From practical perspectives, the original kernel methods require $O(n^2)$ time and $O(n)$ space complexity even if applying SGD as an optimization method. Thus, in many situations, we do not deal with the original RKHS. In Section 4, we consider applying random features [47], which is the most widely adopted kernel approximation method. Combining that with SGD yields $O(nM)$ time and $O(M)$ space complexity, where M denotes the number of features, that is, the dimension of kernel approximation. Thus it is frequently applied in practical large-scale learning problems where n is very large. From a theoretical perspective, there is a trade-off between the error induced by the approximation and its computational cost and it has been shown that random features actually reduce computational cost in a regression setting [16]. In particular, they showed that $O(\sqrt{n}\log n)$ features are sufficient to achieve the ordinal $O(1/\sqrt{n})$ learning rate of the expected risk. However, for classification setting, it has only been shown that $O(n)$ features are required to achieve $O(1/\sqrt{n})$ rate of the expected risk, which means computational gain comes at expense of the approximation accuracy. In Section 4, we show that if we consider the convergence of the excess classification error, the constant number of features is sufficient to achieve the exponential convergence under the strong low noise condition. The relation to the previous research about SGD on RKHS is shown in Table 2.2.

	Original RKHS	Random features
Regression	[22, 44]	[16]
Classification (strong low noise)	[43, 42]	Chapter 4

Table 2.2: Previous research about SGD and random features.

Chapter 3

Stochastic Gradient Descent under Low Noise Condition

3.1 Motivation

In this study, we consider optimizing binary classification problems on RKHS using stochastic gradient descent (SGD). As introduced in Section 2.1, the final objective of classification problems is to minimize expected classification errors and we usually do that by optimizing the convex surrogate loss instead of 0-1 loss. Such an approximation is theoretically justified by the consistency property [10, 66] and this property is satisfied for many practically used loss functions.

When one considers convergence rates of excess classification errors, it can be simply derived from those of excess risks and it can be accelerated by assuming the low noise condition as discussed in Section 2.1. However, the resulting convergence rates of excess classification errors are generally slower than those of the excess risks as shown in (2.4). Since the optimal convergence rates of excess risks are known to $O(1/\sqrt{T})$ and $O(1/T)$, where T denotes the number of iterations, for convex and strongly convex objectives respectively [1, 39], it means that the convergence of excess classification errors cannot be faster than these rates in general.

In [6], it is shown that using empirical risk minimization (ERM) under the low noise condition and additional assumptions, the convergence rates of excess classification errors can be faster than those of excess risks and can be *super-fast*, that is, faster than $O(1/n)$ where n denotes the sample size. However, these results are somewhat inadequate to explain the practical success of SGD and it remains unclear such a fast convergence can be achieved by SGD.

Recently, it has been shown that SGD on RKHS achieves the exponential convergence rate of excess classification errors in the latter half of the training by assuming the strong low noise condition holds. In particular, [43] showed that for squared loss function and [42] extended that to general smooth loss functions including logistic loss. However, the strong low noise condition is somewhat restrictive, since it does not permit a conditional probability on any sample to stay around $1/2$.

Contributions. In this study, we extend the results in [42] and show that the super-fast convergence rates of excess classification errors are actually achieved by SGD under the general (not strong) low noise condition when optimizing general smooth loss functions. Also, our fast rates hold for all iterations T unlike the exponential rates under the strong low noise condition, where such fast convergence starts from the middle of the training.

Chapter organization. This chapter is organized as follows. In Section 3.2, the problem setting of binary classification and the algorithm of SGD treated in this study are briefly reviewed. In Section 3.3, we present our primary result describing the fast convergence rate of the classification error under the low noise condition. Next, a more concrete condition that is required to derive fast convergence is investigated in Section 3.4.

3.2 Problem Setting

In this section, we briefly describe a problem setting and assumptions for the binary classification. Let $\mathcal{X}$ and $\mathcal{Y} = \{-1, 1\}$ be a feature space and the set of binary labels, respectively; ρ denotes a probability measure on $\mathcal{X} \times \mathcal{Y}$, by $\rho_{\mathcal{X}}$ the marginal distribution on X, and by $\rho(\cdot|X)$ the conditional distribution on Y, where $(X, Y) \sim \rho$.

As introduced in Chapter 2, our final objective is to minimize the expected classification error $\mathcal{R}(g)$ defined below amongst all measurable functions:

$$\mathcal{R}(g) = \mathbb{E}_{(X,Y)\sim\rho}\left[I(\mathrm{sgn}(g(X)), Y)\right], \tag{3.1}$$

where $\mathrm{sgn}(x) = 1$ if $x > 0$ and -1 otherwise, and I represents 0-1 loss:

$$I(y, y') = \begin{cases} 1 & (y \neq y') \\ 0 & (y = y'). \end{cases}$$

By definition, $g(x) = \mathbb{E}[Y|x] = 2\rho(1|x) - 1$ minimizes $\mathcal{R}$. However, directly minimizing (3.1) to obtain the Bayes classifier is intractable because of its non-convexity. Thus, we generally use the convex surrogate loss $l(\zeta, y)$ instead of the 0-1 loss and minimize the expected risk $\mathcal{L}(g)$ of l:

$$\mathcal{L}(g) = \mathbb{E}_{(X,Y)\sim\rho}\left[l(g(X), Y)\right]. \tag{3.2}$$

In general, the loss function l has a form $l(\zeta, y) = \phi(\zeta y)$ where $\phi : \mathbb{R} \to \mathbb{R}$ is a non-negative convex function. The typical examples are logistic loss, where $\phi(v) = \log(1 + \exp(-v))$ and hinge loss, where $\phi(v) = \max\{0, 1 - v\}$. Minimizing the expected risk (3.2) ensures minimizing the expected classification error (3.1) if l is *classification-calibrated* [10], which has been proven for several practically implemented losses including hinge loss and logistic loss.

As in Chapter 2, we consider a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated with a positive definite kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the hypothesis space. Recall that it satisfies $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}$ holds for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$, where $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ denotes the inner product of $\mathcal{H}$. In addition, let $\|\cdot\|_{\mathcal{H}}$ denote the norm of $\mathcal{H}$ induced by the inner product. Under these settings, we attempt to solve the following minimization problem:

$$\min_{g\in\mathcal{H}} \left\{ \mathcal{L}_\lambda(g) \stackrel{\mathrm{def}}{=} \mathcal{L}(g) + \frac{\lambda}{2}\|g\|_{\mathcal{H}}^2 \right\}. \tag{3.3}$$

We consider solving the objective (3.3) using averaged stochastic gradient descent, which is described in Algorithm 3.1. Recall that a stochastic gradient of $\mathcal{L}_\lambda$ at $g \in \mathcal{H}$ on sample $Z = (X, Y) \in \mathcal{X} \times \mathcal{Y}$ is given as follows:

$$G_\lambda(g, Z) \stackrel{\mathrm{def}}{=} \partial_\zeta l(g(X), Y)k(\cdot, X) + \lambda g.$$

As introduced in Section 2.2.2, we consider the following learning rates η_t and averaging weights α_t:

$$\alpha_t = \frac{2(\gamma + t - 1)}{(2\gamma + T)(T + 1)}, \quad \eta_t = \frac{2}{\lambda(\gamma + t)}.$$

where $\gamma > 0$ is an offset parameter for the time index.

Algorithm 3.1 Averaged Stochastic Gradient Descent (ASGD)

Input: regularization parameter λ, number of iterations T, learning rates $\{\eta_t\}_{t=1}^{T}$, averaging weights $\{\alpha_t\}_{t=1}^{T+1}$.
Output: classifier $\overline{g}_{T+1}$
  Initialize $g_1 \in \mathcal{H}$
  for $t = 1, \ldots, T$ **do**
    Randomly draw a sample $z_t = (x_t, y_t) \sim \rho$
    $g_{t+1} \leftarrow g_t - \eta_t G_\lambda(g_t, z_t)$
  end for
  return $\overline{g}_{T+1} = \sum_{t=1}^{T+1} \alpha_t g_t$

To ensure the convergence of SGD, we make several assumptions. First, we impose the smoothness of loss functions and boundedness of kernel functions as introduced in Section 2.3.

Assumption 3.1. *$l(\cdot, y)$ is convex, differentiable, G-Lipschitz continuous and L-Lipschitz smooth. That is, for any $\zeta, \zeta' \in \mathbb{R}$ and $y \in \mathcal{Y}$,*

$$|\partial_\zeta l(\zeta, y)| \leq G,$$
$$|\partial_\zeta l(\zeta, y) - \partial_\zeta l(\zeta', y)| \leq L|\zeta - \zeta'|.$$

Assumption 3.2. *There exists $R > 0$ such that $k(x, x) \leq R^2$ for any $x \in \mathcal{X}$.*

To characterize the Bayes rule (the minimizer of $\mathcal{L}$), we impose several assumptions on ϕ. Recall the definition of the link function, which connects the hypothesis space and the probability measure:

$$h_*(\mu) = \arg\min_{\alpha \in \mathbb{R}} \{\mu\phi(\alpha) + (1 - \mu)\phi(-\alpha)\}.$$

In addition, its corresponding value is denoted by l_*:

$$l_*(\mu) = \min_{\alpha \in \mathbb{R}} \{\mu\phi(\alpha) + (1 - \mu)\phi(-\alpha)\}.$$

Although $h_*(\mu)$ may not be uniquely determined nor well-defined in general, the following assumption ensures these properties.

Assumption 3.3. *$\rho(1|X)$ takes values in $(0, 1)$, $\rho_\mathcal{X}$-almost surely; ϕ is differentiable and h_* is well-defined, L'-Lipschitz continuous, differentiable, monotonically increasing, and invertible over $(0, 1)$. Moreover, it follows that*

$$\operatorname{sgn}(\mu - 1/2) = \operatorname{sgn}(h_*(\mu)).$$

As introduced in Section 2.1, for logistic loss we have $h_*(\mu) = \log(\mu/(1 - \mu))$ and the above condition is satisfied with $L' = 4$. Since it is known that l_* is a concave function, we

introduce Bregman divergence for concave function l_* to ensure the uniqueness of Bayes rule g_* of $\mathcal{L}$:

$$d_{l_*}(\eta_1, \eta_2) = -l_*(\eta_2) + l_*(\eta_1) + l'_*(\eta_1)(\eta_2 - \eta_1).$$

Assumption 3.4. *Bregman divergence d_{l_*} derived by l_* is positive, that is, $d_{l_*}(\eta_1, \eta_2) = 0$ if and only if $\eta_1 = \eta_2$. For the expected risk $\mathcal{L}$, a unique Bayes rule g_* (up to zero measure sets) exists in $\mathcal{H}$.*

For logistic loss, we have shown in Section 2.1 that d_{l_*} coincides with Kullbuck-Leibler divergence, and thus, the positivity of the divergence holds. Here recall that if ϕ is differentiable and h_* is differentiable and invertible, the excess risk can be expressed using d_{l_*} from Proposition 2.2:

$$\mathcal{L}(g) - \mathcal{L}(g_*) = \mathbb{E}_X[d_{l_*}(h_*^{-1}(g(X)), \rho(1|X))].$$

Thus combining Assumptions 3.3 and 3.4 implies that Bayes rule g_* is equal to $h_*(\rho(1|X))$, $\rho_{\mathcal{X}}$-almost surely and contained in the original RKHS $\mathcal{H}$.

The next is the low noise condition on sample distribution introduced in Section 2.1.

Assumption 3.5. *The probability distribution ρ satisfies the low noise condition. That is, there exists $C_1, \alpha > 0$ such that the following inequality holds for all $\delta > 0$:*

$$\mathbb{P}\left[|\rho(Y = 1|X) - 1/2| \leq \delta\right] \leq C_1 \delta^{\alpha}.$$

Finally, we characterize the induced bias by the regularization parameter λ on generated hypothesis g_λ defined as follows:

$$g_\lambda \stackrel{\text{def}}{=} \operatorname*{arg\,min}_{g \in \mathcal{H}} \mathcal{L}_\lambda(g).$$

Assumption 3.6. *There exists $C_2, \kappa > 0$ such that the following inequality holds for all $\lambda > 0$:*

$$\|g_\lambda - g_*\|_{L^\infty(d\rho_{\mathcal{X}})} \leq C_2 \lambda^{\kappa}.$$

Although it can be shown that arbitrary small bias is achieved with sufficiently small regularization parameter λ (Proposition 3.1), generally it is difficult to show such a convergence has a order parameter κ. We investigate the several sufficient condition to satisfy Assumption 3.6 in Section 3.4.

3.3 Main Results

Here we present our main result, which shows that the convergence rate of the excess classification errors can be faster by assuming the low noise condition.

Theorem 3.1. *Suppose Assumption 3.1-3.6 holds. Consider Algorithm 4.1 with $\lambda = \left(288 C_2^{-2} G^2 R^4\right)^{\frac{1}{2+2\kappa}} T^{-\frac{1}{2+2\kappa}}$, $\eta_t = \frac{2}{\lambda(\gamma+t)}$ and $\alpha_t = \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)}$ where γ is a positive value such that $\|g_1\|_{\mathcal{H}} \leq (2\eta_1 + 1/\lambda)GR$ and $\eta_1 \leq \min\{1/LR^2, 1/2\lambda\}$. Then there exists constant $C > 0$ such that the following inequality holds:*

$$\mathbb{E}\left[\mathcal{R}(\overline{g}_{T+1})\right] - \mathcal{R}(g_*) \leq C T^{-\frac{(\alpha+1)\kappa}{2+2\kappa}}.$$

Proof is found in Appendix D. We can see that when $\alpha > 1 + 2/\kappa$, the resulting convergence rate is faster than $1/T$ (super fast rate). Note that the convergence rates

derived by the one of excess risk is the following form (Theorem 10 in [10]):

$$C\left(\mathcal{R}(g)-\mathcal{R}_*\right)^{1-\frac{1}{\alpha}}\psi\left(\frac{\left(\mathcal{R}(g)-\mathcal{R}_*\right)^{\frac{1}{\alpha}}}{2C}\right)\leq\mathcal{L}(g)-\mathcal{L}_*,$$

where ψ is a transform function defined in Section 2.1. As discussed in Section 2.1, the obtained convergence rate of excess classification errors from the above inequality does not be faster than the one of excess risk even if we set α to ∞. Thus this never gives a rate faster than $1/T$ since the convergence of right hand side is slower than $1/T$ in general.

We also note that derived fast rate under low noise condition is valid for all $T\geq 1$, whereas under the strong low noise condition, the exponential convergence only occurs in the final phase of learning [42, 43].

3.4 Dependence on Regularization Parameter

In this section, we consider a sufficient condition to satisfy Assumption 3.6 and discuss its validity. First we describe a result from [42], which provides the consistency of g_λ as λ goes to 0.

Proposition 3.1 (Proposition A in [42])**.** *Suppose Assumption 3.2, 3.5, 3.6 hold. Then, for any ϵ there exists sufficiently small $\lambda>0$ such that $\|g_\lambda-g_*\|_{L^\infty(d\rho_\mathcal{X})}\leq\epsilon$.*

Although g_λ converges to g_* as $\lambda\to 0$ as shown in above, specifying its convergence rate is difficult in general. To derive its rate, first we need the *local strong convexity*, which is a strong convexity on a arbitrary compact set.

Assumption 3.7. *$\phi:\mathbb{R}\to\mathbb{R}$ is $\mu(U)$-strongly convex on a bounded set $[-U,U]\subset\mathbb{R}$, i.e.,*

$$\phi(\zeta_1)-\phi(\zeta_2)-\phi'(\zeta_2)(\zeta_1-\zeta_2)\geq\frac{\mu(U)}{2}(\zeta_1-\zeta_2)^2.$$

holds for any $\zeta_1,\zeta_2\in[-U,U]$.

For logistic loss, we have $\phi''(\xi)=\frac{1}{2+e^{\xi}+e^{-\xi}}$ and thus Assumption 3.7 is satisfied with $\mu(U)=\frac{1}{2+e^{U}+e^{-U}}$. Furthermore, to ensure the convergence in terms of L^∞-norm, we impose the norm condition on RKHS $\mathcal{H}$.

Assumption 3.8. *There exists $0\leq p<1$, and a constant $C_3>0$ depends on $0<\delta\leq 1$ that satisfies, for any $f\in\mathcal{H}$,*

$$\|f\|_{L^\infty(d\rho_\mathcal{X})}\leq C_3\|f\|_\mathcal{H}^p\|f\|_{L^2(d\rho_\mathcal{X})}^{1-p}.$$

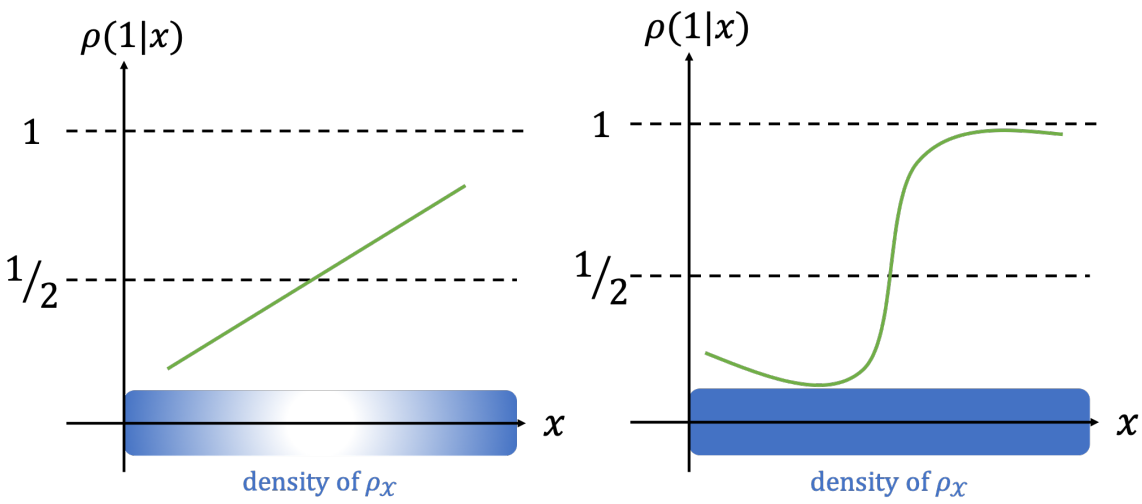
This condition is common in the analysis of kernel methods ([59, 35]) and has an important relation to interpolation spaces, see Proposition B.1. If k is Gaussian kernel and $\rho_\mathcal{X}$ has a density with respect to Lebesgue measure which is uniformly bounded from 0 and ∞, Theorem 4.2 in Chapter 4 shows that it holds with any $0<p<1$.

Theorem 3.2. *Suppose Assumption 3.2, 3.7 and 3.8 holds. Then it holds that*

$$\|g_\lambda-g_*\|_{L^\infty(d\rho_\mathcal{X})}\leq 2^pC_3\|g_*\|_\mathcal{H}\left(\frac{\lambda}{2\mu(R\|g_*\|_\mathcal{H})}\right)^{\frac{1-p}{2}}.$$

Thus Assumption 3.6 is satisfied with $\kappa=\frac{1-p}{2}$.

Proof is found in Appendix D. Applying these results to Theorem 3.1, we obtain the following result.



(a) (b)

Fig. 3.1: Two different situations of the low noise condition.

Corollary 3.1. *Suppose Assumption 3.1-3.5, 3.7 and 3.8 holds. Consider Algorithm 4.1 with* $\lambda \approx T^{-\frac{1}{3-p}}$, $\eta_t = \frac{2}{\lambda(\gamma+t)}$ *and* $\alpha_t = \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)}$ *where* γ *is a positive value such that* $\|g_1\|_{\mathcal{H}} \leq (2\eta_1 + 1/\lambda)GR$ *and* $\eta_1 \leq \min\{1/LR^2, 1/2\lambda\}$. *Then there exists constant* $C > 0$ *such that the following inequality holds:*

$$\mathbb{E}\left[\mathcal{R}(\overline{g}_{T+1})\right] - \mathcal{R}(g_*) \leq CT^{-\frac{(\alpha+1)(1-p)}{6-2p}}.$$

Remark. There are two cases that satisfies the low noise condition as shown in Figure 3.1. If we assume that the conditional probability $\rho(1|x)$ is smooth, the density of $\rho_{\mathcal{X}}$ must be vanished around the point where $\rho(1|x)$ crosses the level $1/2$ (a). On the other hand, if we assume $\rho_{\mathcal{X}}$ has a density which is uniformly bounded away from 0, the conditional probability $\rho(1|x)$ must take off the level $1/2$ abruptly (b). Thus, when we consider this density condition to satisfy Assumption 3.8, the regression objective $\rho(1|x)$ becomes to be less likely to be contained in RKHS. In the case where $\mathcal{H}$ is a function class with β-Hölder smoothness, [6] shows the trade-off of smoothness parameter β and low noise parameter α:

Proposition 3.2 (Proposition 3.4 in [6])**.** *Suppose* $\rho_{\mathcal{X}}$ *has a density with respect to Lebesgue measure which is uniformly bounded away from 0 and* ∞ *and* $\rho(1|\cdot)$ *is* β*-Hölder smooth. In addition, suppose Assumption 3.6 is satisfied with* α *such that* $\min\{\alpha, \alpha\beta\} > 1$. *Then* $\rho(1|\cdot)$ *cannot crosses the level* $1/2$ *in the interior of the support of* $\rho_{\mathcal{X}}$.

We note that there exists a distribution which satisfy the above condition but never crosses the level $1/2$. Example of such distribution ρ is given in [6], such as the one with $\rho_{\mathcal{X}}$ is the uniform distribution on a ball centered at 0 in $\mathbb{R}^d$ and $\rho(1|x) = 1/2 - C\|x\|^2$ with an appropriate $C > 0$. In this case, $\rho(1|\cdot)$ belongs to β-Hölder class with arbitrary large β and Assumption 3.5 is satisfied with $\alpha = d/2$. Further investigation about the trade-off between norm condition (Assumption 3.8) and the low noise condition (Assumption 3.5) is an important future work.

3.5 Conclusion

In this study, we show that the convergence of the expected classification errors faster than $O(1/n)$ is achievable with SGD under such a condition. This result is rather surprising, since it is known that the optimal convergence rates of the expected risk is known to $O(1/\sqrt{n})$ or $O(1/n)$. Additionally, since we need additional assumptions to derive such fast rates, we investigate sufficient conditions to satisfy these assumptions and discuss trade offs between assumptions.

The primal future direction is further investigations and characterizations of such trade offs. Additionally, it could also be interesting to explore the convergence speed of more sophisticated variants of SGD, such as stochastic accelerated methods and stochastic variance reduced methods [52, 30, 20, 2].

Chapter 4

Learning with SGD and Random Features under Strong Low Noise Condition

4.1 Motivation

Kernel methods are commonly used to solve a wide range of problems in machine learning, as they provide flexible non-parametric modeling techniques and come with well-established theories about their statistical properties [15, 59, 35]. However, computing estimators in kernel methods can be prohibitively expensive in terms of memory requirements for large datasets.

There are two popular approaches to scaling up kernel methods. The first is sketching, which reduces data-dimensionality by random projections. A random features method [47] is a representative, which approximates a reproducing kernel Hilbert space (RKHS) by a finite-dimensional space in a data-independent manner. The second is stochastic gradient descent (SGD), which allows data points to be processed individually in each iteration to calculate gradients. Both of these methods are quite effective in reducing memory requirements and are widely used in practical tasks.

For the theoretical properties of random features, several studies have investigated the approximation quality of kernel functions [57, 61, 62], but only a few have considered the generalization properties of learning with random features. For the regression problem, its generalization properties in ERM and SGD settings, respectively, have been studied extensively in [50] and [16]. In particular, they showed that $O(\sqrt{n}\log n)$ features are sufficient to achieve the usual $O(1/\sqrt{n})$ learning rate, indicating that there is a computational benefit to using random features.

However, it remains unclear whether or not it is computationally efficient for other tasks. In [48], the generalization properties were studied with Lipschitz loss functions under ℓ_∞-constraint in hypothesis space, and it was shown that $O(n\log n)$ features are required for $O(1/\sqrt{n})$ learning bounds. Also, in [33], learning with Lipschitz loss and standard regularization was considered instead of ℓ_∞-constraint, and similar results were attained. Both results suggest that computational gains come at the expense of learning accuracy if one considers general loss functions.

In this study, learning classification problems with random features and SGD are considered, and the generalization property is analyzed in terms of the *classification error*. Recently, it was shown that the convergence rate of the excess classification error can be made exponentially faster by assuming the *strong low-noise condition* [63, 31] that conditional label probabilities are uniformly bounded away from 1/2 [43, 42]. We extend these

analyses to a random features setting to show that the exponential convergence is achieved if a sufficient number of features are sampled. Unlike when considering the convergence of loss function, the resulting convergence rate of the classification error is independent of the number of features. In other words, an arbitrary small classification error is achievable as long as there is a sufficient number of random features. So our result suggests that there is indeed a computational benefit to use random features in classification problems under the strong low-noise condition.

Remark. Although several studies consider the optimal sampling distributions of features in terms of the worst-case error [9, 7, 33], we do not explore this direction and treat the original random features algorithm because these distributions are generally intractable or require much computational cost to sample [9]. Note that for the case of Gaussian kernel, a constant time sampling algorithm from the optimized distribution is proposed in [7], but their proposed distribution depends on the sample size n and the goal of their analysis is a good approximation of a gram matrix on given data points. In contrast to this, in our one-pass SGD setting, the sample size cannot be determined beforehand and SGD is aimed to optimize a population risk, not an empirical risk, so we cannot apply this distribution directly. Moreover, there is still a gap to apply those analysis to general kernel functions because they utilized the exponential decaying property about eigenvalues of Gaussian kernel.

In addition, we should refer to Nyström method [64], which is also a popular method to scale up kernel methods. In contrast to random features, Nyström method approximates kernel function in data-dependent way. As a result, similar to calculating an optimized sampling distribution on random features, Nyström method also requires data points before actual training starts and needs $O(nM)$ memory, which is more expensive than $O(M)$ in random features. These are reasons why we dealt with original algorithm of random features in this study.

Contributions. Our contributions are twofold. First, we analyze the error induced by the approximation of random features in terms of the distance between the generated hypothesis including population risk minimizers and empirical risk minimizers when using general Lipschitz loss functions. Our results can be framed as an extension of the analysis in [17, 61], which analyzed the error in terms of the distance between empirical risk minimizers when using a hinge loss.

Second, using the above result, we prove that the exponential convergence rate of the excess classification error under the strong low-noise condition is achieved if a sufficient number of features are sampled. Based on this, we show that there is a significant computational gain in using random features rather than a full kernel method for obtaining a relatively small classification error.

Chapter organization. This chapter is organized as follows. In Section 4.2, the algorithm of random features and SGD treated in this study are briefly reviewed. In Section 4.3, an error analysis of the generated hypothesis using random features is presented, after which a more sophisticated analysis is given for the case of a Gaussian kernel. Our primary result describing the exponential convergence rate of the classification error is given in Section 4.4. Finally, numerical experiments using synthetic datasets are presented in Section 4.5.

4.2 Problem Setting

In this section, we provide notations to describe a problem setting and assumptions for the binary classification and kernel method treated in this study.

4.2.1 Binary Classification Problem

Let $\mathcal{X}$ and $\mathcal{Y} = \{-1, 1\}$ be a feature space and the set of binary labels, respectively; ρ denotes a probability measure on $\mathcal{X} \times \mathcal{Y}$, by $\rho_{\mathcal{X}}$ the marginal distribution on X, and by $\rho(\cdot|X)$ the conditional distribution on Y, where $(X, Y) \sim \rho$.

In the classification problem, our final objective is to choose a discriminant function $g : \mathcal{X} \to \mathbb{R}$ such that the sign of $g(X)$ is an accurate prediction of Y. Therefore, we intend to minimize the expected classification error $\mathcal{R}(g)$ defined below amongst all measurable functions:

$$\mathcal{R}(g) = \mathbb{E}_{(X,Y)\sim\rho}\left[I(\mathrm{sgn}(g(X)), Y)\right], \tag{4.1}$$

where $\mathrm{sgn}(x) = 1$ if $x > 0$ and -1 otherwise, and I represents 0-1 loss:

$$I(y, y') = \begin{cases} 1 & (y \neq y') \\ 0 & (y = y'). \end{cases}$$

By definition, $g(x) = \mathbb{E}[Y|x] = 2\rho(1|x) - 1$ minimizes $\mathcal{R}$. However, directly minimizing (4.1) to obtain the Bayes classifier is intractable because of its non-convexity. Thus, we generally use the convex surrogate loss $l(\zeta, y)$ instead of the 0-1 loss and minimize the expected risk $\mathcal{L}(g)$ of l:

$$\mathcal{L}(g) = \mathbb{E}_{(X,Y)\sim\rho}\left[l(g(X), Y)\right]. \tag{4.2}$$

In general, the loss function l has a form $l(\zeta, y) = \phi(\zeta y)$ where $\phi : \mathbb{R} \to \mathbb{R}$ is a non-negative convex function. The typical examples are logistic loss, where $\phi(v) = \log(1 + \exp(-v))$ and hinge loss, where $\phi(v) = \max\{0, 1 - v\}$. Minimizing the expected risk (4.2) ensures minimizing the expected classification (4.1) if l is *classification-calibrated* [10], which has been proven for several practically implemented losses including hinge loss and logistic loss.

4.2.2 Kernel Methods and Random Features

In this study, we consider a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated with a positive definite kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the hypothesis space. It is known [3] that a positive definite kernel k uniquely defines its RKHS $\mathcal{H}$ such that the reproducing property $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}$ holds for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$, where $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ denotes the inner product of $\mathcal{H}$. Let $\|\cdot\|_{\mathcal{H}}$ denote the norm of $\mathcal{H}$ induced by the inner product. Under these settings, we attempt to solve the following minimization problem:

$$\min_{g\in\mathcal{H}} \ \mathcal{L}(g) + \frac{\lambda}{2}\|g\|_{\mathcal{H}}^2 \tag{4.3}$$

where $\lambda > 0$ is a regularization parameter.

However, because solving the original problem (4.3) is usually computationally inefficient for large-scale datasets, the approximation method is applied in practice. Random

features [47] is a widely used method for scaling up kernel methods because of its simplicity and ease of implementation. Additionally, it approximates the kernel in a data-independent manner, making it easy to combine with SGD. In random features, a kernel function k is assumed to have the following expansion in some space Ω with a probability measure τ and a feature function $\varphi : \mathcal{X} \times \Omega \to \mathbb{C}$:

$$k(x, y) = \int_{\Omega} \varphi(x, \omega)\overline{\varphi(y, \omega)} d\tau(\omega). \tag{4.4}$$

The main idea behind random features is to approximate the integral (4.4) by its Monte-Carlo estimate:

$$k_M(x, y) = \frac{1}{M} \sum_{i=1}^{M} \varphi(x, \omega_i)\overline{\varphi(y, \omega_i)}, \quad \omega_i \overset{i.i.d.}{\sim} \tau. \tag{4.5}$$

For example, if k is a shift invariant kernel, by Bochner's theorem (Proposition A.2), the expansion (4.4) is achieved with $\varphi(x, \omega) = C' e^{i\omega^\top x}$, where C' is a normalization constant. Then, the approximation (4.5) is called random Fourier features [47], which is the most widely used variant of random features.

We denote the RKHS associate with k and k_M by $\mathcal{H}$ and $\mathcal{H}_M$, respectively. These spaces then admit the following explicit representation [9, 8]:

$$\mathcal{H} = \left\{ \int_{\Omega} \beta(\omega)\varphi(\cdot, \omega) d\tau(\omega) \;\middle|\; \beta \in L^2(d\tau) \right\},$$
$$\mathcal{H}_M = \left\{ \sum_{i=1}^{M} \frac{\beta_i}{\sqrt{M}} \varphi(\cdot, \omega_i) \;\middle|\; |\beta_i| < \infty \right\}.$$

We note that the approximation space $\mathcal{H}_M$ is not necessarily contained in the original space $\mathcal{H}$. For $g \in \mathcal{H}$ and $h \in \mathcal{H}_M$, the following RKHS norm relations hold:

$$\|g\|_{\mathcal{H}} = \inf \left\{ \|\beta\|_{L^2(d\tau)} \;\middle|\; g = \int_{\Omega} \beta(\omega)\varphi(\cdot, \omega) d\tau(\omega) \right\},$$
$$\|h\|_{\mathcal{H}_M} = \inf \left\{ \|\beta\|_2 \;\middle|\; h = \sum_{i=1}^{M} \frac{\beta_i}{\sqrt{M}} \varphi(\cdot, \omega_i) \right\}.$$

As a result, the problem (4.3) in the approximation space $\mathcal{H}_M$ is reduced to the following generalized linear model:

$$\min_{\beta \in \mathbb{R}^M} \; \mathcal{L}(\beta^\top \phi_M) + \frac{\lambda}{2} \|\beta\|_2^2 \tag{4.6}$$

where ϕ_M is a feature vector:

$$\phi_M = \frac{1}{\sqrt{M}} [\varphi(\cdot, \omega_1), \ldots, \varphi(\cdot, \omega_M)]^\top.$$

In this paper, we consider solving the problem (4.6) using the averaged SGD. The details are discussed in the following section.

4.2.3 Averaged Stochastic Gradient Descent

SGD is the most popular method to solve large scale learning problems. In this section, we discuss a specific form of SGD following Section 2.2.2. For the optimization problem (4.6), its gradient with respect to β is given as follows:

$$\mathbb{E}\left[\partial_\zeta l(\beta^\top \phi_M(X), Y)\phi_M(X) + \lambda\beta\right],$$

where ∂_ζ is a partial derivative with respect to the first variable of l. Thus, the stochastic gradient with respect to β is given by $\partial_\zeta l(\beta^\top \phi_M(X), Y)\phi_M(X) + \lambda\beta$. We note that the update on the β parameter corresponds to the update on the function space $\mathcal{H}_M$ instead on $\mathcal{H}$ as introduced in Section 2.2.2, because a gradient on $\mathcal{H}_M$ is given by

$$\mathbb{E}\left[\partial_\zeta l(\beta^\top \phi_M(X), Y)\phi_M(X) + \lambda\beta\right]^\top \phi_M.$$

The algorithm of random features and averaged SGD is described in Algorithm 4.1.

Algorithm 4.1 Random Features + SGD

Input: number of features M, regularization parameter λ, number of iterations T, learning rates $\{\eta_t\}_{t=1}^T$, averaging weights $\{\alpha_t\}_{t=1}^{T+1}$
Output: classifier $\overline{g}_{T+1}$
Randomly draw feature variables $\omega_1, \ldots, \omega_M \sim \tau$
Initialize $\beta_1 \in \mathbb{R}^M$
for $t = 1, \ldots, T$ **do**
  Randomly draw samples $(x_t, y_t) \sim \rho$
  $\beta_{t+1} \leftarrow \beta_t - \eta_t \left(\partial_\zeta l(\beta_t^\top \phi_M(x_t), y_t)\phi_M(x_t) + \lambda\beta_t\right)$
end for
$\overline{\beta}_{T+1} = \sum_{t=1}^{T+1} \alpha_t \beta_t$
return $\overline{g}_{T+1} = \overline{\beta}_{T+1}^\top \phi_M$

As introduced in Section 2.2.2, we set the learning rate and the averaging weight as follows:

$$\eta_t = \frac{2}{\lambda(\gamma + t)}, \quad \alpha_t = \frac{2(\gamma + t - 1)}{(2\gamma + T)(T + 1)},$$

where γ is an offset parameter for the time index. We note that an averaged iterate $\overline{\beta}_t$ can be updated iteratively as follows:

$$\overline{\beta}_1 = \beta_1,$$
$$\overline{\beta}_{t+1} = (1 - \theta_t)\overline{\beta}_t + \theta_t \beta_{t+1}, \quad \theta_t = \frac{2(\gamma + t)}{(t + 1)(2\gamma + t)}.$$

Using this formula, we can compute the averaged output without storing all internal iterate $(\beta_t)_{t=1}^{T+1}$.

4.2.4 Computational Complexity

If we assume the evaluation of a feature map $\varphi(x, \omega)$ to have a constant cost, one iteration in Algorithm 4.1 requires $O(M)$ operations. As a result, one pass SGD on n samples requires $O(Mn)$ computational time. On the other hand, the full kernel method without

approximation requires $O(n)$ computations per iteration; thus, the overall computation time is $O(n^2)$, which is much more expensive than random features.

For the memory requirements, random features needs to store M coefficients, and it does not depend on the sample size n. On the other hand, we have to store n coefficients in the full kernel method, so it is also advantageous to use random features in large-scale learning problems.

4.3 Error Analysis of Random Features

Our primary purpose here is to bound the distance between the hypothesis generated by solving the problems in each space $\mathcal{H}$ and $\mathcal{H}_M$. Population risk minimizers in spaces $\mathcal{H}, \mathcal{H}_M$ are defined as below:

$$g_\lambda = \operatorname*{arg\,min}_{g \in \mathcal{H}} \left(\mathcal{L}(g) + \frac{\lambda}{2} \|g\|_{\mathcal{H}}^2 \right),$$
$$g_{M,\lambda} = \operatorname*{arg\,min}_{g \in \mathcal{H}_M} \left(\mathcal{L}(g) + \frac{\lambda}{2} \|g\|_{\mathcal{H}_M}^2 \right).$$

The uniqueness of minimizers is guaranteed by the regularization term.

First, the $L^\infty(d\rho_\mathcal{X})$-norm is bound between g_λ and $g_{M,\lambda}$ when the loss function $l(\cdot, y)$ is Lipschitz continuous. Then, a more concrete analysis is provided when k is a Gaussian kernel.

4.3.1 Error analysis for population risk minimizers

Before beginning the error analysis, some assumptions about the loss function and kernel function are imposed.

Assumption 4.1. *$l(\cdot, y)$ is convex and G-Lipscitz continuous, that is, there exists $G > 0$ such that for any $\zeta, \zeta' \in \mathbb{R}$ and $y \in \mathcal{Y}$,*

$$|l(\zeta, y) - l(\zeta', y)| \leq G|\zeta - \zeta'|.$$

This assumption implies G-Lipschitzness of $\mathcal{L}$ with respect to the $L^2(d\rho_\mathcal{X})$ norm, because

$$\begin{aligned} |\mathcal{L}(g) - \mathcal{L}(h)| &\leq G \int |g(x) - h(x)| d\rho_\mathcal{X}(x) \\ &\leq G \|g - h\|_{L^2(d\rho_\mathcal{X})} \end{aligned}$$

for any $g, h \in L^2(d\rho_\mathcal{X})$. For several practically used losses, such as logistic loss or hinge loss, this assumption is satisfied with $G = 1$.

To control continuity and boundedness of the induced kernel, the following assumptions are required:

Assumption 4.2. *The function φ is continuous and there exists $R > 0$ such that $|\varphi(x, \omega)| \leq R$ for any $x \in \mathcal{X}, \omega \in \Omega$.*

If k is Gaussian and φ is its random Fourier features, it is satisfied with $R = 1$. This assumption implies $\sup_{x,y \in \mathcal{X}} k(x, y) \leq R^2, \sup_{x,y \in \mathcal{X}} k_M(x, y) \leq R^2$ and it leads to an important relationship $R\|\cdot\|_\mathcal{H} \geq \|\cdot\|_{L^\infty(\mathcal{X})}, R\|\cdot\|_{\mathcal{H}_M} \geq \|\cdot\|_{L^\infty(\mathcal{X})}$.

For the two given kernels k and k_M, $k + k_M$ is also a positive definite kernel, and its RKHS includes $\mathcal{H}$ and $\mathcal{H}_M$. The last assumption imposes a specific norm relationship in its combined RKHS of $\mathcal{H}$ and $\mathcal{H}_M$.

Assumption 4.3. *Let $\mathcal{H}_M^+$ be RKHS with the kernel function $k + k_M$. Then there exists $0 \leq p < 1$, and a constant $C(\delta) > 0$ depends on $0 < \delta \leq 1$ that satisfies, for any $f \in \mathcal{H}_M^+$,*

$$\|f\|_{L^\infty(d\rho_\mathcal{X})} \leq C(\delta)\|f\|_{\mathcal{H}_M^+}^p \|f\|_{L^2(d\rho_\mathcal{X})}^{1-p}$$

with probability at least $1 - \delta$.

For a fixed kernel function, the Assumption 4.3 is a commonly used condition in an analysis of kernel methods [59, 35]. It is satisfied, for example, that the eigenfunctions of the kernel are uniformly bounded and the eigenvalues $\{\mu_i\}_i$ decay at the rate $i^{-1/p}$ [35]. In Theorem 4.2, specific p and $C(\delta)$ that satisfy the condition for the case of a Gaussian kernel and its random Fourier features approximation are derived.

Here, we introduce our primary result, which bounds the distance between g_λ and $g_{M,\lambda}$ in terms of $L^\infty(d\rho_\mathcal{X})$-norm. The complete statement, including proof and all constants, are found in Appendix E.

Theorem 4.1 (Simplified.)**.** *Under Assumption 4.1-4.3, with probability at least $1 - 2\delta$ with respect to the sampling of features, the following inequality holds:*

$$\|g_\lambda - g_{M,\lambda}\|_{L^\infty(d\rho_\mathcal{X})} \lesssim \left(\frac{R^4 \log \frac{R}{\delta}}{M}\right)^{\min\{(1-p)/4, 1/8\}} \frac{C(\delta) R G^{3/4} \|g_*\|_\mathcal{H}}{\lambda^{3/4}}.$$

The resulting error rate is $O(M^{-\min\{(1-p)/4, 1/8\}})$. It can be easily shown that a consistent error rate of $O(M^{-1/8})$ is seen for $L^2(d\rho_\mathcal{X})$-norm without Assumption 4.3.

Comparison to previous results. In [17, 61], the distance between empirical risk minimizers of SVM (i.e., l is hinge loss) were studied in terms of the error induced by Gram matrices. Considering K and K_M to be Gram matrices of kernel k and k_M, respectively, they showed that $\|g_\lambda - g_{M,\lambda}\|_{L^\infty(d\rho_\mathcal{X})} \lesssim O(\|K - K_M\|_{\mathrm{op}}^{1/4})$, where $\|\cdot\|_{\mathrm{op}}$ is an operator norm, defined in Chapter 1. Because the Gram matrix can be considered as the integral operator on the empirical measure, we can apply Lemma A.1 and obtain $\|K - K_M\|_{\mathrm{op}} \lesssim O(M^{-1/2})$, so the resulting rate is $O(M^{-1/8})$. This coincides with our result, because when $\rho_\mathcal{X}$ is an empirical measure, Assumption 3 holds with $p = 0$. From this perspective, our result is an extension of these previous results, because we treat the more general Lipschitz loss function l and general measure $\rho_\mathcal{X}$ including empirical measure.
In [50, 16], the case of squared loss was studied. In particular, in Lemma 8 of [50], the L^2 distance between g_λ and $g_{M,\lambda}$ is shown as $O(M^{-1/2})$ (without decreasing λ). While this is a better rate than ours, our theory covers a wider class of loss functions, and a similar phenomenon is observed in the case of empirical risk minimizers for the squared loss and hinge loss [17].
In [9], approximations of functions in $\mathcal{H}$ by functions in $\mathcal{H}_M$ were considered, but this result cannot be applied here because $g_{M,\lambda}$ is not the function closest to g_λ in $\mathcal{H}_M$. Finally, we note that our result cannot be obtained from the approximation analysis of Lipschitz loss functions [48, 33], where the rate was shown to be $O(M^{-1/2})$ under several assumptions, because the closeness of the loss values does not imply that of the hypothesis.

4.3.2 Further analysis for Gaussian kernels

The following theorem shows that if k is a Gaussian kernel and k_M is its random Fourier features approximation, then the norm condition in Assumption 4.3 is satisfied for any $0 < p < 1$.

Theorem 4.2. *Assume* $\mathrm{supp}(\rho_{\mathcal{X}}) \subset \mathbb{R}^d$ *is a bounded set and* $\rho_{\mathcal{X}}$ *has a density with respect to Lebesgue measure, which is uniformly bounded away from 0 and* ∞ *on* $\mathrm{supp}(\rho_{\mathcal{X}})$*. Let* k *be a Gaussian kernel and* $\mathcal{H}$ *be its RKHS; then, for any* $m \geq d/2$*, there exists a constant* $C_{m,d} > 0$ *such that*

$$\|f\|_{L^\infty(d\rho_{\mathcal{X}})} \leq C_{m,d} \|f\|_{\mathcal{H}}^{d/2m} \|f\|_{L^2(d\rho_{\mathcal{X}})}^{1-d/2m} \tag{4.7}$$

for any $f \in \mathcal{H}$*. Also, for any* $M \geq 1$*, let* k_M *be a random Fourier features approximation of* k *with* M *features and* $\mathcal{H}_M^+$ *be a RKHS of* $k + k_M$*. Then, with probability at least* $1 - \delta$ *with respect to a sampling of features,*

$$\|f\|_{L^\infty(d\rho_{\mathcal{X}})} \leq C_{m,d} \left(1 + \frac{1}{\delta}\right)^{d/4m} \|f\|_{\mathcal{H}_M^+}^{d/2m} \|f\|_{L^2(d\rho_{\mathcal{X}})}^{1-d/2m} \tag{4.8}$$

for any $f \in \mathcal{H}_M^+$*.*

We note that the norm relation of the Gaussian RKHS (4.7) is a known result in [59] and our analysis extends this to the combined RKHS $\mathcal{H}_M^+$. The proof is based on the following fact:

Let us denote $\mathrm{supp}(\rho_{\mathcal{X}})$ by $\mathcal{X}'$. First, from [59] we have

$$\left[L^2(\mathcal{X}'), W^m(\mathcal{X}')\right]_{d/2m,1} = B_{2,1}^{d/2}(\mathcal{X}')$$

and there exists a constant $C_1 > 0$ such that

$$\|f\|_{[L^2(\mathcal{X}'),W^m(\mathcal{X}')]_{d/2m,1}} \leq C_1 \|f\|_{W^m(\mathcal{X}')}^{d/2m} \|f\|_{L^2(\mathcal{X}')}^{1-d/2m},$$

where $W^m(\mathcal{X}')$ and $B_{2,1}^{d/2}(\mathcal{X}')$ denote Sobolev and Besov space, respectively, and $[E, F]_{\theta,r}$ denotes real interpolation of Banach spaces E and F (see [58]). Also, by Sobolev's embedding theorem for Besov space, $B_{2,1}^{d/2}(\mathcal{X}')$ can be continuously embedded in $L^\infty(\mathcal{X}')$. Finally, from the condition on $\rho_{\mathcal{X}}$, there exists a constant $C_2 > 0$ such that

$$\begin{aligned} \|f\|_{L^\infty(d\rho_{\mathcal{X}})} &= \|f\|_{L^\infty(\mathcal{X}')}, \\ \|f\|_{L^2(d\rho_{\mathcal{X}})} &\geq C_2 \|f\|_{L^2(\mathcal{X}')}. \end{aligned}$$

Therefore, if it can be shown that RKHS $\mathcal{H}_M^+$ is continuously embedded in $W^m(\mathcal{X}')$, the norm relation (4.8) holds. The complete proof is found in Appendix E.

Remark. Although we consider k as Gaussian, the statement itself holds if the probability measure τ has finite every order moments (see proof in Appendix E). In particular, for a shift-invariant kernel k, if $\psi(x - y) = k(x, y)$ belongs to the Schwartz class (including the case of a Gaussian kernel), τ (Fourier transform of ψ) also belongs to it, indicating that every moment is finite from the property of the Schwartz class [65] and the statement of Theorem 4.2 holds.

Using this theorem, it can be shown that in the case of a Gaussian kernel and its random Fourier features approximation, Assumption 4.3 is satisfied with $p = 1/2$ and $C(\delta) = C_{d,d}(1 + 1/\delta)^{1/4}$, and the resulting rate in Theorem 4.1 is $O(M^{-1/8})$.

4.4 Main Results

In this section, we show that learning classification problems with SGD and random features achieve the exponential convergence of the expected classification error under

certain conditions. Before providing our results, several assumptions are imposed on the classification problems and loss function.

The first is the smoothness of the loss function.

Assumption 4.4. *$l(\cdot, y)$ is differentiable and L-Lipschitz smooth. That is, for any $\zeta, \zeta' \in \mathbb{R}$ and $y \in \mathcal{Y}$,*

$$|\partial_\zeta l(\zeta, y) - \partial_\zeta l(\zeta', y)| \leq L|\zeta - \zeta'|.$$

This yields LR^2-smoothness of l with respect to $\|\cdot\|_{\mathcal{H}_M}$ norm from the same arguments as Section 2.3. The second is the margin condition on the conditional label probability, which is introduced in Section 2.1.

Assumption 4.5. *The strong low-noise condition holds:*

$$\exists \delta \in (0, 1/2)\,, \quad |\rho(Y = 1|x) - 1/2| > \delta \quad (\rho_{\mathcal{X}}\text{-}a.s.)$$

The third is the condition on the *link function* h_* [10, 66], which connects the hypothesis space and the probability measure:

$$h_*(\mu) = \arg\min_{\alpha \in \mathbb{R}} \{\mu\phi(\alpha) + (1-\mu)\phi(-\alpha)\}\,.$$

Its corresponding value is denoted by l_*:

$$l_*(\mu) = \min_{\alpha \in \mathbb{R}} \{\mu\phi(\alpha) + (1-\mu)\phi(-\alpha)\}\,.$$

It is known that l_* is a concave function [66]. Although $h_*(\mu)$ may not be uniquely determined nor well-defined in general, the following assumption ensures these properties.

Assumption 4.6. *$\rho(1|X)$ takes values in $(0, 1)$, $\rho_{\mathcal{X}}$-almost surely; ϕ is differentiable and h_* is well-defined, differentiable, monotonically increasing, and invertible over $(0, 1)$. Moreover, it follows that*

$$\operatorname{sgn}(\mu - 1/2) = \operatorname{sgn}(h_*(\mu)).$$

For logistic loss, $h_*(\mu) = \log(\mu/(1-\mu))$, and the above condition is satisfied. Next, following [66], we introduce Bregman divergence for concave function l_* to ensure the uniqueness of Bayes rule g_*:

$$d_{l_*}(\eta_1, \eta_2) = -l_*(\eta_2) + l_*(\eta_1) + l_*'(\eta_1)(\eta_2 - \eta_1).$$

Assumption 4.7. *Bregman divergence d_{l_*} derived by l_* is positive, that is, $d_{l_*}(\eta_1, \eta_2) = 0$ if and only if $\eta_1 = \eta_2$. For the expected risk $\mathcal{L}$, a unique Bayes rule g_* (up to zero measure sets) exists in $\mathcal{H}$.*

For logistic loss, it is known that d_{l_*} coincides with Kullbuck-Leibler divergence, and thus, the positivity of the divergence holds. If ϕ is differentiable and h_* is differentiable and invertible, the excess risk can be expressed using d_{l_*} [66]:

$$\mathcal{L}(g) - \mathcal{L}(g_*) = \mathbb{E}_X[d_{l_*}(h_*^{-1}(g(X)), \rho(1|X))].$$

So, combining Assumptions 4.6 and 4.7 implies that Bayes rule g_* is equal to $h_*(\rho(1|X))$, $\rho_{\mathcal{X}}$-almost surely and contained in the original RKHS $\mathcal{H}$.

Finally, we introduce the following notation:

$$m(\delta) = \max\{h_*(0.5 + \delta), |h_*(0.5 - \delta)|\}.$$

Using this notation, Assumption 4.5 can be reduced to the Bayes rule condition, that is, $|g_*(X)| \geq m(\delta)$, $\rho_{\mathcal{X}}$-almost surely. For logistic loss, we have $m(\delta) = \log((1+2\delta)/(1-2\delta))$.

Remark. As in Section 3, we can also impose Lipschitz continuity on h_*^{-1} instead of specifying $m(\delta)$. However, the use of Lipschitz continuity would yield looser bounds since h_*^{-1} is generally almost flat when an input is far away from 0 and a deviations on such a large value does not affect the corresponding probability value.

Under these assumptions and notations, the exponential convergence of the expected classification error is shown.

Theorem 4.3. *Suppose Assumptions 4.1–4.7 hold. There exists a sufficiently small $\lambda > 0$ such that the following statement holds:*
Taking the number of random features M that satisfies

$$M \gtrsim \left(\frac{R^4 C^4(\delta') G^3 \|g_*\|_{\mathcal{H}}^4}{\lambda^3 m^4(\delta)} \right)^{\max\left\{\frac{1}{1-p}, 2\right\}} R^4 \log \frac{R}{\delta'}.$$

Consider Algorithm 4.1 with $\eta_t = \frac{2}{\lambda(\gamma+t)}$ and $\alpha_t = \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)}$ where γ is a positive value such that $\|g_1\|_{\mathcal{H}_M} \leq (2\eta_1 + 1/\lambda)GR$ and $\eta_1 \leq \min\{1/LR^2, 1/2\lambda\}$. Then, with probability $1 - 2\delta'$, for sufficiently large T such that

$$\max\left\{ \frac{36G^2R^2}{\lambda^2(2\gamma+T)}, \frac{\gamma(\gamma-1)\|g_1 - g_{M,\lambda}\|_{\mathcal{H}_M}^2}{(2\gamma+T)(T+1)} \right\} \leq \frac{m^2(\delta)}{64R^2},$$

we have the following inequality for any $t \geq T$:

$$\mathbb{E}\left[\mathcal{R}(\overline{g}_{t+1}) - \mathcal{R}(\mathbb{E}[Y|x])\right] \leq 2\exp\left(-\frac{\lambda^2(2\gamma+t)m^2(\delta)}{2^{12} \cdot 9G^2R^4} \right).$$

The complete statement and proof are given in Appendix E. We note that although a certain number of features are required to achieve the exponential convergence, the resulting rate does not depend on M. In contrast to this, when one considers the convergence rate of the loss function, its rate depends on M in general [50, 16, 48, 33]. From this fact, we can show that random features can save computational cost in a relatively small classification error regime. A detailed discussion is presented below.

As a corollary, we show a simplified result when learning with random Fourier features approximation of a Gaussian kernel and logistic loss, which can be obtained by setting $m(\delta) = \log((1+2\delta)/(1-2\delta))$, $R = G = 1$ and $L = 1/4$ in Theorem 4.3 and applying Theorem 4.2.

Corollary 4.1. *Assume $\mathrm{supp}(\rho_{\mathcal{X}}) \subset \mathbb{R}^d$ is a bounded set and $\rho_{\mathcal{X}}$ has a density with respect to Lebesgue measure, which is uniformly bounded away from 0 and ∞ on $\mathrm{supp}(\rho_{\mathcal{X}})$. Let k be a Gaussian kernel and l be logistic loss. Under Assumption 4.5-4.7, there exists a sufficiently small $\lambda > 0$ such that the following statement holds:*
Taking a number of random features M that satisfies

$$M \gtrsim \left(\frac{\left(1 + \frac{1}{\delta'}\right) \|g_*\|_{\mathcal{H}}^4}{\lambda^3 \log^4 \frac{1+2\delta}{1-2\delta}} \right)^2 \log \frac{1}{\delta'}.$$

Consider Algorithm 4.1 with $\eta_t = \frac{2}{\lambda(\gamma+t)}$ and $\alpha_t = \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)}$ where γ is a positive value such that $\|g_1\|_{\mathcal{H}_M} \leq (2\eta_1 + 1/\lambda)$ and $\eta_1 \leq \min\{4, 1/2\lambda\}$. Then, with probability $1 - 2\delta'$, for a sufficiently large T such that

$$\max\left\{ \frac{36}{\lambda^2(2\gamma+T)}, \frac{\gamma(\gamma-1)\|g_1 - g_{M,\lambda}\|_{\mathcal{H}_M}^2}{(2\gamma+T)(T+1)} \right\} \leq \frac{\log^2 \frac{1+2\delta}{1-2\delta}}{64},$$

we have the following inequality for any $t \geq T$:

$$\mathbb{E}\left[\mathcal{R}(\overline{g}_{t+1}) - \mathcal{R}(\mathbb{E}[Y|x])\right] \leq 2\exp\left(-\frac{\lambda^2(2\gamma+t)}{2^{12}\cdot 9}\log^2\frac{1+2\delta}{1-2\delta}\right).$$

Computational viewpoint. As shown in Theorem 4.3, once a sufficient number of features are sampled, the convergence rate of the excess classification error does not depend on the number of features M. This is unexpected because when considering the convergence of the loss function, the approximation error induced by random features usually remains [50, 33, 48]. Thus, to obtain the best convergence rate, we have to sample more M as the sample size n increases.

From this fact, it can be shown that to achieve a relatively small classification error, learning with random features is indeed more computationally efficient than learning with a full kernel method without approximation. As shown in Section 4.2.4, if one runs SGD in Algorithm 4.1 with more than M iterations, both the time and space computational costs of a full kernel method exceed those of random features. In particular, if one can achieve a classification error ϵ such that

$$\epsilon \lesssim \exp\left(-\log^{2\max\{(1+p)/(1-p),3\}} m(\delta)\right),$$

then the required number of iterations n exceeds the required number of features M in Theorem 4.3, and the overall computational cost become larger in a full kernel method. Theoretical results which suggest the efficiency of random features in terms of generalization error have only been derived in the regression setting [50, 16]; this is the first time the superiority of random features has been demonstrated in the classification setting. Moreover, this result shows that an arbitrary small classification error is achievable as long as there is a sufficient number of random features unlike the regression setting where a required number of random features depend on the target accuracy.

4.5 Experiments

In this section, the behavior of the SGD with random features studied on synthetic datasets is described. We considered logistic loss as a loss function, a Gaussian kernel as an original kernel function, and its random Fourier features as an approximation method. Two-dimensional synthetic datasets were used, as shown in Figure 4.1. The dataset support is composed of four parts: $[-1.0,-0.1]\times[-1.0,-0.1], [-1.0,-0.1]\times[0.1,1.0], [0.1,1.0]\times[-1,-0.1], [0.1,1.0]\times[0.1,1.0]$. For two of them, the conditional probability is $\rho(1|X) = 0.8$, and for the other two, $\rho(1|X) = 0.2$. This distribution satisfies the strong low-noise condition with $\delta = 0.3$. For hyper-parameters, we set $\gamma = 500$ and $\lambda = 0.001$. The averaged stochastic descent was run 100 times with 12,000 iterations and the classification error and loss function were calculated on 100,000 test samples. The average of each run is reported with standard deviations.

First, the learning curves of the expected classification error and the expected risk are drawn when the number of features $M = 1000$, as shown in Figure 4.2. Our theoretical result suggests that with sufficient features, the classification error converges exponentially fast, whereas the loss function converges sub-linearly. We can indeed observe a much faster decrease in the classification error (left) than in the loss function (right).

Next, we show the learning curves of the expected classification error when the number of features are varied as $M = 100, 200, 500, 1000$ in Figure 4.3. We can see that the exact convergence of the classification error is not attained with relatively few features such as $M = 100$, which also coincides with our results.

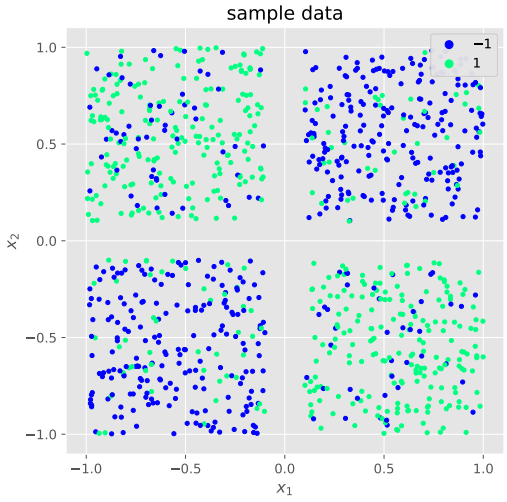
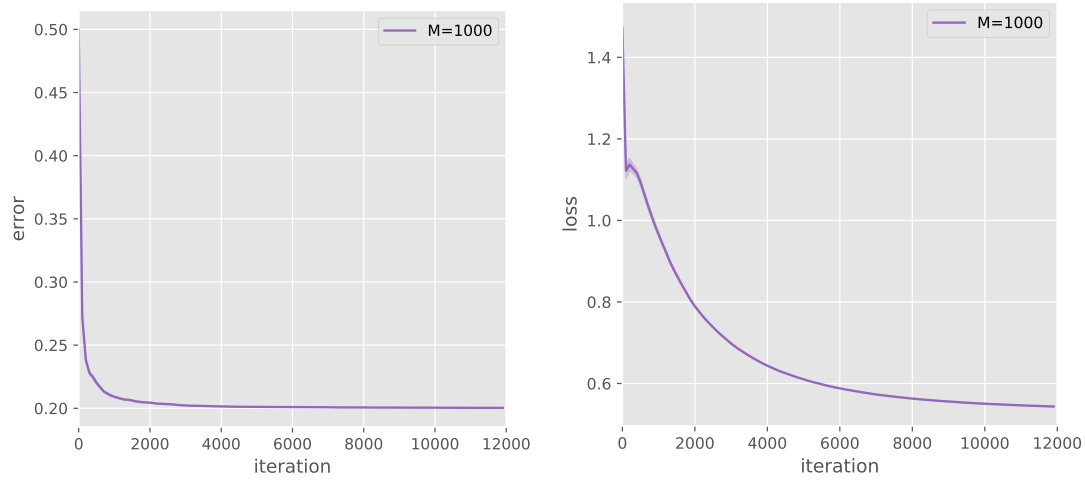


Fig. 4.1: Subsample of data used in the experiment.



(a) Classification errors (b) Loss functions

Fig. 4.2: Learning curves of the expected classification error (left) and the expected risk (right) by averaged SGD with 1000 features.

Finally, the convergence of the classification error is compared in terms of computational cost between the random features model with $M = 500, 1000$ and the full kernel model without approximation. In Figure 4.4, the learning curves are drawn with respect to the number of parameter updates; the full kernel model requires increasing numbers of updates in later iterations, whereas the random features model requires a constant number of updates. It can be observed that both random features models require fewer parameter updates to achieve the same classification error than the full kernel model for a relatively small classification error. This implies that random features approximation is indeed computationally efficient under a strong low-noise condition.

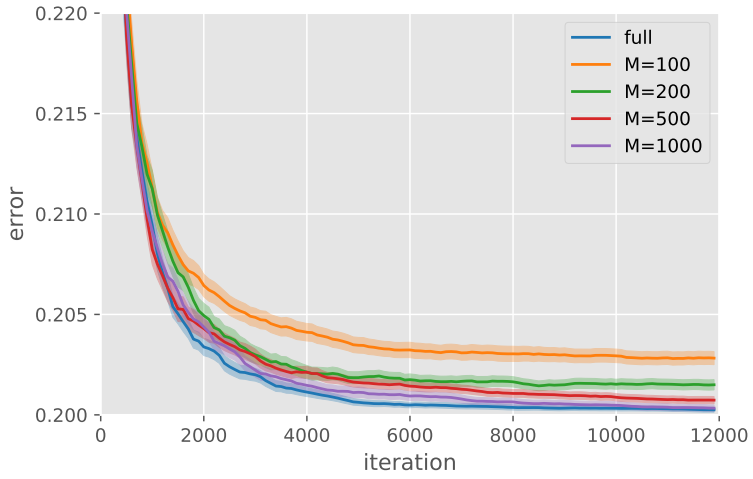


Fig. 4.3: Comparison of learning curves of the expected classification error with varying numbers of features.

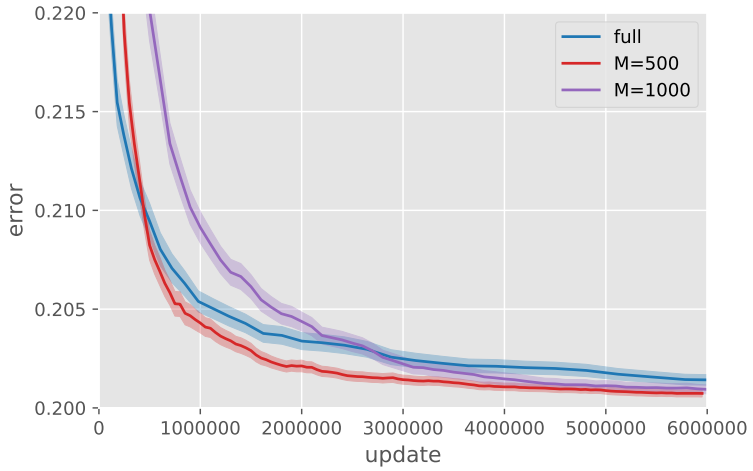


Fig. 4.4: Comparison of learning curves with respect to number of parameter updates.

4.6 Conclusion

This study shows that learning with SGD and random features could achieve exponential convergence of the classification error under a strong low-noise condition. Unlike when considering the convergence of a loss function, the resulting convergence rate of the classification error is independent of the number of features, indicating that an arbitrary small classification error is achievable as long as there is a sufficient number of random features. Our results suggest, for the first time, that random features is theoretically computationally efficient even for classification problems under certain settings. Our theoretical analysis has been verified by numerical experiments.

One possible future direction is to extend our analysis to general low-noise conditions to derive faster rates than $O(1/\sqrt{n})$, as in [43] in the case of the squared loss. It could also be interesting to explore the convergence speed of more sophisticated variants of SGD, such as stochastic accelerated methods and stochastic variance reduced methods [52, 30, 20, 2].

Chapter 5

Conclusion

5.1 Concluding Remarks

In this thesis, we have analyzed the validity of two approximation methods in binary classification problems on RKHS under the low noise condition. The first is SGD, which is an approximation from optimization perspectives. In Chapter 3, we showed that SGD can achieve a faster rate than $O(1/n)$ under the general low noise condition. It can be seen as interpolation of recent studies [43, 42], that show the exponential convergence under the strong low noise condition and conventional results [10, 66] that reveals the connection between the convergence rate of loss functions and that of classification errors under the general setting. In Chapter 4, we considered applying random features, which is the second approximation from the modeling perspectives. We showed that the required number of features to get an exact convergence is independent of sample size under the strong low noise condition, thus it means there is indeed a computational benefit to apply random features approximation. We believe these results give a better explanation of the practical success of these widely used approximation methods.

5.2 Future Perspective

Currently, we are aware of at least three future extensions of these results. First, it is important to consider different hypothesis spaces. Throughout this thesis, we consider kernel methods as a learning model, whose hypothesis space is RKHS. Although these models are general as linear estimators, it is somewhat inadequate when considering the great success of nonlinear estimators, in particular, deep neural networks (DNN). Recent studies have analyzed two-layer neural networks (i.e., those with one hidden layer) with over-parametrization where the number of neurons exceeds the sample size, and revealed the connection to kernel models [28, 23, 5]. Such an interpretation is called *neural tangent kernel* (NTK) and convergence properties of gradient descent on such models have been extensively studied for both regression and classification settings. Then one natural direction is to extend our analysis on classification problems under the low noise condition to such models optimized by SGD. These extensions may contribute to an understanding of the superior performance of SGD on DNN.

Secondly, it is worth applying more sophisticated variants of SGD to our analysis. Specifically, it can improve the stability of SGD with the use of variance reduction technique [30, 20, 41]. In particular, [26] considers the streaming (online) version of the variance reduction technique and analyze its generalization properties. These analyses may yield a better convergence of the expected classification error under the low noise condition. Other refinements, such as acceleration [2, 37] also have the possibility to improve the convergence.

Lastly, investigating whether it is possible or not to derive generalization error on classification problems optimized by SGD without regularization terms is also of great significance. Some recent studies [22, 43] shows that the optimal rates of least squares regression on RKHS is achievable without regularization terms by using step sizes instead of regularization parameter to balance generalization and optimization errors. Although our analysis heavily relies on the strong convexity of the objective, this extension, if possible, may give a better explanation of the superiority of SGD.

Acknowledgement

I am deeply grateful to my supervisor, associate professor Taiji Suzuki, whose comments and advice have been invaluable throughout the course of my study. His attitude towards research has always been inspiring to me. I would also like to offer my special thanks to assistant professor Atsushi Nitanda for all of his help and encouragement. Without having discussions with him, this thesis would not have materialized. I appreciate the feedback offered by professor Kenji Yamanishi and associate professor Takaaki Ohnishi throughout my master course. The members of Yamanishi, Suzuki and Ohnishi laboratory also gave me insightful comments and suggestions in days at the laboratory. These were very helpful to improve the quality of my research. Finally, I would like to express my gratitude to my family and friends for their moral support and warm encouragement.

Bibliography

[1] Alekh Agarwal, Martin J Wainwright, Peter L Bartlett, and Pradeep K Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pages 1–9, 2009.

[2] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

[3] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[4] Sanjeev Arora, László Babai, Jacques Stern, and Z Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences*, 54(2):317–331, 1997.

[5] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332, 2019.

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

[7] Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International Conference on Machine Learning*, pages 253–262, 2017.

[8] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[9] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

[10] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[11] Colin Bennett and Robert C Sharpley. *Interpolation of Operators*, volume 129. Academic Press, 1988.

[12] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM, 1992.

[13] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pages 161–168, 2008.

[14] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

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

[16] Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with SGD and random features. In *Advances in Neural Information Processing Systems*, pages 10192–10203, 2018.

[17] Corinna Cortes, Mehryar Mohri, and Ameet Talwalkar. On the impact of kernel approximation on learning accuracy. In *International Conference on Artificial Intel-*

ligence and Statistics, pages 113–120, 2010.

[18] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

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

[20] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[21] Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *International Conference on Knowledge Discovery and Data Mining*, pages 551–556. ACM, 2004.

[22] Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

[23] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

[24] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[25] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337–407, 2000.

[26] Roy Frostig, Rong Ge, Sham M Kakade, and Aaron Sidford. Competing with the empirical risk minimizer in a single pass. In *Conference on Learning Theory*, pages 728–763, 2015.

[27] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

[28] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

[29] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798, 2019.

[30] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[31] Vladimir Koltchinskii and Olexandra Beznosova. Exponential convergence rates in classification. In *International Conference on Computational Learning Theory*, pages 295–307. Springer, 2005.

[32] Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

[33] Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random Fourier features. In *International Conference on Machine Learning*, pages 3905–3914, 2019.

[34] Junhong Lin and Lorenzo Rosasco. Optimal learning for multi-pass stochastic gradient methods. In *Advances in Neural Information Processing Systems*, pages 4556–4564, 2016.

[35] Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, 2010.

[36] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing*

Systems, pages 451–459, 2011.

[37] Tomoya Murata and Taiji Suzuki. Doubly accelerated stochastic variance reduced dual averaging method for regularized empirical risk minimization. In *Advances in Neural Information Processing Systems*, pages 608–617, 2017.

[38] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[39] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley & Sons, 1983.

[40] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Springer Publishing Company, 2014.

[41] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621, 2017.

[42] Atsushi Nitanda and Taiji Suzuki. Stochastic gradient descent with exponential convergence rates of expected classification errors. In *International Conference on Artificial Intelligence and Statistics*, pages 1417–1426, 2019.

[43] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Exponential convergence of testing error for stochastic gradient methods. In *Conference on Learning Theory*, pages 250–296, 2018.

[44] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pages 8114–8124, 2018.

[45] Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

[46] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[47] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

[48] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems*, pages 1313–1320, 2009.

[49] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[50] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3215–3225, 2017.

[51] David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

[52] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

[53] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pages 583–588. Springer, 1997.

[54] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.

[55] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(10):2635–2670, 2010.

[56] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan

Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[57] Bharath Sriperumbudur and Zoltán Szabó. Optimal rates for random Fourier features. In *Advances in Neural Information Processing Systems*, pages 1144–1152, 2015.

[58] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

[59] Ingo Steinwart, Don R Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Conference on Learning Theory*, pages 79–93, 2009.

[60] Ingo Steinwart and Clint Scovel. Mercer' s theorem on general domains: on the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

[61] Dougal J Sutherland and Jeff Schneider. On the error of random Fourier features. In *Conference on Uncertainty in Artificial Intelligence*, pages 862–871, 2015.

[62] Zoltán Szabó and Bharath Sriperumbudur. On kernel derivative approximation with random Fourier features. In *International Conference on Artificial Intelligence and Statistics*, pages 827–836, 2019.

[63] Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[64] Christopher KI Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, pages 682–688, 2001.

[65] Kösaku Yoshida. *Functional Analysis*. Springer-Verlag Berlin Heidelberg, 1995.

[66] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

A

Reproducing Kernel Hilbert Space (RKHS)

In this section, we introduce some properties of RKHS. For more thoroughly introduction, we refer to Chapter 4 in [58] and [22].

Definition of RKHS

Firstly, we introduce a reproducing kernel Hilbert space (RKHS), which is widely adapted in a non-parametric statistical learning. Here we define a positive definite kernel.

Definition A.1 (Positive definite kernel)**.** *$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive definite kernel if for any finite set $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, the $n \times n$ matrix whose (i, j) entry is $k(x_i, x_j)$ is positive semi-definite.*

Example. The following kernels are positive definite:

- Polynomial kernel on $\mathcal{X} = \mathbb{R}^d$ with a integer degree parameter $j \geq 0$:

$$k(x_1, x_2) = (1 + x_1 \cdot x_2)^j$$

- Binomial kernel on $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\|_1 < 1\}$ with a integer degree parameter $j > 0$:

$$k(x_1, x_2) = (1 - x_1 \cdot x_2)^{-j}$$

- Gaussian kernel on $\mathcal{X} = \mathbb{R}^d$ with a band width parameter $\sigma > 0$:

$$k(x_1, x_2) = \exp(-\|x_1 - x_2\|_2^2/\sigma^2)$$

- Laplacian kernel on $\mathcal{X} = \mathbb{R}^d$ with a band width parameter $\sigma > 0$:

$$k(x_1, x_2) = \exp(-\|x_1 - x_2\|_2/\sigma)$$

For every positive definite kernel k, a Hilbert space spanned by k is defined:

Definition A.2 (Reproducing kernel Hilbert space (RKHS))**.** *A reproducing kernel Hilbert space $\mathcal{H}$ associates with a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ satisfying the following properties:*

1. *$k(\cdot, x) \in \mathcal{H}$ for any $x \in \mathcal{X}$.*
2. *$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$ for any $x \in \mathcal{X}$ and $f \in \mathcal{H}$.*

It is known that every positive definite kernel has a unique RKHS and every RKHS has a unique reproducing kernel (Theorem 4.20, 4.21 in [58]). Here we introduce some facts and theorems which is related to the thesis. The following proposition characterizes a RKHS of sum of two kernel functions.

Proposition A.1 (RKHS of sum of kernels)**.** *Let $k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be positive definite kernels and $\mathcal{H}_1$ and $\mathcal{H}_2$ be their associated RKHSs, respectively. Then, $k_1 + k_2$ is also a positive definite kernel and its RKHS $\mathcal{H}^+$ is given by*

$$\mathcal{H}^+ = \{f_1 + f_2 \mid f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2\}$$
$$\|f\|_{\mathcal{H}^+} = \inf\{\|f_1\|_{\mathcal{H}_1} + \|f_2\|_{\mathcal{H}_2} \mid f = f_1 + f_2, f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2\}.$$

The next proposition shows an important characteristic of shift-invariant kernels such as Gaussian and Laplacian kernel introduced above, and it is utilized in kernel approximations by random Fourier features [47] in Chapter 4.

Proposition A.2 (Bochner's theorem [65])**.** *Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a kernel and assume that there exists $\psi : \mathbb{R}^d \to \mathbb{R}$ such that $k(x_1, x_2) = \psi(x_1 - x_2)$ for all $x_1, x_2 \in \mathbb{R}^d$. Then k is positive definite if and only if there exists a unique finite Borel measure μ on $\mathbb{R}^d$ such that*

$$\psi(t) = \int_{\mathbb{R}^d} e^{it\cdot\lambda} d\mu(\lambda).$$

Characterization of RKHS by integral operators

It is well known [58] that RKHS can be characterized using an integral operator associated with a kernel function. In this section, we introduce these connection used in the analysis. We note that although it is common for analyses of kernel methods to assume $\mathcal{X}$ is compact, $\rho_{\mathcal{X}}$ has the full support and k is continuous because under such assumptions we utilize Mercer's theorem to characterize RKHS [19, 3], such an assumption may not be adopted under classification problems we consider in the thesis. In particular, when the low noise condition is considered, $\rho_{\mathcal{X}}$ may not have full support. So we explain some basic properties of reproducing kernel Hilbert space (RKHS) under more general settings based on [22, 60].
First we impose a boundedness on k as same as previous sections.

Assumption A.1. *There exists $R > 0$ such that $k(x, x) \leq R^2$ for any $x \in \mathcal{X}$.*

For given kernel function k and its RKHS $\mathcal{H}$, we define a covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$ as follows:

$$\langle f, \Sigma g\rangle_{\mathcal{H}} = \langle f, g\rangle_{L^2(d\rho_{\mathcal{X}})}, \quad \forall f, g \in \mathcal{H}.$$

It is well-defined through Riesz' representation theorem [65]. Using reproducing property, we have

$$\Sigma = \mathbb{E}_{X\sim\rho_{\mathcal{X}}}[k(\cdot, X) \otimes_{\mathcal{H}} k(\cdot, X)],$$
$$(\Sigma f)(z) = \mathbb{E}_{X\sim\rho_{\mathcal{X}}}[f(X)k(X, z)], \quad \forall f \in \mathcal{H}. \tag{A.1}$$

where expectation is defined via a Bochner integration, which is an extension of Lebesgue integration to general Banach spaces [65]. From the representation (A.1), we can extend the covariance operater to $f \in L^2(d\rho_{\mathcal{X}})$. We denote this by $T : L^2(d\rho_{\mathcal{X}}) \to L^2(d\rho_{\mathcal{X}})$ as follows:

$$(Tf)(z) = \mathbb{E}_{X\sim\rho_{\mathcal{X}}}[f(X)k(X, z)], \quad \forall f \in L^2(d\rho_{\mathcal{X}}).$$

$\mathrm{Im}(T) \subset L^2(d\rho_{\mathcal{X}})$ is verified since $k(\cdot, x)$ is uniformly bounded by Assumption A.1. In addition, we denote a set of square integral function itself (without dividing by almost sure equivalence class) by $\mathcal{L}^2(d\rho_{\mathcal{X}})$. Then we define the extended covariance operator $\mathcal{T} : L^2(d\rho_{\mathcal{X}}) \to \mathcal{L}^2(d\rho_{\mathcal{X}})$ as follows:

$$(\mathcal{T}f)(z) = \mathbb{E}_{X \sim \rho_{\mathcal{X}}}[f(X)k(X, z)], \quad \forall f \in L^2(d\rho_{\mathcal{X}}).$$

Here we present some properties of these covariance operators $\Sigma, T, \mathcal{T}$ from Appendix I in [22].

Proposition A.3.

1. *Σ is self-adjoint, continuous operator and* $\mathrm{Ker}(\Sigma) = \{f \in \mathcal{H} \mid \|f\|_{L^2(d\rho_{\mathcal{X}})} = 0\}$.
2. *T is continuous, self-adjoint, positive semi-definite operator.*
3. *$\mathcal{T}^{1/2} : \mathrm{Ker}(T)^{\perp} \to \mathrm{Ker}(\Sigma)^{\perp}$ is well-defined and an isometry. In particular, for any $f \in \mathrm{Ker}(\Sigma)^{\perp} \subset \mathcal{H}$, there exists $g \in \mathrm{Ker}(T)^{\perp} \subset L^2(d\rho_{\mathcal{X}})$ such that $\|f\|_{\mathcal{H}} = \|g\|_{L^2(d\rho_{\mathcal{X}})}$.*

Thus, a square root of an integral operator $\mathcal{T}$ represents an isometric correspondence between $L^2(d\rho_{\mathcal{X}})$ and $\mathcal{H}$ up to zero measure sets. Next, we consider the case where a kernel k is represented as a feature expansion form as introduced in Chapter 4.

Assumption A.2. *There exists $\varphi : \mathcal{X} \times \Omega \to \mathbb{C}$ and a probability measure τ on Ω such that the following inequality holds:*

$$k(x, y) = \int_{\Omega} \varphi(x, \omega)\overline{\varphi(y, \omega)}d\tau(\omega). \tag{A.2}$$

Then, we can write T using feature expansion (A.2) as

$$T = \mathbb{E}_{\omega \sim \tau}[\varphi(\cdot, \omega) \otimes_{L^2(d\rho_{\mathcal{X}})} \varphi(\cdot, \omega)], \tag{A.3}$$

since it holds that

$$\begin{aligned}(Tf)(z) &= \mathbb{E}_{X \sim \rho_{\mathcal{X}}}[f(X)\mathbb{E}_{\omega \sim \tau}[\overline{\varphi(X, \omega)}\varphi(z, \omega)]] \\ &= \mathbb{E}_{\omega \sim \tau}[\langle f, \varphi(\cdot, \omega)\rangle_{L^2(d\rho_{\mathcal{X}})}\varphi(z, \omega)].\end{aligned}$$

Moreover, define a finite dimensional approximation of (A.2) as

$$k_M(x, y) = \sum_{i=1}^{M} \varphi(x, \omega_i)\overline{\varphi(y, \omega_i)}, \quad \omega_i \sim \tau.$$

and denote its RKHS by $\mathcal{H}_M$ and extended covariate operators by $T_M : L^2(d\rho_{\mathcal{X}}) \to L^2(d\rho_{\mathcal{X}})$ and $\mathcal{T}_M : L^2(d\rho_{\mathcal{X}}) \to \mathcal{L}^2(d\rho_{\mathcal{X}})$. As with (A.3), we have

$$\begin{aligned}T_M &= \frac{1}{M}\sum_{i=1}^{M} \varphi(\cdot, \omega_i) \otimes_{L^2(d\rho_{\mathcal{X}})} \varphi(\cdot, \omega_i), \\ \mathbb{E}[T_M] &= T.\end{aligned}$$

In the following section, we show some inequalities about the approximation error of $\mathcal{H}$ by $\mathcal{H}_M$, which is used to derive Theorem 4.1.

Finite dimensional approximation of RKHS

The following inequality shows that the difference between the square root of two self-adjoint positive semi-definite operators is bounded by the square root of the difference of them.

Proposition A.4. *Let V be a separable Hilbert space. For any compact, positive semi-definite, self-adjoint operators $S, \widetilde{S} : V \to V$, the following inequality holds:*

$$\|S^{1/2} - \widetilde{S}^{1/2}\|_{\mathrm{op}} \leq \|S - \widetilde{S}\|_{\mathrm{op}}^{1/2} \tag{A.4}$$

Proof. Since $S^{1/2} - \widetilde{S}^{1/2}$ is also a compact and self-adjoint operator, it allows eigendecomposition of itself. Then let $\lambda_{\max}$ be the eigenvalue with largest absolute value and v be the corresponding normalized eigenfunction of $S^{1/2} - \widetilde{S}^{1/2}$, i.e.,

$$(S^{1/2} - \widetilde{S}^{1/2})v = \lambda_{\max} v.$$

Since (A.4) obviously holds if $S = \widetilde{S}$, we can assume that $\lambda_{\max} > 0$ without loss of generality. Because $S^{1/2}$ is positive semi-definite, we have

$$\begin{aligned}
\langle v, Sv\rangle_V &= \|S^{1/2}v\|_V^2 \\
&= \|\widetilde{S}^{1/2}v + \lambda_{\max} v\|_V^2 \\
&= \langle v, \widetilde{S}v\rangle_V + \lambda_{\max}^2 + 2\lambda_{\max}\langle v, S^{1/2}v\rangle_V \\
&\geq \langle v, \widetilde{S}v\rangle_V + \lambda_{\max}^2.
\end{aligned}$$

Thus we have

$$\begin{aligned}
\|S - \widetilde{S}\|_{\mathrm{op}} &\geq \langle v, (S - \widetilde{S})v\rangle_V \\
&\geq \lambda_{\max}^2 = \|S^{1/2} - \widetilde{S}^{1/2}\|_{\mathrm{op}}^2,
\end{aligned}$$

which completes the proof. □

The following inequality is a generalization of the Bernstein inequality to random operators on separable Hilbert space and used in Lemma A.1 to derive the concentration of integral operators.

Proposition A.5 (Proposition 3 in [50])**.** *Let V be a separable Hilbert space and let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed self-adjoint random operators on V. Assume that $\mathbb{E}X_i = 0$ and there exists $B > 0$ such that $\|X_i\|_{\mathrm{op}} \leq B$ almost surely for any $1 \leq i \leq n$. Let S be the positive operator such that $\mathbb{E}X_i^2 \leq S$. Then for any $\delta \in (0, 1]$, the following inequality holds with probability at least $1 - \delta$:*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \right\|_{\mathrm{op}} \leq \frac{2B\beta}{3n} + \sqrt{\frac{2\|S\|_{\mathrm{op}}\beta}{n}},$$

where $\beta = \log \frac{2\mathrm{tr}S}{\|S\|_{\mathrm{op}}\delta}$.

The next lemma provides a probabilistic bounds about the difference of the two covariate operators T and T_M.

Lemma A.1. *For any $\delta \in [0,1)$, the following inequality holds with probability at least $1-\delta$:*

$$\|T - T_M\|_{\mathrm{op}} \leq R^2 \left(\frac{2\beta}{3M} + \sqrt{\frac{2\beta}{M}} \right)$$

where $\beta = \log \frac{2R^2}{\|T\|_{\mathrm{op}}\delta}$.

Proof. Let $X_i = T - \varphi(\cdot,\omega_i) \otimes_{L^2(d\rho_{\mathcal{X}})} \varphi(\cdot,\omega_i)$. Then $T - T_M = \frac{1}{M}\sum_{i=1}^{M} X_i$. Also, we have

$$\begin{aligned}
\mathbb{E}X_i &= 0, \\
X_i &\preceq T \preceq R^2 I, \\
X_i &\succeq -\varphi(\cdot,\omega_i) \otimes_{L^2(d\rho_{\mathcal{X}})} \varphi(\cdot,\omega_i) \succeq -R^2 I, \\
||X_i||_{\mathrm{op}} &\leq R^2, \text{as a result of two previous inequalities,} \\
\mathbb{E}X_i^2 &= \mathbb{E}\left[\varphi(\cdot,\omega_i) \otimes_{L^2(d\rho_{\mathcal{X}})} \varphi(\cdot,\omega_i)\right]^2 - T^2 \\
&\preceq \mathbb{E}\left[\varphi(\cdot,\omega_i) \otimes_{L^2(d\rho_{\mathcal{X}})} \varphi(\cdot,\omega_i)\right]^2 \\
&\preceq \mathbb{E}\left[\langle \varphi(\cdot,\omega_i), \varphi(\cdot,\omega_i)\rangle_{L^2(d\rho_{\mathcal{X}})} \varphi(\cdot,\omega_i) \otimes_{L^2(d\rho_{\mathcal{X}})} \varphi(\cdot,\omega_i)\right] \\
&\preceq R^2 T, \\
\mathrm{tr}T &= \int_{\mathcal{X}} k(x,x) d\rho_{\mathcal{X}}(x) \leq R^2.
\end{aligned}$$

Let $B = R^2$ and $S = R^2 T$ in Proposition A.5, we have

$$\begin{aligned}
\|T - T_M\|_{\mathrm{op}} &= \left\| \frac{1}{M}\sum_{i=1}^{M} X_i \right\|_{\mathrm{op}} \\
&\leq \frac{2R^2\beta}{3M} + \sqrt{\frac{2R^2\|T\|_{\mathrm{op}}\beta}{M}} \\
&\leq R^2 \left(\frac{2\beta}{3M} + \sqrt{\frac{2\beta}{M}} \right),
\end{aligned}$$

which completes the proof. □

Let $\mathcal{H}$ and $\mathcal{H}_M$ be RKHSs associate with kernels k and k_M, respectively. Using Proposition A.3 and Lemma A.1, we have the following proposition, which is essential in the proof of Theorem 4.1.

Lemma A.2. *For any $\delta \in (0,1]$ and $\xi > 0$, if*

$$M \geq \max\left\{ \frac{8}{3}\left(\frac{R}{\xi}\right)^2, 32\left(\frac{R}{\xi}\right)^4 \right\} \log \frac{2R^2}{\|T\|_{\mathrm{op}}\delta}$$

holds, the following statement holds with probability at least $1-\delta$:
For any $g \in \mathcal{H}$, there exists $\widetilde{g} \in \mathcal{H}_M$ that satisfies

- $\|g - \widetilde{g}\|_{L^2(d\rho_{\mathcal{X}})} \leq \xi \|g\|_{\mathcal{H}}$
- $\|g\|_{\mathcal{H}} \geq \|\widetilde{g}\|_{\mathcal{H}_M}$.

Also, for any $\widetilde{g} \in \mathcal{H}_M$, there exists $g \in \mathcal{H}$ that satisfies

- $\|g - \widetilde{g}\|_{L^2(d\rho_{\mathcal{X}})} \leq \xi \|\widetilde{g}\|_{\mathcal{H}_M}$

- $\|g\|_{\mathcal{H}} \leq \|\widetilde{g}\|_{\mathcal{H}_M}$.

Proof. We show the first part of the statement. The latter half can be shown in the same manner.
For $g \in \mathcal{H}$, set $\widetilde{g} = \mathcal{T}_M^{1/2} \mathcal{P}_{\mathrm{Ker}(T_M)^\perp} \mathcal{T}^{-1/2} \mathcal{P}_{\mathrm{Ker}(\Sigma)^\perp} g \in \mathcal{H}_M$. Then we have

$$\begin{aligned}
\|\widetilde{g}\|_{\mathcal{H}_M} &= \|\mathcal{P}_{\mathrm{Ker}(T_M)^\perp} \mathcal{T}^{-1/2} \mathcal{P}_{\mathrm{Ker}(\Sigma)^\perp} g\|_{L^2(d\rho_{\mathcal{X}})} \\
&\leq \|\mathcal{T}^{-1/2} \mathcal{P}_{\mathrm{Ker}(\Sigma)^\perp} g\|_{L^2(d\rho_{\mathcal{X}})} \\
&= \|\mathcal{P}_{\mathrm{Ker}(\Sigma)^\perp} g\|_{\mathcal{H}} \\
&\leq \|g\|_{\mathcal{H}}.
\end{aligned}$$

Moreover, by Proposition A.4 and Lemma A.1, with probability at least $1 - \delta$, we have

$$\begin{aligned}
\|g - \widetilde{g}\|_{L^2(d\rho_{\mathcal{X}})} &= \|\mathcal{P}_{\mathrm{Ker}(\Sigma)^\perp} g - \widetilde{g}\|_{L^2(d\rho_{\mathcal{X}})} \quad (\because \text{Proposition A.3.1}) \\
&= \|\mathcal{T}^{1/2} h - \mathcal{T}_M^{1/2} \mathcal{P}_{\mathrm{Ker}(T_M)^\perp} h\|_{L^2(d\rho_{\mathcal{X}})} \\
&= \|T^{1/2} h - T_M^{1/2} h\|_{L^2(d\rho_{\mathcal{X}})} \\
&\leq \|T^{1/2} - T_M^{1/2}\|_{\mathrm{op}} \|h\|_{L^2(d\rho_{\mathcal{X}})} \\
&\leq \|T - T_M\|_{\mathrm{op}}^{1/2} \|g\|_{\mathcal{H}} \\
&\leq \left(R^2 \left(\frac{2\beta}{3M} + \sqrt{\frac{2\beta}{M}} \right) \right)^{1/2} \|g\|_{\mathcal{H}} \\
&\leq R \left(\left(\frac{2\beta}{3M} \right)^{1/2} + \left(\frac{2\beta}{M} \right)^{1/4} \right) \|g\|_{\mathcal{H}}
\end{aligned}$$

where $h = \mathcal{T}^{-1/2} \mathcal{P}_{\mathrm{Ker}(\Sigma)^\perp} g \in L^2(d\rho_{\mathcal{X}})$ and $\beta = \log \frac{2R^2}{\|T\|_{\mathrm{op}} \delta}$.
Solving the equation $\max \left\{ \left(\frac{2\beta}{3M} \right)^{1/2}, \left(\frac{2\beta}{M} \right)^{1/4} \right\} \leq \frac{\xi}{2R}$, we get a desired result. □

B

Interpolation Space

In this section, we introduce real interpolation of two Banach spaces.

Definition B.1 (Interpolation space [11, 59])**.** *For given Banach spaces E and F such that $F \subset E$ and $\mathrm{id} : F \to E$ is continuous, we define the K-functional of $x \in E$ by*

$$K(x,t) \stackrel{\mathrm{def}}{=} \inf_{y \in F} \left(\|x-y\|_E + t\|y\|_F \right), \quad t > 0.$$

Then, the real interpolation space $[E,F]_{\theta,r}$, where $0 < \theta < 1$ and $1 \leq r \leq \infty$, is a Banach space that consists of those $x \in E$ with finite norm

$$\|x\|_{[E,F]_{\theta,r}} \stackrel{\mathrm{def}}{=} \begin{cases} \left(\int_0^\infty \left(t^{-\theta} K(x,t) \right)^r t^{-1} dt \right)^{\frac{1}{r}} & (r < \infty) \\ \sup_{t>0} t^{-\theta} K(x,t) & (r = \infty). \end{cases}$$

Moreover, the limiting cases are defined by

$$[E,F]_{0,\infty} \stackrel{\mathrm{def}}{=} E, \quad [E,F]_{1,\infty} \stackrel{\mathrm{def}}{=} F.$$

It is known that for all $0 < \theta < 1$ and $1 \leq r \leq r' \leq \infty$, $[E,F]_{\theta,r}$ is continuously embedded in $[E,F]_{\theta,r'}$, that is, there exists a constant $C > 0$ such that for any $f \in [E,F]_{\theta,r}$

$$\|f\|_{[E,F]_{\theta,r'}} \leq C \|f\|_{[E,F]_{\theta,r}}.$$

In analyses of kernel methods, the case where $E = L^2(d\rho_{\mathcal{X}})$ and $F = \mathcal{H}$ (RKHS) is usually considered to characterize how difficult the target function f is to learn with functions in $\mathcal{H}$ [59]. When $r = 1$, there is a useful relation about norms, which are used in the proof of Theorem 4.2.

Proposition B.1 (Proposition 2.10 in [11])**.** *For any $f \in [E,F]_{\theta,1}$, there exists a constant $C > 0$ such that*

$$\|f\|_{[E,F]_{\theta,1}} \leq C \|f\|_E^{1-\theta} \|f\|_F^{\theta}.$$

Setting $E = L^2(d\rho_{\mathcal{X}})$ and $F = \mathcal{H}$, the norm condition (Assumption 4.3)

$$\|f\|_{L^\infty(\mathcal{X})} \leq C \|f\|_{L^2(d\rho_{\mathcal{X}})}^{1-p} \|f\|_{\mathcal{H}}^{p}$$

is reduced to the condition that $[L^2(d\rho_{\mathcal{X}}), \mathcal{H}]_{p,1}$ is continuously embedded to $L^\infty(\mathcal{X})$.

C

Proofs in Chapter 2

In this section, we show the missing proofs in Chapter 2. They are essentially appeared in [42].

Proof of Proposition 2.3

Proposition 2.3 (Proposition C in [42])**.** *Suppose Assumption 2.1, 2.2 holds. Consider Algorithm 2.2 with* $\eta_t = \frac{2}{\lambda(\gamma+t)}$ *and* $\alpha_t = \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)}$ *and assume assume* $\|\overline{g}_1\|_{\mathcal{H}} \leq (2\gamma_1 + 1/\lambda)GR$ *and* $\eta_1 \leq \min\{1/L, 1/2\lambda\}$*. Then, it follows that*

$$\|\mathbb{E}[\overline{g}_{T+1}] - g_\lambda\|_{\mathcal{H}}^2 \leq \frac{2}{\lambda}\left(\frac{18G^2R^2}{\lambda(2\gamma+T)} + \frac{\lambda\gamma(\gamma-1)}{2(2\gamma+T)(T+1)}\|\overline{g}_1 - g_\lambda\|_{\mathcal{H}}^2\right).$$

Proof. First, we show that the norm of $G_\lambda(g_t, Z_t)$ can be uniformly bounded for all t in this setting. By Assumption 2.1 and 2.2, we have the following bound on the stochastic gradient of l in $\mathcal{H}$:

$$\|\partial_\zeta l(g(x), y)k(\cdot, x)\|_{\mathcal{H}} \leq GR.$$

Therefore, if $\|g_t\|_{\mathcal{H}} \geq \frac{1}{\lambda}GR$, then

$$\begin{aligned}\|g_{t+1}\|_{\mathcal{H}} &= \|g_t - \eta_t\partial_\zeta l(g(X_t), Y_t)k(\cdot, X_t) - \eta_t\lambda g_t\|_{\mathcal{H}} \\ &\leq (1-\eta_t\lambda)\|g_t\|_{\mathcal{H}} + \eta_t GR \\ &\leq \|g_t\|_{\mathcal{H}}.\end{aligned}$$

Otherwise, we have

$$\begin{aligned}\|g_{t+1}\|_{\mathcal{H}} &= \|g_t - \eta_t\partial_\zeta l(g(X_t), Y_t)k(\cdot, X_t) - \eta_t\lambda g_t\|_{\mathcal{H}} \\ &\leq \|g_t\|_{\mathcal{H}} + 2\eta_t GR \\ &\leq \|g_t\|_{\mathcal{H}} + 2\eta_1 GR\end{aligned}$$

From this, we can see that a generated sequence $\{g_t\}_{t=1,\ldots,T+1}$ is contained in a close ball centered at the origin with radius $(2\eta_1 + 1/\lambda)GR$ as long as an initial function $\overline{g}_1$ is contained in this ball. Thus, the norm of the stochastic gradient $G_\lambda(g_t, Z_t)$ is bounded as

$$\begin{aligned}\|G_\lambda(g_t, Z_t)\|_{\mathcal{H}} &\leq \|\partial_\zeta l(g_t(X_t), Y_t)k(\cdot, X_t)\|_{\mathcal{H}} + \lambda\|g_t\|_{\mathcal{H}} \\ &\leq GR + \lambda(2\eta_1 + 1/\lambda)GR \\ &\leq 3GR, \end{aligned} \tag{C.1}$$

where the last inequality follows from the condition $\eta_1 \leq 1/2\lambda$.

By (C.1) and the strong convexity of $\mathcal{L}_\lambda$, we have

$$\begin{aligned}\mathbb{E}\|g_{t+1} - g_\lambda\|_{\mathcal{H}}^2 &= \mathbb{E}\|g_t - g_\lambda\|_{\mathcal{H}}^2 - 2\eta_t \mathbb{E}[\langle g_t - g_\lambda, G_\lambda(g_t, Z_t)\rangle_{\mathcal{H}}] + \eta_t^2 \mathbb{E}\|G_\lambda(g_t, Z_t)\|_{\mathcal{H}}^2 \\ &\leq \mathbb{E}\|g_t - g_\lambda\|_{\mathcal{H}}^2 - 2\eta_t \mathbb{E}[\langle g_t - g_\lambda, \nabla\mathcal{L}_\lambda(g_t)\rangle_{\mathcal{H}}] + 9\eta_t^2 G^2 R^2 \\ &\leq \mathbb{E}\|g_t - g_\lambda\|_{\mathcal{H}}^2 - 2\eta_t \left(\mathbb{E}[\mathcal{L}_\lambda(g_t)] - \mathcal{L}_\lambda(g_\lambda) + \frac{\lambda}{2}\mathbb{E}\|g_t - g_\lambda\|_{\mathcal{H}}^2\right) + 9\eta_t^2 G^2 R^2.\end{aligned}$$

Thus we obtain

$$\begin{aligned}\mathbb{E}[\mathcal{L}_\lambda(g_t)] - \mathcal{L}_\lambda(g_\lambda) &\leq \frac{1 - \lambda\eta_t}{2\eta_t}\mathbb{E}\|g_t - g_\lambda\|_{\mathcal{H}}^2 - \frac{1}{2\eta_t}\mathbb{E}\|g_{t+1} - g_\lambda\|_{\mathcal{H}}^2 + \frac{9\eta_t G^2 R^2}{2} \\ &= \frac{\lambda(\gamma + t - 2)}{4}\mathbb{E}\|g_t - g_\lambda\|_{\mathcal{H}}^2 - \frac{\lambda(\gamma + t)}{4t}\mathbb{E}\|g_{t+1} - g_\lambda\|_{\mathcal{H}}^2 + \frac{9G^2R^2}{\lambda(\gamma + t)}.\end{aligned}$$

By multiplying $\gamma + t - 1$ and taking sum over $t \in \{1, \ldots, T+1\}$, we get

$$\begin{aligned}&\sum_{t=1}^{T+1}(\gamma + t - 1)(\mathbb{E}[\mathcal{L}_\lambda(g_t)] - \mathcal{L}_\lambda(g_\lambda)) \\ &\leq \frac{\lambda}{4}\sum_{t=1}^{T+1}\{(\gamma + t - 1)(\gamma + t - 2)\mathbb{E}\|g_t - g_\lambda\|_{\mathcal{H}}^2 - (\gamma + t)(\gamma + t - 1)\mathbb{E}\|g_{t+1} - g_\lambda\|_{\mathcal{H}}^2\} \\ &\quad + \frac{9G^2R^2(T+1)}{\lambda} \\ &\leq \frac{\lambda}{4}\gamma(\gamma - 1)\|g_1 - g_\lambda\|_{\mathcal{H}}^2 + \frac{9G^2R^2(T+1)}{\lambda}.\end{aligned}$$

Dividing by $(2\gamma + T)(T+1)/2$ and applying Jensen's inequality, we obtain

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}_\lambda(\overline{g}_{T+1})\right] - \mathcal{L}_\lambda(g_\lambda) &= \mathbb{E}\left[\mathcal{L}_\lambda\left(\sum_{t=1}^{T+1}\frac{2(\gamma + t - 1)}{(2\gamma + T)(T+1)}g_t\right)\right] - \mathcal{L}_\lambda(g_\lambda) \\ &\leq \sum_{t=1}^{T+1}\frac{2(\gamma + t - 1)}{(2\gamma + T)(T+1)}\mathbb{E}\left[\mathcal{L}_\lambda\left(g_t\right)\right] - \mathcal{L}_\lambda(g_\lambda) \\ &\leq \frac{\lambda\gamma(\gamma - 1)}{2(2\gamma + T)(T+1)}\|g_1 - g_\lambda\|_{\mathcal{H}}^2 + \frac{18G^2R^2}{\lambda(2\gamma + T)}.\end{aligned}$$

Using the strong convexity of $\mathcal{L}_\lambda$ and applying Jensen's inequality again, we get

$$\begin{aligned}\|\mathbb{E}[\overline{g}_{T+1}] - g_\lambda\|_{\mathcal{H}}^2 &\leq \frac{2}{\lambda}\left(\mathcal{L}_\lambda(\mathbb{E}[\overline{g}_{T+1}]) - \mathcal{L}_\lambda(g_\lambda)\right) \\ &\leq \frac{2}{\lambda}\left(\mathbb{E}\left[\mathcal{L}_\lambda(\overline{g}_{T+1})\right] - \mathcal{L}_\lambda(g_\lambda)\right)\end{aligned}$$

and obtain a desired bound. □

Proof of Proposition 2.4

Proposition 2.4 (Proposition 2 and D in [42])**.** *Suppose Assumption 2.1, 2.2 holds. Consider Algorithm 2.2 with* $\eta_t = \frac{2}{\lambda(\gamma+t)}$ *and* $\alpha_t = \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)}$ *and assume* $\|\overline{g}_1\|_{\mathcal{H}} \leq (2\gamma_1 + 1/\lambda)GR$ *and* $\eta_1 \leq \min\{1/L, 1/2\lambda\}$*. Then, it follows that*

$$\mathbb{P}\left[\left\|\overline{g}_{T+1} - \mathbb{E}[\overline{g}_{T+1}]\right\|_{\mathcal{H}} \geq \epsilon\right] \leq 2\exp\left(-\frac{\lambda^2(2\gamma+T)}{2^6 \cdot 3^2 G^2 R^2}\epsilon^2\right).$$

Proof. From (C.1) and the fact that $g_t = g_t^t$, we have

$$\|g_{t+1} - g_{t+1}^t\|_{\mathcal{H}} = \eta_t \|G_\lambda(g_t, Z_t) - G_\lambda(g_t, Z_t')\|_{\mathcal{H}} \leq 6\eta_t GR. \tag{C.2}$$

For simplicity, let $l(g, z)$ denote $l(g(x), y)$ for $z = (x, y)$ and $\partial_g l(g, z)$ denote the gradient of $l(g, z)$ with respect to $g \in \mathcal{H}$. Then Assumption 2.1 and Theorem 2.1.5 in [40] yields that for any $g, g' \in \mathcal{H}$,

$$\langle \partial_g l(g, z) - \partial_g l(g', z), g - g' \rangle_{\mathcal{H}} \geq \frac{1}{LR^2} \|\partial_g l(g, z) - \partial_g l(g', z)\|_{\mathcal{H}}^2.$$

Thus we have that for $s \geq t+1$,

$$\begin{aligned}
\|g_{s+1} - g_{s+1}^t\|_{\mathcal{H}}^2 &= (1-\eta_s\lambda)^2 \|g_s - g_s^t\|_{\mathcal{H}}^2 - 2\eta_s(1-\eta_s\lambda)\langle \partial_g l(g_s, Z_s) - \partial_g l(g_s^t, Z_s), g_s - g_s^t \rangle_{\mathcal{H}} \\
&\quad + \eta_s^2 \|\partial_g l(g_s, Z_s) - \partial_g l(g_s^t, Z_s)\|_{\mathcal{H}}^2 \\
&\leq (1-\eta_s\lambda)^2 \|g_s - g_s^t\|_{\mathcal{H}}^2 - \eta_s \left(\frac{1}{LR^2} - \eta_s\right) \|\partial_g l(g_s, Z_s) - \partial_g l(g_s^t, Z_s)\|_{\mathcal{H}}^2 \\
&\leq (1-\eta_s\lambda)^2 \|g_s - g_s^t\|_{\mathcal{H}}^2,
\end{aligned}$$

where the last inequality follows from the condition $\eta_s \leq \eta_1 \leq 1/LR^2$. By substituting $\eta_s = \frac{2}{\lambda(\gamma+s)}$ and (C.2), we have

$$\begin{aligned}
\|g_{s+1} - g_{s+1}^t\|_{\mathcal{H}} &\leq \prod_{r=t+1}^{s} \frac{\gamma+r-2}{\gamma+r} \|g_{t+1} - g_{t+1}^t\|_{\mathcal{H}} \\
&\leq \frac{12GR}{\lambda(\gamma+s)}.
\end{aligned}$$

Recall that $\overline{g}_{t+1} = (1-\theta_t)\overline{g}_t + \theta_t g_{t+1}$, where $\theta_t = \frac{2(\gamma+t)}{(t+1)(2\gamma+t)}$. Then we have

$$\begin{aligned}
\|\overline{g}_{T+1} - \overline{g}_{T+1}^t\|_{\mathcal{H}} &\leq (1-\theta_T)\|\overline{g}_T - \overline{g}_T^t\|_{\mathcal{H}} + \theta_t \|g_{T+1} - g_{T+1}^t\|_{\mathcal{H}}. \\
&\leq \sum_{s=t}^{T} \left\{ \prod_{r=s+1}^{T} (1-\theta_r) \right\} \theta_s \|g_{s+1} - g_{s+1}^t\|_{\mathcal{H}} \\
&\leq \sum_{s=t}^{T} \frac{(2\gamma+s)}{(T+1)(2\gamma+T)} \frac{12GR}{\lambda(\gamma+s)} \\
&\leq \sum_{s=t}^{T} \frac{24GR}{\lambda(T+1)(2\gamma+T)} \\
&\leq \frac{24GR(T-t+1)}{\lambda(T+1)(2\gamma+T)}
\end{aligned}$$

Finally, we obtain

$$\begin{aligned}\sum_{t=1}^{T+1} \|D_t\|_\infty^2 &\leq \frac{24^2 G^2 R^2}{\lambda^2 (T+1)^2 (2\gamma+T)^2} \sum_{t=1}^{T+1} (T-t+1)^2 \\ &\leq \frac{24^2 G^2 R^2}{\lambda^2 (T+1)^2 (2\gamma+T)^2} \frac{T(T+1)(2T+1)}{6} \\ &\leq \frac{24 \cdot 4 G^2 R^2 (2T+1)}{\lambda^2 (2\gamma+T)^2} \\ &\leq \frac{288 G^2 R^2}{\lambda^2 (2\gamma+T)}.\end{aligned}$$

By substituting $c_T^2 = \frac{288 G^2 R^2}{\lambda^2 (2\gamma+T)}$ in Lemma 2.1, we obtain a desired bound. □

D

Proofs in Chapter 3

Proof of Theorem 3.1

The following lemma shows that the low noise condition on $\rho(1|X)$ yields that on $g_*(X)$ with the use of the Lipschitz continuity of h_*^{-1}.

Lemma D.1. *Under Assumption 3.3 and 3.5, there exists $C_4 > 0$ such that the following inequality holds for all $\delta > 0$:*

$$\mathbb{P}\left[|g_*(X)| \leq \delta\right] \leq C_4 \delta^{\alpha}$$

Proof.

$$\begin{aligned}\mathbb{P}\left[|g_*(X)| \leq \delta\right] &\leq \mathbb{P}\left[|h_*^{-1}(g_*(X)) - h_*^{-1}(0)| \leq L'\delta\right] \\ &= \mathbb{P}\left[|\rho(Y=1|x) - 1/2| \leq L'\delta\right] \\ &\leq C_2 L'^{\alpha} \delta^{\alpha},\end{aligned}$$

and setting $C_4 = C_2 L'^{\alpha}$, we get a desired result. □

Using the above lemma, the fast convergence of expected classification errors is shown.

Theorem 3.1. *Suppose Assumption 3.1-3.6 holds. Consider Algorithm 4.1 with $\lambda = \left(288 C_2^{-2} G^2 R^4\right)^{\frac{1}{2+2\kappa}} T^{-\frac{1}{2+2\kappa}}$, $\eta_t = \frac{2}{\lambda(\gamma+t)}$ and $\alpha_t = \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)}$ where γ is a positive value such that $\|g_1\|_{\mathcal{H}} \leq (2\eta_1 + 1/\lambda)GR$ and $\eta_1 \leq \min\{1/LR^2, 1/2\lambda\}$. Then there exists constant $C > 0$ such that the following inequality holds:*

$$\mathbb{E}\left[\mathcal{R}(\overline{g}_{T+1})\right] - \mathcal{R}(g_*) \leq C T^{-\frac{(\alpha+1)\kappa}{2+2\kappa}}.$$

Proof. For simplicity, we denote the conditional label probability $\rho(1|X)$ by $\eta(X)$ and the training samples by $z_{1:T} = \{(x_1, y_1), \ldots, (x_T, y_T)\}$. For given $\delta > 0$, choose $\lambda > 0$ so that

$$\|g_\lambda - g_*\|_{L^\infty(d\rho_{\mathcal{X}})} \leq \delta. \tag{D.1}$$

Consider the sets $A_j \subset \mathcal{X}, j = 0, 1, \ldots$, defined as

$$\begin{aligned}A_0 &= \left\{x \in \mathcal{X} \mid \operatorname{sgn}(g_*(x)) g_\lambda(x) \leq \delta\right\}, \\ A_j &= \left\{x \in \mathcal{X} \;\middle|\; 2^{j-1}\delta < \operatorname{sgn}(g_*(x)) g_\lambda(x) \leq 2^j \delta\right\} \quad \text{for } j \geq 1.\end{aligned}$$

Then we have the following inequality:

$$\begin{aligned}\mathbb{P}\left[\operatorname{sgn}(g_*(X)) g_\lambda(X) \leq 2^j\delta\right] &\leq \mathbb{P}\left[|g_\lambda(X)| \leq 2^j\delta\right] \\ &\leq \mathbb{P}\left[|g_*(X)| \leq 2^{j+1}\delta\right] \\ &\leq 2^{\alpha(j+1)} C_4 \delta^{\alpha},\end{aligned} \tag{D.2}$$

where the first inequality follows from the fact that we have $|g_\lambda(X)| \leq \delta$ in the case of $\operatorname{sgn}(g_*(X))g_\lambda(X) < 0$ because of (D.1), and the last inequality follows from Assumption 3.5 and Lemma A.2. Also, we have

$$X \in A_j \Longrightarrow |g_*(X)| \leq 2^{j+1}\delta \Longrightarrow \left|\eta(X) - \frac{1}{2}\right| \leq 2^{j+1}L'\delta \tag{D.3}$$

from Assumption 3.3. Now let us fix $T > 0$ such that

$$\max\left\{\frac{36L^2R^2}{\lambda^2(2\gamma+T)}, \frac{\gamma(\gamma-1)\|g_1 - g_\lambda\|_{\mathcal{H}}^2}{(2\gamma+T)(T+1)}\right\} \leq \frac{\delta^2}{8R^2}, \tag{D.4}$$

then we obtain

$$\left\|\mathbb{E}[\overline{g}_{T+1}] - g_\lambda\right\|_{L^\infty(d\rho_{\mathcal{X}})} \leq R\|\mathbb{E}[\overline{g}_{T+1}] - g_\lambda\|_{\mathcal{H}} \leq \frac{\delta}{2}$$

from Proposition 2.3 and

$$\begin{aligned}
\mathbb{P}\left[\|\overline{g}_{T+1} - g_\lambda\|_{L^\infty(d\rho_{\mathcal{X}})} \geq 2^{j-1}\delta\right] &\leq \mathbb{P}\left[\|\overline{g}_{T+1} - g_\lambda\|_{\mathcal{H}} \geq 2^{j-1}\delta/R\right] \\
&\leq \mathbb{P}\left[\|\overline{g}_{T+1} - \mathbb{E}[\overline{g}_{T+1}]\|_{\mathcal{H}} \geq 2^{j-2}\delta/R\right] \\
&\leq 2\exp\left(-\frac{2^{2j-10}\lambda^2(2\gamma+T)}{3^2L^2R^4}\delta^2\right)
\end{aligned} \tag{D.5}$$

from Proposition 2.4. Moreover we have

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{R}(\overline{g}_{T+1}) - \mathcal{R}(g_*)\right] \\
&= \mathbb{E}_{X,z_{1:T}}\left[|2\eta(X)-1| \cdot \mathbb{1}_{\operatorname{sgn}(\overline{g}_{T+1}(X)) \neq \operatorname{sgn}(g_*(X))}\right] \\
&= \mathbb{E}_{X,z_{1:T}} \sum_{j=0}^{\infty}\left[|2\eta(X)-1| \cdot \mathbb{1}_{\operatorname{sgn}(\overline{g}_{T+1}(X)) \neq \operatorname{sgn}(g_*(X))} \cdot \mathbb{1}_{A_j}\right] \\
&\leq \mathbb{E}_{X,z_{1:T}}\left[|2\eta(X)-1| \cdot \mathbb{1}_{A_0}\right] + \mathbb{E}_{X,z_{1:T}} \sum_{j=1}^{\infty}\left[|2\eta(X)-1| \cdot \mathbb{1}_{\operatorname{sgn}(\overline{g}_{T+1}(X)) \neq \operatorname{sgn}(g_*(X))} \cdot \mathbb{1}_{A_j}\right] \\
&\leq 4L'\delta \cdot 2^\alpha C_4\delta^\alpha \\
&\quad + \sum_{j=1}^{\infty} 2^{j+2}L'\delta\mathbb{E}_X\left[\mathbb{E}_{z_{1:T}}\left[\mathbb{1}_{\operatorname{sgn}(\overline{g}_{T+1}(X)) \neq \operatorname{sgn}(g_*(X))} \cdot \mathbb{1}_{2^{j-1}\delta < \operatorname{sgn}(g_*(X))g_\lambda(X)}\right]\mathbb{1}_{\operatorname{sgn}(g_*(X))g_\lambda(X) \leq 2^j\delta}\right] \\
&\leq 2^{\alpha+2}L'C_4\delta^{\alpha+1} + \sum_{j=1}^{\infty} 2^{j+3}L'\delta\exp\left(-\frac{2^{2j-10}\lambda^2(2\gamma+T)}{3^2L^2R^4}\delta^2\right) \cdot 2^{\alpha(j+1)}C_4\delta^\alpha \quad (\because \text{(D.2), (D.3), (D.5)}) \\
&= 2^{\alpha+2}L'C_4\delta^{\alpha+1}\left(1 + \sum_{j=1}^{\infty} 2^{(\alpha+1)j+1}\exp\left(-\frac{2^{2j-10}\lambda^2(2\gamma+T)}{3^2L^2R^4}\delta^2\right)\right).
\end{aligned}$$

Now we set λ and δ as

$$\lambda = C_5T^{-\frac{1}{2+2\kappa}}, \qquad \delta = C_6T^{-\frac{\kappa}{2+2\kappa}},$$

where

$$C_5 = \left(288 C_2^{-2} L^2 R^4\right)^{\frac{1}{2+2\kappa}},$$
$$C_6 = \max\left\{\left(288^{\kappa} C_2^2 L^{2\kappa} R^{4\kappa}\right)^{\frac{1}{2+2\kappa}}, \frac{4\gamma(\gamma-1)R^2\left(\|g_1\|_{\mathcal{H}} + \|g_*\|_{\mathcal{H}}\right)^2}{2\gamma+1}\right\}.$$

Then we can check that it satisfies the conditions (D.1) and (D.4) using Assumption 3.6. Indeed for (D.1) we have

$$\begin{aligned}\|g_\lambda - g_*\|_{L^\infty(d\rho(\mathcal{X}))} &\le C_2\lambda^{\kappa} \quad (\because \text{Assumption 3.6})\\ &= C_2 C_5^{\kappa} T^{-\frac{\kappa}{2+2\kappa}}\\ &\le \delta,\end{aligned}$$

and for (D.4), we have

$$\begin{aligned}\frac{36L^2R^2}{\lambda^2(2\gamma+T)} &\le 36L^2R^2C_5^{-2}T^{-\frac{2\kappa}{2+2\kappa}}\\ &\le 8^{-1}C_6^2R^{-2}T^{-\frac{\kappa}{2+2\kappa}}\\ &= \frac{\delta^2}{8R^2}\end{aligned}$$

and

$$\begin{aligned}\frac{\gamma(\gamma-1)\|g_1-g_\lambda\|_{\mathcal{H}}^2}{(2\gamma+T)(T+1)} &\le \frac{\gamma(\gamma-1)(\|g_1\|_{\mathcal{H}}+\|g_*\|_{\mathcal{H}})^2}{(2\gamma+1)T}\\ &\le \frac{\gamma(\gamma-1)(\|g_1\|_{\mathcal{H}}+\|g_*\|_{\mathcal{H}})^2(C_6^{-1}\delta)^{\frac{2+2\kappa}{\kappa}}}{2\gamma+1}\\ &\le \frac{\delta^2}{8R^2}.\end{aligned}$$

By substituting them, we obtain

$$\mathbb{E}\left[\mathcal{R}(\overline{g}_{T+1}) - \mathcal{R}(g_*)\right] \le 2^{\alpha+2}L'C_4C_6^{\alpha+1}T^{-\frac{(\alpha+1)\kappa}{2+2\kappa}} \left(1 + \sum_{j=1}^{\infty} 2^{(\alpha+1)j+1}\exp\left(-\frac{2^{2j-10}C_5^2C_6^2(2\gamma+1)}{3^2L^2R^4}\right)\right)$$

for any $T \ge 1$. Since the last sum converges, we get a desired result. □

Proof of Theorem 3.2

Theorem 3.2. *Suppose Assumption 3.2, 3.7 and 3.8 holds. Then it holds that*

$$\|g_\lambda - g_*\|_{L^\infty(d\rho_{\mathcal{X}})} \le 2^p C_3 \|g_*\|_{\mathcal{H}} \left(\frac{\lambda}{\mu(R\|g_*\|_{\mathcal{H}})}\right)^{\frac{1-p}{2}}.$$

Thus Assumption 3.6 is satisfied with $\kappa = \frac{1-p}{2}$.

Proof. By definition of g_λ, we have

$$\mathcal{L}(g_*) + \frac{\lambda}{2}\|g_*\|_{\mathcal{H}}^2 \geq \mathcal{L}(g_\lambda) + \frac{\lambda}{2}\|g_\lambda\|_{\mathcal{H}}^2, \tag{D.6}$$

$$\|g_*\|_{\mathcal{H}} \geq \|g_\lambda\|_{\mathcal{H}}. \tag{D.7}$$

In addition, it holds that

$$\begin{aligned} g_*(x) &\leq R\|g_*\|_{\mathcal{H}}, \\ g_\lambda(x) &\leq R\|g_\lambda\|_{\mathcal{H}} \leq R\|g_*\|_{\mathcal{H}} \end{aligned} \tag{D.8}$$

for all $x \in \mathcal{X}$. Furthermore, since g_* attains infimum of $\mathcal{L}$ among all measurable functions, we have

$$\int_{\mathcal{Y}} \partial_\zeta l(g_*(\cdot), y) d\rho(y|\cdot) \equiv 0, \tag{D.9}$$

where ∂_ζ denotes a partial derivative of l with respect to the first variable.
Then we obtain

$$\begin{aligned} \|g_\lambda - g_*\|_{L^2(d\rho_{\mathcal{X}})}^2 &= \int_{\mathcal{X}} |g_\lambda(x) - g_*(x)|^2 \, d\rho_{\mathcal{X}}(x) \\ &\leq \int_{\mathcal{X}\times\mathcal{Y}} \frac{2}{\mu(R\|g_*\|_{\mathcal{H}})} \{l(g_\lambda(x), y) - l(g_*(x), y) \\ &\qquad - \partial_\zeta l(g_*(x), y)(g_\lambda(x) - g_*(x))\} d\rho(x, y) \quad (\because \text{(D.8) and Assumption 3.7}) \\ &= \int_{\mathcal{X}\times\mathcal{Y}} \frac{2}{\mu(R\|g_*\|_{\mathcal{H}})} \{l(g_\lambda(x), y) - l(g_*(x), y)\} \, d\rho(x, y) \quad (\because \text{(D.9)}) \\ &= \frac{2}{\mu(R\|g_*\|_{\mathcal{H}})} (\mathcal{L}(g_\lambda) - \mathcal{L}(g_*)) \\ &\leq \frac{\lambda}{\mu(R\|g_*\|_{\mathcal{H}})} \left(\|g_*\|_{\mathcal{H}}^2 - \|g_\lambda\|_{\mathcal{H}}^2\right) \quad (\because \text{(D.6)}) \\ &\leq \frac{\lambda}{\mu(R\|g_*\|_{\mathcal{H}})} \|g_*\|_{\mathcal{H}}^2 \quad (\because \text{(D.7)}). \end{aligned}$$

Finally, applying Assumption 3.8 we have

$$\begin{aligned} \|g_\lambda - g_*\|_{L^\infty(d\rho_{\mathcal{X}})} &\leq C_3 \|g_\lambda - g_*\|_{\mathcal{H}}^p \|g_\lambda - g_*\|_{L^2(d\rho_{\mathcal{X}})}^{1-p} \\ &\leq 2^p C_3 \left(\frac{\lambda}{\mu(R\|g_*\|_{\mathcal{H}})}\right)^{\frac{1-p}{2}} \|g_*\|_{\mathcal{H}}. \end{aligned}$$

□

E

Proofs in Chapter 4

Proof of Theorem 4.1

In this section, we give the complete statement and proof of Theorem 4.1.

Theorem 4.1. *Define $\xi > 0$ such that*

$$\xi = \min\left\{\left(\frac{\epsilon}{2^{p+1}C(\delta)\|g_*\|_{\mathcal{H}}}\right)^{1/1-p}, \frac{\lambda\epsilon^2}{2^4\cdot 3R^2L\|g_*\|_{\mathcal{H}}}, \left(\frac{\lambda^3\epsilon^4}{2^7\cdot 3^2R^4L^2\mathcal{L}(g_*)}\right)^{1/2}, \left(\frac{\lambda^3\epsilon^4}{2^7\cdot 3^2R^4L^3\|g_*\|_{\mathcal{H}}}\right)^{1/3}\right\}.$$

Then a number of random features M which satisfies

$$M \geq \max\left\{\frac{8}{3}\left(\frac{R}{\xi}\right)^2, 32\left(\frac{R}{\xi}\right)^4\right\}\log\frac{2R^2}{\|T\|_{\mathrm{op}}\delta}$$

is enough to guarantee, with probability at least $1-2\delta$, that

$$\|g_\lambda - g_{M,\lambda}\|_{L^\infty(d\rho_{\mathcal{X}})} \leq \epsilon.$$

Proof. By Lemma A.2, for given $\xi > 0$, if we have a number of feature M such that

$$M \geq \max\left\{\frac{8}{3}\left(\frac{R}{\xi}\right)^2, 32\left(\frac{R}{\xi}\right)^4\right\}\log\frac{2R^2}{\|T\|_{\mathrm{op}}\delta},$$

we can take $\widetilde{g}_\lambda \in \mathcal{H}_M, \widetilde{g}_{M,\lambda} \in \mathcal{H}$ which satisfy the following conditions:

$$\|g_\lambda\|_{\mathcal{H}} \geq \|\widetilde{g}_\lambda\|_{\mathcal{H}_M} \tag{E.1}$$
$$\|g_{M,\lambda}\|_{\mathcal{H}_M} \geq \|\widetilde{g}_{M,\lambda}\|_{\mathcal{H}} \tag{E.2}$$
$$\|\widetilde{g}_{M,\lambda} - g_{M,\lambda}\|_{L^2(d\rho_{\mathcal{X}})} \leq \xi\|g_{M,\lambda}\|_{\mathcal{H}_M} \tag{E.3}$$
$$\|\widetilde{g}_\lambda - g_\lambda\|_{L^2(d\rho_{\mathcal{X}})} \leq \xi\|g_\lambda\|_{\mathcal{H}} \tag{E.4}$$

By λ-strong convexity with respect to RKHS norm, we have

$$\mathcal{L}(g_\lambda) + \frac{\lambda}{2}\|g_\lambda\|_{\mathcal{H}}^2 + \frac{\lambda}{2}\|g_\lambda - \widetilde{g}_{M,\lambda}\|_{\mathcal{H}}^2 \leq \mathcal{L}(\widetilde{g}_{M,\lambda}) + \frac{\lambda}{2}\|\widetilde{g}_{M,\lambda}\|_{\mathcal{H}}^2 \tag{E.5}$$
$$\mathcal{L}(g_{M,\lambda}) + \frac{\lambda}{2}\|g_{M,\lambda}\|_{\mathcal{H}_M}^2 + \frac{\lambda}{2}\|g_{M,\lambda} - \widetilde{g}_\lambda\|_{\mathcal{H}_M}^2 \leq \mathcal{L}(\widetilde{g}_\lambda) + \frac{\lambda}{2}\|\widetilde{g}_\lambda\|_{\mathcal{H}_M}^2. \tag{E.6}$$

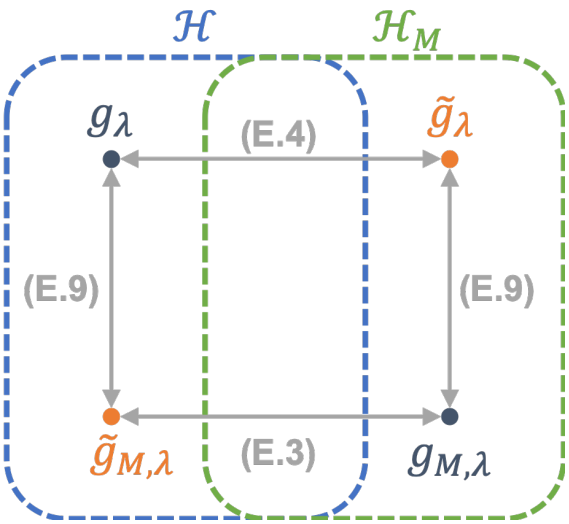


Fig. E.1: A schematic drawing of the proof of Theorem 4.1.

In addition, by L-Lipschitzness of $\mathcal{L}$ with respect to $L^2(d\rho_{\mathcal{X}})$ norm in Assumption 4.1 and (E.3)(E.4), we have

$$\begin{aligned}\mathcal{L}(\widetilde{g}_{M,\lambda}) &\le \mathcal{L}(g_{M,\lambda}) + L\|\widetilde{g}_{M,\lambda} - g_{M,\lambda}\|_{L^2(d\rho_{\mathcal{X}})} \\ &\le \mathcal{L}(g_{M,\lambda}) + L\xi\|g_{M,\lambda}\|_{\mathcal{H}_M} \qquad \text{(E.7)} \\ \mathcal{L}(\widetilde{g}_\lambda) &\le \mathcal{L}(g_\lambda) + L\|\widetilde{g}_\lambda - g_\lambda\|_{L^2(d\rho_{\mathcal{X}})} \\ &\le \mathcal{L}(g_\lambda) + L\xi\|g_\lambda\|_{\mathcal{H}} \qquad \text{(E.8)}\end{aligned}$$

By inequalities (E.5)(E.6)(E.7)(E.8) and (E.1)(E.2), we have

$$\begin{aligned}&\mathcal{L}(g_\lambda) + \frac{\lambda}{2}\|g_\lambda\|_{\mathcal{H}}^2 + \frac{\lambda}{2}\left(\|g_\lambda - \widetilde{g}_{M,\lambda}\|_{\mathcal{H}}^2 + \|g_{M,\lambda} - \widetilde{g}_\lambda\|_{\mathcal{H}_M}^2\right) \\ &\le \mathcal{L}(\widetilde{g}_{M,\lambda}) + \frac{\lambda}{2}\|\widetilde{g}_{M,\lambda}\|_{\mathcal{H}}^2 + \frac{\lambda}{2}\|g_{M,\lambda} - \widetilde{g}_\lambda\|_{\mathcal{H}_M}^2 \\ &\le \mathcal{L}(g_{M,\lambda}) + L\xi\|g_{M,\lambda}\|_{\mathcal{H}_M} + \frac{\lambda}{2}\|g_{M,\lambda}\|_{\mathcal{H}_M}^2 + \frac{\lambda}{2}\|g_{M,\lambda} - \widetilde{g}_\lambda\|_{\mathcal{H}_M}^2 \\ &\le \mathcal{L}(\widetilde{g}_\lambda) + \frac{\lambda}{2}\|\widetilde{g}_\lambda\|_{\mathcal{H}_M}^2 + L\xi\|g_{M,\lambda}\|_{\mathcal{H}_M} \\ &\le \mathcal{L}(g_\lambda) + \frac{\lambda}{2}\|g_\lambda\|_{\mathcal{H}}^2 + L\xi\left(\|g_\lambda\|_{\mathcal{H}} + \|g_{M,\lambda}\|_{\mathcal{H}_M}\right).\end{aligned}$$

Thus we have

$$\|g_\lambda - \widetilde{g}_{M,\lambda}\|_{\mathcal{H}}^2 + \|g_{M,\lambda} - \widetilde{g}_\lambda\|_{\mathcal{H}_M}^2 \le \frac{2L\xi}{\lambda}\left(\|g_\lambda\|_{\mathcal{H}} + \|g_{M,\lambda}\|_{\mathcal{H}_M}\right). \qquad \text{(E.9)}$$

In addition, by (E.6) and (E.8), we have

$$\begin{aligned}\frac{\lambda}{2}\|g_{M,\lambda}\|_{\mathcal{H}_M}^2 &\le \mathcal{L}(\widetilde{g}_\lambda) + \frac{\lambda}{2}\|\widetilde{g}_\lambda\|_{\mathcal{H}_M}^2 \\ &\le \mathcal{L}(g_\lambda) + L\xi\|g_\lambda\|_{\mathcal{H}} + \frac{\lambda}{2}\|g_\lambda\|_{\mathcal{H}}^2. \qquad \text{(E.10)}\end{aligned}$$

Combining (E.9) and (E.10), we obtain

$$\begin{aligned}\|g_\lambda - \widetilde{g}_{M,\lambda}\|_{\mathcal{H}}^2 + \|g_{M,\lambda} - \widetilde{g}_\lambda\|_{\mathcal{H}_M}^2 &\leq \frac{2L\xi}{\lambda}\left(\|g_\lambda\|_{\mathcal{H}} + \left(\frac{2}{\lambda}\mathcal{L}(g_\lambda) + \frac{2L\xi}{\lambda}\|g_\lambda\|_{\mathcal{H}} + \|g_\lambda\|_{\mathcal{H}}^2\right)^{1/2}\right) \\ &\leq \frac{2L\xi}{\lambda}\left(\|g_*\|_{\mathcal{H}} + \left(\frac{2}{\lambda}\mathcal{L}(g_*) + \frac{2L\xi}{\lambda}\|g_*\|_{\mathcal{H}} + \|g_*\|_{\mathcal{H}}^2\right)^{1/2}\right) \\ &\leq \frac{2L\xi}{\lambda}\left(2\|g_*\|_{\mathcal{H}} + \left(\frac{2}{\lambda}\mathcal{L}(g_*)\right)^{1/2} + \left(\frac{2L\xi}{\lambda}\|g_*\|_{\mathcal{H}}\right)^{1/2}\right).\end{aligned}$$

In the second inequality, we used $\|g_*\|_{\mathcal{H}} \geq \|g_\lambda\|_{\mathcal{H}}$ and $\mathcal{L}(g_*) + \frac{\lambda}{2}\|g_*\|_{\mathcal{H}}^2 \geq \mathcal{L}(g_\lambda) + \frac{\lambda}{2}\|g_\lambda\|_{\mathcal{H}}^2$. In the third inequality, we used $\sqrt{a} + \sqrt{b} \geq \sqrt{a+b}$ for $a, b > 0$. Then by Assumption 4.2, we obtain

$$\begin{aligned}\|g_{M,\lambda} - \widetilde{g}_\lambda\|_{L^\infty(d\rho_{\mathcal{X}})} \leq R\max\Bigg\{&\left(\frac{12L\xi}{\lambda}\|g_*\|_{\mathcal{H}}\right)^{1/2}, \\ &\left(\frac{72L^2\xi^2}{\lambda^3}\mathcal{L}(g_*)\right)^{1/4}, \left(\frac{72L^3\xi^3}{\lambda^3}\|g_*\|_{\mathcal{H}}\right)^{1/4}\Bigg\}. \quad \text{(E.11)}\end{aligned}$$

On the other hand, by Assumption 4.3, we have

$$\begin{aligned}\|g_\lambda - \widetilde{g}_\lambda\|_{L^\infty(d\rho_{\mathcal{X}})} &\leq C(\delta)\|g_\lambda - \widetilde{g}_\lambda\|_{\mathcal{H}_M^+}^p \|g_\lambda - \widetilde{g}_\lambda\|_{L^2(d\rho_{\mathcal{X}})}^{1-p} \\ &\leq C(\delta)(\|g_\lambda\|_{\mathcal{H}} + \|\widetilde{g}_\lambda\|_{\mathcal{H}_M})^p(\xi\|g_\lambda\|_{\mathcal{H}})^{1-p} \\ &\leq 2^p C(\delta)\xi^{1-p}\|g_*\|_{\mathcal{H}} \quad \text{(E.12)}\end{aligned}$$

with probability at least $1-\delta$. In the second inequality, we used the fact that

$$\|g\|_{\mathcal{H}_M^+} = \inf\{\|g_1\|_{\mathcal{H}} + \|g_2\|_{\mathcal{H}_M} \mid g = g_1 + g_2, g_1 \in \mathcal{H}, g_2 \in \mathcal{H}_M\}$$

from Proposition A.1. Combining (E.11) and (E.12), we have

$$\begin{aligned}\|g_\lambda - g_{M,\lambda}\|_{L^\infty(d\rho_{\mathcal{X}})} &\leq \|g_\lambda - \widetilde{g}_\lambda\|_{L^\infty(d\rho_{\mathcal{X}})} + \|\widetilde{g}_\lambda - g_{M,\lambda}\|_{L^\infty(d\rho_{\mathcal{X}})} \\ &\leq \max\Bigg\{2^{p+1}C(\delta)\|g_*\|_{\mathcal{H}}\xi^{1-p}, R\left(\frac{2^4\cdot 3L\xi}{\lambda}\|g_*\|_{\mathcal{H}}\right)^{1/2}, \\ &\qquad R\left(\frac{2^7\cdot 3^2L^2\xi^2}{\lambda^3}\mathcal{L}(g_*)\right)^{1/4}, R\left(\frac{2^7\cdot 3^2L^3\xi^3}{\lambda^3}\|g_*\|_{\mathcal{H}}\right)^{1/4}\Bigg\}.\end{aligned}$$

As a result, define $\xi > 0$ which satisfies

$$\begin{aligned}\xi = \min\Bigg\{&\left(\frac{\epsilon}{2^{p+1}C(\delta)\|g_*\|_{\mathcal{H}}}\right)^{1/1-p}, \frac{\lambda\epsilon^2}{2^4\cdot 3R^2L\|g_*\|_{\mathcal{H}}}, \\ &\left(\frac{\lambda^3\epsilon^4}{2^7\cdot 3^2R^4L^2\mathcal{L}(g_*)}\right)^{1/2}, \left(\frac{\lambda^3\epsilon^4}{2^7\cdot 3^2R^4L^3\|g_*\|_{\mathcal{H}}}\right)^{1/3}\Bigg\},\end{aligned}$$

then we have $\|g_\lambda - g_{M,\lambda}\|_{L^\infty(d\rho_{\mathcal{X}})} \leq \epsilon$ with probability at least $1-2\delta$. □

Proof of Theorem 4.2

The following theorem shows that if k is a Gaussian kernel and k_M is its random Fourier features approximation, then the norm condition in the assumption is satisfied. The proof is inspired by the analysis of Theorem 4.48 in [58].

Theorem 4.2. *Assume* $\mathrm{supp}(\rho_{\mathcal{X}}) \subset \mathbb{R}^d$ *is a bounded set and* $\rho_{\mathcal{X}}$ *has a density with respect to Lebesgue measure which is uniformly bounded away from 0 and* ∞ *on* $\mathrm{supp}(\rho_{\mathcal{X}})$*. Let* k *be a Gaussian kernel and* $\mathcal{H}$ *be its RKHS, then for any* $m \geq d/2$*, there exists a constant* $C_{m,d} > 0$ *such that*

$$\|f\|_{L^\infty(d\rho_{\mathcal{X}})} \leq C_{m,d} \|f\|_{\mathcal{H}}^{d/2m} \|f\|_{L^2(d\rho_{\mathcal{X}})}^{1-d/2m}$$

for any $f \in \mathcal{H}$*. Also, for any* $M \geq 1$*, let* k_M *be a random Fourier features approximation of* k *with* M *features and* $\mathcal{H}_M^+$ *be a RKHS of* $k + k_M$*. Then with probability at least* $1 - \delta$ *with respect to a sampling of features,*

$$\|f\|_{L^\infty(d\rho_{\mathcal{X}})} \leq C_{m,d} \left(1 + \frac{1}{\delta}\right)^{d/4m} \|f\|_{\mathcal{H}_M^+}^{d/2m} \|f\|_{L^2(d\rho_{\mathcal{X}})}^{1-d/2m}$$

for any $f \in \mathcal{H}_M^+$*.*

Proof. For notational simplicity, we denote $\mathrm{supp}(\rho_{\mathcal{X}})$ by $\mathcal{X}'$. From the boundedness of $\mathcal{X}'$ and the condition on $\rho_{\mathcal{X}}$, the following relation holds for any $f \in L^\infty(d\rho_{\mathcal{X}})$:

$$\|f\|_{L^\infty(d\rho_{\mathcal{X}})} = \|f\|_{L^\infty(\mathcal{X}')} \tag{E.13}$$

$$\|f\|_{L^2(d\rho_{\mathcal{X}})} \geq C_1 \|f\|_{L^2(\mathcal{X}')}, \tag{E.14}$$

where $C_1 > 0$ is a constant. From the discussion after Theorem 4.2, for any $f \in W^m(\mathcal{X}')$ $(m \geq d/2)$ there exists a constant $C_2 > 0$ such that the following inequality holds:

$$\|f\|_{L^\infty(\mathcal{X}')} \leq C_2 \|f\|_{W^m(\mathcal{X}')}^{d/2m} \|f\|_{L^2(\mathcal{X}')}^{1-d/2m}. \tag{E.15}$$

Here $W^m(\mathcal{X}')$ is Sobolev space with order m defined as follows:

$$W^m(\mathcal{X}') = \left\{ f \in L^2(\mathcal{X}') \;\middle|\; \partial^{(\alpha)} f \in L^2(\mathcal{X}') \text{ exists for all } \alpha \in \mathbb{N}^d \text{ with } |\alpha| \leq m \right\},$$

where $\partial^{(\alpha)}$ is the α-th weak derivative for a multi-index $\alpha = (\alpha^{(1)}, \ldots, \alpha^{(d)}) \in \mathbb{N}^d$ with $|\alpha| = \sum_{i=1}^d \alpha^{(i)}$.
Combining (E.13), (E.14) and (E.15), we have

$$\|f\|_{L^\infty(d\rho_{\mathcal{X}})} \leq C \|f\|_{W^m(\mathcal{X}')}^{d/2m} \|f\|_{L^2(d\rho_{\mathcal{X}})}^{1-d/2m}, \tag{E.16}$$

where $C > 0$ is a constant. So it suffices to show that $\mathcal{H}$ and $\mathcal{H}_M^+$ can be continuously embedded in $W^m(\mathcal{X}')$. For $\mathcal{H}$, it can be shown in the same manner as Theorem 4.48 in [58]. For $\mathcal{H}_M^+$, we first define a spectral measure of the kernel function $k + k_M$ as

$$\tau^+(\omega) = \frac{1}{M} \sum_{i=1}^M \delta(\omega - \omega_i) + \tau(\omega),$$

where δ is a Dirac measure on Ω. Then a kernel function $k + k_M$ can be written as

$$(k + k_M)(x, x') = \int_\Omega \varphi(x, \omega)\overline{\varphi(x', \omega)} d\tau^+(\omega),$$

and from [9], for any $f \in \mathcal{H}_M^+$, there exists $g \in L^2(d\tau^+)$ such that

$$\begin{aligned} f(x) &= \int_\Omega g(\omega)\varphi(x, \omega) d\tau^+(\omega), \\ \|f\|_{\mathcal{H}_M^+} &= \|g\|_{L^2(d\tau^+)}. \end{aligned}$$

Let us fix a multi-index $\alpha = (\alpha^{(1)}, \ldots, \alpha^{(d)}) \in \mathbb{N}^d$ and $|\alpha| = m$. For $\alpha \in \mathbb{N}^d$, we write $\partial^\alpha = \partial_1^{\alpha^{(1)}} \cdots \partial_d^{\alpha^{(d)}}$. We then have

$$\begin{aligned} \|\partial^\alpha f\|_{L^2(\mathcal{X}')}^2 &= \int_{\mathcal{X}'} \left(\partial_x^\alpha \int_\Omega g(\omega)\varphi(x, \omega) d\tau^+(\omega) \right)^2 dx \\ &\leq \int_{\mathcal{X}'} \left(\int_\Omega |g(\omega)| \partial_x^\alpha \varphi(x, \omega) d\tau^+(\omega) \right)^2 dx \\ &\leq \|g\|_{L^2(d\tau^+)}^2 \int_{\mathcal{X}'} \int_\Omega |\partial_x^\alpha \varphi(x, \omega)|^2 d\tau^+(\omega) dx. \end{aligned}$$

Because we consider φ as a random Fourier feature, $\Omega = \mathbb{R}^d$ and

$$\begin{aligned} \varphi(x, \omega) &= C' e^{i\omega^\top x}, \\ \partial_x^\alpha \varphi(x, \omega) &= \omega^\alpha C' e^{i\omega^\top x} \end{aligned}$$

where $C' > 0$ is a normalization constant and $\omega^\alpha = \prod_{i=1}^d {\omega^{(i)}}^{\alpha_i}$ for $\omega = (\omega^{(1)}, \ldots, \omega^{(d)}) \in \mathbb{R}^d$ and $\alpha = (\alpha^{(1)}, \ldots, \alpha^{(d)}) \in \mathbb{N}^d$. So we have

$$\begin{aligned} \|\partial^\alpha f\|_{L^2(\mathcal{X}')}^2 &\leq \|g\|_{L^2(d\tau^+)}^2 \int_{\mathcal{X}'} C'^2 \int_\Omega \omega^{2\alpha} d\tau^+(\omega) dx \\ &\leq C'^2 \mathrm{vol}(\mathcal{X}') \|f\|_{\mathcal{H}_M^+}^2 \left(\mathbb{E}_{\omega \sim \tau} \left[\omega^{2\alpha} \right] + \frac{1}{M} \sum_{i=1}^M \omega_i^{2\alpha} \right). \end{aligned}$$

We note that because τ is Gaussian, $\mathbb{E}_{\omega \sim \tau} \left[\omega^{2\alpha} \right]$ is finite for any $\alpha \in \mathbb{N}^d$. Because $\omega_i \sim \tau$ and $\omega_i^{2\alpha}$ is non negative, from Markov's inequality we have

$$\frac{1}{M} \sum_{i=1}^M \omega_i^{2\alpha} \leq \frac{1}{\delta} \mathbb{E}_{\omega \sim \tau} \left[\omega^{2\alpha} \right]$$

with probability at least $1 - \delta$. As a result, we have

$$\|\partial^\alpha f\|_{L^2(\mathcal{X}')}^2 \leq \left(1 + \frac{1}{\delta} \right) C'^2 \mathrm{vol}(\mathcal{X}') \|f\|_{\mathcal{H}_M^+}^2 \mathbb{E}_{\omega \sim \tau} \left[\omega^{2\alpha} \right].$$

So we can compute Sobolev norms of f as follows:

$$\begin{aligned} \|f\|_{W^m(\mathcal{X}')}^2 &= \sum_{|\alpha| \leq m} \|\partial^\alpha f\|_{L^2(\mathcal{X}')}^2 \\ &\leq \left(1 + \frac{1}{\delta} \right) C'^2 \mathrm{vol}(\mathcal{X}') \|f\|_{\mathcal{H}_M^+}^2 \sum_{|\alpha| \leq m} \mathbb{E}_{\omega \sim \tau} \left[\omega^{2\alpha} \right]. \end{aligned} \tag{E.17}$$

Substitute (E.17) to (E.16) and define $C_{m,d} = C\left(C'^2 \mathrm{vol}(\mathcal{X}') \sum_{|\alpha|\leq m} \mathbb{E}_{\omega\sim\tau}\left[\omega^{2\alpha}\right]\right)^{d/4m}$, we get a desired result. □

Remark. We note that the assumption that k is Gaussian is only used to derive $\mathbb{E}_{\omega\sim\tau}\left[\omega^{2\alpha}\right]$ is finite for all $\alpha \in \mathbb{N}^d$. This means that if $\psi(x-y) = k(x-y)$ belongs to Schwartz class (a space of rapidly decreasing function) [65], its Fourier transform τ also belongs to this class, thus the above finite moment property is satisfied.

Proof of Theorem 4.3

In this section, we provide the complete statement and the proof of Theorem 4.3. First, we provide a useful proposition which is appeared in [42]. this result suggests that there exists a sufficiently small $\lambda > 0$ such that g_λ is also the Bayes classifier.

Proposition E.1 (Proposition A in [42])**.** *Suppose Assumption 4.3, 4.5, 4.6, 4.7 hold. Then, there exists $\lambda > 0$ such that $\|g_\lambda - g_*\|_{L^\infty(d\rho_\mathcal{X})} \leq m(\delta)/2$.*

Our main result about the exponential convergence of the expected classification error is shown as follows.

Theorem 4.3. *Suppose Assumptions 4.1-4.7 holds. There exists a sufficiently small $\lambda > 0$ such that the following statement holds:*
Taking the number of random features M that satisfies

$$M \geq \max\left\{\frac{8}{3}\left(\frac{R}{\xi}\right)^2, 32\left(\frac{R}{\xi}\right)^4\right\} \log\frac{2R^2}{\|T\|_{\mathrm{op}}\delta} \tag{E.18}$$

where $\xi > 0$ is defined as below:

$$\xi = \min\left\{\left(\frac{m(\delta)}{2^{p+3}C(\delta')\|g_*\|_\mathcal{H}}\right)^{1/1-p}, \frac{\lambda m^2(\delta)}{2^8 \cdot 3R^2 G\|g_*\|_\mathcal{H}}, \left(\frac{\lambda^3 m^4(\delta)}{2^{15}\cdot 3^2 R^4 G^2 \mathcal{L}(g_*)}\right)^{1/2}, \left(\frac{\lambda^3 m^4(\delta)}{2^{15}\cdot 3^2 R^4 G^3 \|g_*\|_\mathcal{H}}\right)^{1/3}\right\}.$$

Consider Algorithm 4.1 with $\eta_t = \frac{2}{\lambda(\gamma+t)}$ and $\alpha_t = \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)}$ where γ is a positive value such that $\|g_1\|_{\mathcal{H}_M} \leq (2\eta_1 + 1/\lambda)GR$ and $\eta_1 \leq \min\{1/L, 1/2\lambda\}$. Then, with probability $1-2\delta'$, for sufficiently large T such that

$$\max\left\{\frac{36G^2R^2}{\lambda^2(2\gamma+T)}, \frac{\gamma(\gamma-1)\|g_1 - g_{M,\lambda}\|^2_{\mathcal{H}_M}}{(2\gamma+T)(T+1)}\right\} \leq \frac{m^2(\delta)}{64R^2},$$

we have the following inequality for any $t > T$:

$$\mathbb{E}\left[\mathcal{R}(\overline{g}_{T+1}) - \mathcal{R}(\mathbb{E}[Y|x])\right] \leq 2\exp\left(-\frac{\lambda^2(2\gamma+t)m^2(\delta)}{2^{12}\cdot 9G^2R^4}\right).$$

Proof. Fix $\lambda > 0$ satisfying the condition in Proposition E.1. From Theorem 4.1, if we set a number of features M satisfying (E.18), we have

$$\begin{aligned}\|g_{M,\lambda} - g_*\|_{L^\infty(d\rho_\mathcal{X})} &\leq \|g_{M,\lambda} - g_\lambda\|_{L^\infty(d\rho_\mathcal{X})} + \|g_\lambda - g_*\|_{L^\infty(d\rho_\mathcal{X})} \\ &\leq \frac{m(\delta)}{4} + \frac{m(\delta)}{2} = \frac{3m(\delta)}{4}.\end{aligned}$$

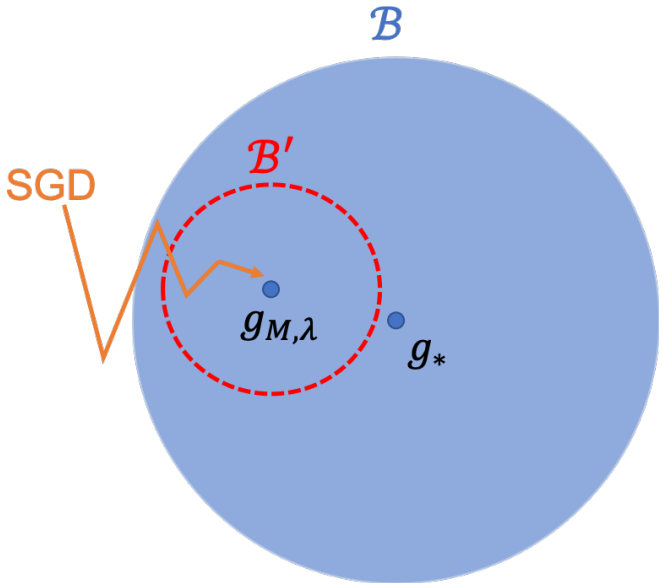


Fig. E.2: A schematic drawing of the proof of Theorem 4.3;
We can take a ball $\mathcal{B}'$ around $g_{M,\lambda}$ in which all hypotheses are the Bayes classifiers if $g_{M,\lambda}$ is sufficiently close to g_*.

Then $\mathrm{sgn}(g(X)) = \mathrm{sgn}(g_*(X))$ almost surely for any $g \in \mathcal{H}_M$ satisfying $\|g - g_{M,\lambda}\|_{\mathcal{H}_M} \leq m(\delta)/4R$, since

$$\begin{aligned}\|g - g_*\|_{L^\infty(d\rho_\mathcal{X})} &\leq \|g - g_{M,\lambda}\|_{L^\infty(d\rho_\mathcal{X})} + \|g_{M,\lambda} - g_*\|_{L^\infty(d\rho_\mathcal{X})} \\ &\leq R\|g - g_{M,\lambda}\|_{\mathcal{H}_M} + \|g_{M,\lambda} - g_*\|_{L^\infty(d\rho_\mathcal{X})} \\ &\leq \frac{m(\delta)}{4} + \frac{3m(\delta)}{4} = m(\delta)\end{aligned}$$

and $|g_*(X)| \geq m(\delta)$ almost surely. In other words, g is also the Bayes classifier of $\mathcal{R}(g)$. Assume

$$\|\mathbb{E}[\overline{g}_{T+1}] - g_{M,\lambda}\|_{\mathcal{H}_M} \leq \frac{m(\delta)}{8R}. \tag{E.19}$$

Then, substituting $\epsilon = m(\delta)/8R$ in Proposition 2.4, we have

$$\left\|\overline{g}_{T+1} - g_{M,\lambda}\right\|_{\mathcal{H}_M} \leq \left\|\overline{g}_{T+1} - \mathbb{E}[\overline{g}_{T+1}]\right\|_{\mathcal{H}_M} + \|\mathbb{E}[\overline{g}_{T+1}] - g_{M,\lambda}\|_{\mathcal{H}_M} \leq \frac{m(\delta)}{4R}$$

with probability at least $1 - 2\exp\left(-\frac{\lambda^2(2\gamma+T)m^2(\delta)}{2^{12}\cdot 3^2 G^2 R^4}\right)$. In other words, $\overline{g}_{T+1}$ is also the Bayes classifier with same probability. By definition of the expected classification error, we have

$$\mathbb{E}[\mathcal{R}(\overline{g}_{T+1})] - \mathcal{R}(\mathbb{E}[Y|x]) \leq 1 - 2\exp\left(-\frac{\lambda^2(2\gamma+T)m^2(\delta)}{2^{12}\cdot 3^2 G^2 R^4}\right).$$

Finally, to satisfy (E.19), the required number of iteration T is obtained by Proposition 2.3, which completes the proof. □