Exponential Convergence Rates of Classification Errors on Learning with SGD and Random Features

Shingo Yashima[1] syashima9@gmail.com
Atsushi Nitanda[1,2,3] nitanda@mist.i.u-tokyo.ac.jp
Taiji Suzuki[1,2] taiji@mist.i.u-tokyo.ac.jp
[1]Graduate School of Information Science and Technology, The University of Tokyo
[2]Center for Advanced Intelligence Project, RIKEN
[3]PRESTO, Japan Science and Technology Agency

Abstract

Although kernel methods are widely used in many learning problems, they have poor scalability to large datasets. To address this problem, sketching and stochastic gradient methods are the most commonly used techniques to derive computationally efficient learning algorithms. We consider solving a binary classification problem using random features and stochastic gradient descent, both of which are common and widely used in practical large-scale problems. Although there are plenty of previous works investigating the efficiency of these algorithms in terms of the convergence of the objective loss function, these results suggest that the computational gain comes at expense of the learning accuracy when dealing with general Lipschitz loss functions such as logistic loss. In this study, we analyze the properties of these algorithms in terms of the convergence not of the loss function, but the classification error under the strong low-noise condition, which reflects a realistic property of real-world datasets. We extend previous studies on SGD to a random features setting, examining a novel analysis about the error induced by the approximation of random features in terms of the distance between the generated hypothesis to show that an exponential convergence of the expected classification error is achieved even if random features approximation is applied. 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 in classification problems under the strong low-noise condition.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 Introduction

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 (Caponnetto and De Vito, 2007; Steinwart et al., 2009; Mendelson and Neeman, 2010). 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 (Rahimi and Recht, 2008) 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 (Sriperumbudur and Szabó, 2015; Sutherland and Schneider, 2015; Szabó and Sriperumbudur, 2019), 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 by Rudi and Rosasco (2017) and Carratino et al. (2018). 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. By Rahimi and Recht (2009), 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, by Li et al. (2019), 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* (Tsybakov, 2004; Koltchinskii and Beznosova, 2005) that conditional label probabilities are uniformly bounded away from 1/2 (Pillaud-Vivien et al., 2018; Nitanda and Suzuki, 2019). 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 and show the superiority of random features (Bach, 2017b; Rudi and Rosasco, 2017; Li et al., 2019; Sun et al., 2018), we do not explore this direction and treat the original random features because these distributions are generally intractable or require much computational cost to sample (Bach, 2017b) whereas an efficient sampling algorithm is proposed in the case of Gaussian kernel (Avron et al., 2017).

In addition, we should refer to Nyström method (Williams and Seeger, 2001), 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.

Our Contributions Our contributions are twofold. First, we analyze the error induced by the approximation of random features in terms of the L^∞-norm between the generated hypothesis including population risk minimizers and empirical risk minimizers when using general Lipschitz loss functions in Section 3. Our results can be framed as an extension of the analysis of Cortes et al. (2010); Sutherland and Schneider (2015), which analyzed the error in terms of the distance between empirical risk minimizers when using a hinge loss. However, it is not straightforward to extend these results to our case since we cannot access the closed-form solutions, unlike those previous results, when using the general loss functions and treating population risk minimizers. In addition, since the true and the approximated minimizer lie in different function spaces, it is not easy to derive L^∞-norm bound between them. We deal with these difficulties with novel proof techniques. 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 in Section 4. Then 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. We also validate these results through experiments on synthetic datasets in Section 5.

2 Problem Setting

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 general, for a probability measure μ, $L^2(\mu)$ denotes a space of square-integrable functions with respect to μ, and $L^2(\mathcal{X})$ denotes one with respect to the Lebesgue measure. Similarly, $L^\infty(\mu)$ denotes a space of functions for which the essential supremum with respect to μ is bounded, and $L^\infty(\mathcal{X})$ denotes one with respect to Lebesgue measure.

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:

$$\mathcal{R}(g) \stackrel{\text{def}}{=} \mathbb{E}_{(X,Y)\sim\rho}\left[I(\operatorname{sgn}(g(X)), Y)\right], \tag{1}$$

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

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

By definition, $g(x) = \mathbb{E}[Y|x] = 2\rho(1|x) - 1$ minimizes $\mathcal{R}$. However, directly minimizing (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 loss function $\mathcal{L}(g)$ of l:

$$\mathcal{L}(g) \stackrel{\text{def}}{=} \mathbb{E}_{(X,Y)\sim\rho}\left[l(g(X), Y)\right]. \tag{2}$$

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 loss function (2) ensures minimizing the expected classification (1) if l is *classification-calibrated* (Bartlett et al., 2006), which has been proven for several practically implemented losses including hinge loss and logistic loss.

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 (Aronszajn, 1950) 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:

$$\min_{g\in\mathcal{H}} \ \mathcal{L}(g) + \frac{\lambda}{2}\|g\|_{\mathcal{H}}^2 \tag{3}$$

where $\lambda > 0$ is a regularization parameter.
However, because solving the original problem (3) is usually computationally inefficient for large-scale datasets, the approximation method is applied in practice. Random features (Rahimi and Recht, 2008) 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:

$$k(x, y) = \int_\Omega \varphi(x, \omega)\overline{\varphi(y, \omega)}\mathrm{d}\tau(\omega). \tag{4}$$

The main idea behind random features is to approximate the integral (4) by its Monte-Carlo estimate:

$$k_M(x, y) \stackrel{\text{def}}{=} \frac{1}{M}\sum_{i=1}^{M} \varphi(x, \omega_i)\overline{\varphi(y, \omega_i)}, \quad \omega_i \overset{i.i.d.}{\sim} \tau. \tag{5}$$

For example, if k is a shift invariant kernel, by Bochner's theorem (Yoshida, 1995), the expansion (4) is achieved with $\varphi(x, \omega) = C' e^{i\omega^\top x}$, where C' is a normalization constant. Then, the approximation (5) is called random Fourier features (Rahimi and Recht, 2008), which is the most widely used variant of random features.
We denote the RKHS associated with k and k_M by $\mathcal{H}$ and $\mathcal{H}_M$, respectively. These spaces then admit the following explicit representation (Bach, 2017b,a):

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

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:

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

As a result, the problem (3) in the approximation space $\mathcal{H}_M$ is reduced to the following generalized linear model:

$$\min_{\beta\in\mathbb{R}^M} \ \mathcal{L}(\beta^\top \phi_M) + \frac{\lambda}{2}\|\beta\|_2^2 \tag{6}$$

where ϕ_M is a feature vector:

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

In this paper, we consider solving the problem (6) using the averaged SGD. The details are discussed in the following section.

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 an optimization procedure. For the minimization problem (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 $\mathcal{H}_M$, because a gradient on $\mathcal{H}_M$ is given by

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

We consider the averaged variants of SGD, since it is widely known that gradient averaging gives faster convergence than plain SGD on strongly convex problems (Lacoste-Julien et al., 2012). The algorithm of random features and averaged SGD is described in Algorithm 1. Following Nitanda and Suzuki (2019), 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 $(\beta_t)_{t=1}^{T+1}$.

Algorithm 1 Random Feature + 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 $\omega_1, \dots, \omega_M \sim \tau$
Initialize $\beta_1 \in \mathbb{R}^M$
for $t = 1, \dots, T$ **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 $\overline{g}_{T+1} = \overline{\beta}_{T+1}^\top \phi_M$

Computational Complexity If we assume the evaluation of a feature map $\varphi(x, \omega)$ to have a constant cost, one iteration in Algorithm 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.

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 $L^\infty(\rho_\mathcal{X})$-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.

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 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 $y \in \mathcal{Y}$,*

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

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

$$|\mathcal{L}(g) - \mathcal{L}(h)| \le G \int |g(x) - h(x)| \mathrm{d}\rho_\mathcal{X}(x)$$
$$\le G\|g - h\|_{L^2(\rho_\mathcal{X})}$$

for any $g, h \in L^2(\rho_\mathcal{X})$. 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 2. *The function φ is continuous and there exists $R > 0$ such that $|\varphi(x, \omega)| \le R$ 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 $\sup_{x,y\in\mathcal{X}} k(x, y) \le R^2, \sup_{x,y\in\mathcal{X}} k_M(x, y) \le R^2$ and it leads to an important relationship $R\|\cdot\|_\mathcal{H} \ge \|\cdot\|_{L^\infty(\mathcal{X})}, R\|\cdot\|_{\mathcal{H}_M} \ge \|\cdot\|_{L^\infty(\mathcal{X})}$.
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 3. *Let $\mathcal{H}_M^+$ be RKHS with the kernel function $k + k_M$. Then there exists $0 \le p < 1$, and a constant $C(\delta) > 0$ depends on $0 < \delta \le 1$ that satisfies, for any $f \in \mathcal{H}_M^+$,*

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

with probability at least $1-\delta$.

For a fixed kernel function, the Assumption 3 is a commonly used condition in analysis of kernel methods (Steinwart et al., 2009; Mendelson and Neeman, 2010). It is satisfied, for example, when the eigenfunctions of the kernel are uniformly bounded and the eigenvalues $\{\mu_i\}_i$ decay at the rate $i^{-1/p}$ (Mendelson and Neeman, 2010). In Theorem 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 $L^\infty(\rho_\mathcal{X})$-norm. The complete statement, including proof and all constants, are found in Appendix C.

Theorem 1. *Under Assumption 1-3, with probability at least* $1-2\delta$ *with respect to the sampling of features, the following inequality holds:*

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

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 $L^2(\rho_\mathcal{X})$-norm without Assumption 3.

Comparison to Previous Results 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 by Cortes et al. (2010); Sutherland and Schneider (2015). Considering K and K_M to be Gram matrices of kernel k and k_M, respectively, they showed that $\|g_\lambda - g_{M,\lambda}\|_{L^\infty(\rho_\mathcal{X})} \lesssim O(\|K-K_M\|_{\mathrm{op}}^{1/4})$, where $\|\cdot\|_{\mathrm{op}}$ is an operator norm, defined in Appendix A. Because the Gram matrix can be considered as the integral operator on the empirical measure, we can apply Lemma 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. Although it is relatively easy to derive the infinite norm bound in those finite dimentional case, more careful deriviation is needed in our setting (infinite dimentional case) and our analysis is novel.
The case of squared loss was studied by Rudi and Rosasco (2017); Carratino et al. (2018). In particular, in Lemma 8 of Rudi and Rosasco (2017), 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 (Cortes et al., 2010).
Approximation of functions in $\mathcal{H}$ by functions in $\mathcal{H}_M$ is also considered by Bach (2017b), 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 (Rahimi and Recht, 2009; Li et al., 2019), 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.

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 3 is satisfied for any $0<p<1$.

Theorem 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*

$$\|f\|_{L^\infty(\rho_\mathcal{X})} \leq C_{m,d} \|f\|_\mathcal{H}^{d/2m} \|f\|_{L^2(\rho_\mathcal{X})}^{1-d/2m} \tag{7}$$

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,*

$$\|f\|_{L^\infty(\rho_\mathcal{X})} \leq C_{m,d}\left(1+\frac{1}{\delta}\right)^{d/4m} \|f\|_{\mathcal{H}_M^+}^{d/2m} \|f\|_{L^2(\rho_\mathcal{X})}^{1-d/2m} \tag{8}$$

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

We note that the norm relation of the Gaussian RKHS (7) is a known result of Steinwart et al. (2009) 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 Steinwart et al. (2009) 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 Steinwart and Christmann (2008)). Also, by Sobolev's embedding

theorem for Besov space, $B^{d/2}_{2,1}(\mathcal{X}')$ 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

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

Therefore, if it can be shown that RKHS $\mathcal{H}^+_M$ is continuously embedded in $W^m(\mathcal{X}')$, the norm relation (8) holds. The complete proof is found in Appendix D. Using this theorem, it can be shown that in the case of a Gaussian kernel and its random Fourier features approximation, Assumption 3 is satisfied with $p = 1/2$ and $C(\delta) = C_{d,d}(1 + 1/\delta)^{1/4}$, and the resulting rate in Theorem 1 is $O(M^{-1/8})$.

4 Main Result

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

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

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 and 4 yields LR^2-smoothness in $\mathcal{H}$, since

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

holds for any $z \in \mathcal{X} \times \mathcal{Y}$ and it is known as an equivalent condition of smoothness by Theorem 2.1.5 of Nesterov (2014). The second is the margin condition on the conditional label probability.

Assumption 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_* (Bartlett et al., 2006; Zhang, 2004), 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 (Zhang, 2004). Although $h_*(\mu)$ may not be uniquely determined nor well-defined in general, the following assumption ensures these properties.

Assumption 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 Zhang (2004), 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 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_*} (Zhang, 2004):

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

So, combining Assumptions 6 and 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 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 $m(\delta) = \log((1 + 2\delta)/(1 - 2\delta))$. Under these assumptions and notations, the exponential convergence of the expected classification error is shown.

Theorem 3. *Suppose Assumptions 1–7 hold. There exists a sufficiently small $\lambda > 0$ such that the following statement holds:*
Taking the number of random features M that satisfies

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

Consider Algorithm 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 (Rudi and Rosasco, 2017; Carratino et al., 2018; Rahimi and Recht, 2009; Li et al., 2019). 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 later.

Dependence of γ and λ on T As we can see in the condition inequality on T, γ adjusts the step size and consequently affects T, when the exponential convergence phase starts. Indeed, there is a trade-off relation in T, that is, the first part of max in the condition on T is $O(1/\gamma)$ and the second part is $O(\gamma)$. In addition, we note that when we apply non-averaged SGD, the dependence of λ on T is worse than our averaged SGD, although similar exponential convergence can be shown in such case. This comes from the fact that gradient averaging achieves better dependence on the strongly convex parameter. For further details, see Nitanda and Suzuki (2019); Lacoste-Julien et al. (2012).

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 3 and applying Theorem 2. In addition, we can specify a required λ to achieve the convergence in this setting. The complete statement and proof are given in Appendix F.

Corollary 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 5-7, the following statement holds:*
Taking a regularization parameter λ *and a number of random features* M *that satisfies*

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

Consider Algorithm 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 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 (Rudi and Rosasco, 2017; Li et al., 2019; Rahimi and Recht, 2009). 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 2.3, if one runs SGD in Algorithm 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 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 (Rudi and Rosasco, 2017; Carratino et al., 2018); this is the first time the superiority of random features has been demonstrated in the classification setting. Moreover,

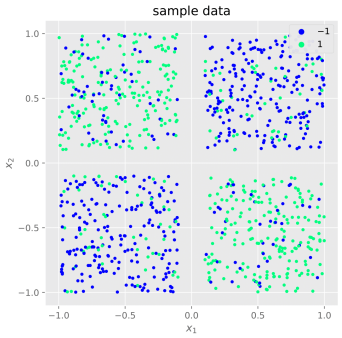
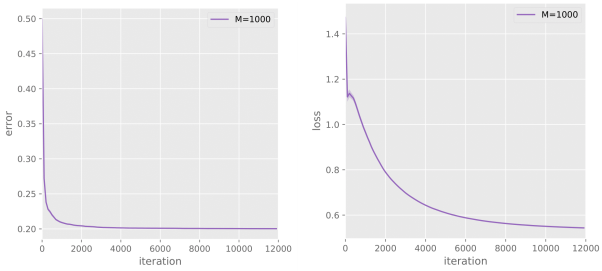


Figure 1: Subsample of data used in the experiment.



(a) Classification errors (b) Loss functions

Figure 2: Learning curves of the expected classification error (left) and the expected loss function (right) by averaged SGD with 1000 features.

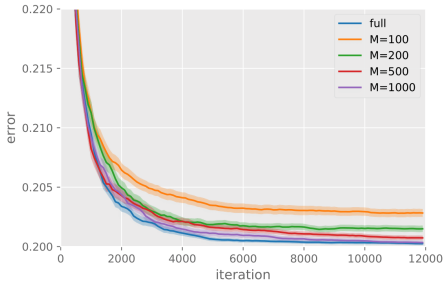


Figure 3: Comparison of learning curves of the expected classification error with varying numbers of features.

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.

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 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 $\lambda = 0.001$. SGD 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 loss function are drawn when the number of features $M = 1000$, as shown in Figure 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 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. 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, 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.

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

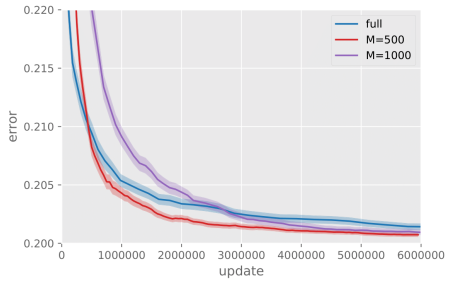


Figure 4: Comparison of learning curves with respect to number of parameter updates.

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 Pillaud-Vivien et al. (2018) did 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 (Schmidt et al., 2017; Johnson and Zhang, 2013; Defazio et al., 2014; Allen-Zhu, 2017).

Acknowledgements

AN was partially supported by JSPS KAKENHI (19K20337) and JST PRESTO. TS was partially supported by JSPS KAKENHI (18K19793, 18H03201, and 20H00576), Japan Digital Design, and JST CREST.

References

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

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

Avron, H., Kapralov, M., Musco, C., Musco, C., Velingker, A., and Zandieh, A. (2017). Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International Conference on Machine Learning*, pages 253–262.

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

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

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

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

Carratino, L., Rudi, A., and Rosasco, L. (2018). Learning with SGD and random features. In *Advances in Neural Information Processing Systems*, pages 10192–10203.

Cortes, C., Mohri, M., and Talwalkar, A. (2010). On the impact of kernel approximation on learning accuracy. In *International Conference on Artificial Intelligence and Statistics*, pages 113–120.

Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654.

Dieuleveut, A. and Bach, F. (2016). Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323.

Koltchinskii, V. and Beznosova, O. (2005). Exponential convergence rates in classification. In *International Conference on Computational Learning Theory*, pages 295–307. Springer.

Lacoste-Julien, S., Schmidt, M., and Bach, F. (2012). A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*.

Li, Z., Ton, J.-F., Oglic, D., and Sejdinovic, D. (2019). Towards a unified analysis of random Fourier features. In *International Conference on Machine Learning*, pages 3905–3914.

Mendelson, S. and Neeman, J. (2010). Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565.

Nesterov, Y. (2014). *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company.

Nitanda, A. and Suzuki, T. (2019). Stochastic gradient descent with exponential convergence rates of expected classification errors. In *International Conference on Artificial Intelligence and Statistics*, pages 1417–1426.

Pillaud-Vivien, L., Rudi, A., and Bach, F. (2018). Exponential convergence of testing error for stochastic gradient methods. In *Conference on Learning Theory*, pages 250–296.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184.

Rahimi, A. and Recht, B. (2009). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems*, pages 1313–1320.

Rudi, A. and Rosasco, L. (2017). Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3215–3225.

Schmidt, M., Le Roux, N., and Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112.

Sriperumbudur, B. and Szabó, Z. (2015). Optimal rates for random Fourier features. In *Advances in Neural Information Processing Systems*, pages 1144–1152.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Science & Business Media.

Steinwart, I., Hush, D. R., and Scovel, C. (2009). Optimal rates for regularized least squares regression. In *Conference on Learning Theory*, pages 79–93.

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

Sun, Y., Gilbert, A., and Tewari, A. (2018). But how does it work in theory? linear svm with random features. In *Advances in Neural Information Processing Systems*, pages 3379–3388.

Sutherland, D. J. and Schneider, J. (2015). On the error of random Fourier features. In *Conference on Uncertainty in Artificial Intelligence*, pages 862–871.

Szabó, Z. and Sriperumbudur, B. (2019). On kernel derivative approximation with random Fourier features. In *International Conference on Artificial Intelligence and Statistics*, pages 827–836.

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

Williams, C. K. and Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, pages 682–688.

Yoshida, K. (1995). *Functional Analysis*. Springer-Verlag Berlin Heidelberg.

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