Stochastic Gradient Descent with Exponential Convergence Rates of Expected Classification Errors

Atsushi Nitanda[1,2] **Taiji Suzuki**[1,2]
[1]Graduate School of Information Science and Technology, The University of Tokyo
[2]Center for Advanced Intelligence Project, RIKEN

Abstract

We consider stochastic gradient descent and its averaging variant for binary classification problems in a reproducing kernel Hilbert space. In traditional analysis using a consistency property of loss functions, it is known that the expected classification error converges more slowly than the expected risk even when assuming a low-noise condition on the conditional label probabilities. Consequently, the resulting rate is sublinear. Therefore, it is important to consider whether much faster convergence of the expected classification error can be achieved. In recent research, an exponential convergence rate for stochastic gradient descent was shown under a strong low-noise condition but provided theoretical analysis was limited to the squared loss function, which is somewhat inadequate for binary classification tasks. In this paper, we show an exponential convergence of the expected classification error in the final phase of the stochastic gradient descent for a wide class of differentiable convex loss functions under similar assumptions. As for the averaged stochastic gradient descent, we show that the same convergence rate holds from the early phase of training. In experiments, we verify our analyses on the L_2-regularized logistic regression.

1 Introduction

The ultimate goal of binary classification problems is to find the Bayes classifier that minimizes the expected classification error in the space of all measurable functions. Usually, this goal is achieved by approximating the classification error with a convex surrogate loss function and solving the expected risk minimization problem defined by this surrogate loss. Such an approximation is theoretically justified by the consistency property (Zhang, 2004; Bartlett et al., 2006) of loss functions, which gives the connection between the *excess risk* (equivalent to the expected risk minus the Bayes risk which is the minimum expected risk over all measurable functions) and the *excess classification error* (equivalent to the expected classification error minus the Bayes classification error which is the error of Bayes classifier).

Stochastic gradient descent (Robbins and Monro, 1951) is the workhorse method for large-scale machine learning problems, including the binary classification, owing to its scalability, wide applicability for various problems, simplicity of implementation, and superior performance. Hence, there is a great deal of research into improving its performance and analyzing the convergence behavior under various problem settings. A popular variant of the method is averaged stochastic gradient descent (Ruppert, 1988; Polyak and Juditsky, 1992; Rakhlin et al., 2012; Lacoste-Julien et al., 2012), which returns a weighted average of parameters obtained by stochastic gradient descent to stabilize the iterates. Moreover, these methods have been generalized into a kernel setting (Cesa-Bianchi et al., 2004; Smale and Yao, 2006; Ying and Zhou, 2006). The convergence rates for the expected risk minimization have been also well studied. For instance, the rates of $O(1/\sqrt{T})$ and $O(1/T)$, where T is the number of iterations, were obtained in Nemirovski et al. (2009); Bach and Moulines (2011); Rakhlin et al. (2012); Lacoste-Julien et al. (2012); Ghadimi and Lan (2013); Bubeck (2015); Bottou et al. (2018) for the general convex and strongly convex problems, and these rates are known to be asymptotically optimal (Nemirovskii and Yudin, 1983; Agarwal et al., 2009). Note that convergence rates of excess classification errors can be simply derived from these rates with the consistency property of loss functions and can be accelerated by some preferable assumptions such as the low-noise condition (Tsybakov, 2004; Bartlett et al., 2006) on the conditional probability of the class label (c.f., Ying and Zhou (2006)), but obtained rates of excess classification errors are generally slower than those of excess risk functions.

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

In Audibert and Tsybakov (2007); Koltchinskii and Beznosova (2005), it is shown that the convergence rate of the excess classification error of the empirical risk minimizer can be exponentially fast by assuming the *strong low-noise condition* that conditional label probabilities are uniformly bounded away from $1/2$, although the excess risk converges at sublinear rate. This is a rather surprising result because exponential convergence is significantly faster than sublinear convergence. However, these theories are insufficient to explain the great success of stochastic gradient descent. More recently, Pillaud-Vivien et al. (2017) has provided direct analysis concerning an exponential convergence property of stochastic gradient descent in a reproducing kernel Hilbert space, but Pillaud-Vivien et al. (2017) adopts the squared loss function, which is somewhat inadequate for the binary classification problems (Rosasco et al., 2004).

Our contribution In this paper, we extend the results in Pillaud-Vivien et al. (2017) to general loss functions which are more appropriate for classification problems by utilizing a different strategy of the proof from the one in Pillaud-Vivien et al. (2017). That is, we show the exponential convergence of the excess classification error in the final phase of the learning procedure using stochastic gradient descent for binary classification problems defined by a wide class of differentiable classification loss functions, including the logistic loss and the exponential loss. Since, a method considered in our analysis corresponds to the common form of stochastic gradient descent, the traditional sublinear convergence rates of $O(1/T^q)$ ($q \in (0,1)$) also hold in the overall learning procedure. In that sense, our result implies acceleration of the excess classification error in the final phase of stochastic gradient descent. In addition, we show a much better convergence result for the averaged stochastic gradient which reduces a threshold for the beginning time (the number of iteration) of exponential convergence. As a result, we conclude that the averaged stochastic gradient exhibits exponential convergence from the early phase of the learning procedure in practice. Moreover, an obtained convergence rate is the same as that in Pillaud-Vivien et al. (2017) for the squared loss function. Although, these results may be further improved by making an additional assumption on a decreasing rate of eigen-values of the covariance operator as shown in Pillaud-Vivien et al. (2017), we do not treat it in this study for the simplicity. Namely, we generalize the result in Pillaud-Vivien et al. (2017) without degradation under general settings.

Technical difficulties We next explain our technical contributions. An obtained result in this work is a generalization of Pillaud-Vivien et al. (2017) and an outline of the proof is essentially the same as that in Pillaud-Vivien et al. (2017), but we emphasize that we use proof techniques that were not argued in Pillaud-Vivien et al. (2017) to overcome several obstacles caused by generalization of loss functions, so that details of the proof are quite different. For instance, the stability argument (Bousquet and Elisseeff, 2002; Hardt et al., 2016; Liu et al., 2017) of stochastic gradient descent is used to bound the error term of it and the property of the link function (Zhang, 2004) is used to show the convergence of the expected classification error from the convergence of the stochastic gradient descent. As a result, (i) obtained convergence rates in this study are much faster than that derived from another concentration inequality such as Kakade and Tewari (2009) and (ii) the overall proof is significantly simplified and shortened without degradation under general settings compared to that of Pillaud-Vivien et al. (2017) which relies on the specific update rule for the squared loss.

We finally note that exponential convergence of the stochastic gradient descent cannot be obtained from analyses (Audibert and Tsybakov, 2007; Koltchinskii and Beznosova, 2005) because these work does not improve the bound of the consistency property of loss functions but only analyze the convergence rate of the empirical risk minimizer. In addition, it is also generally difficult to show that the stochastic gradient descent has the comparative convergence rate to the empirical risk minimizer, thus an analysis of the stochastic gradient descent cannot be reduced to that of the empirical risk minimizer. Even when such an argument is valid, analyses given in (Audibert and Tsybakov, 2007; Koltchinskii and Beznosova, 2005) cannot be utilized under our problem setting because Audibert and Tsybakov (2007) focuses on a more specific model of local polynomial estimators and Koltchinskii and Beznosova (2005) assumes additional assumptions such as the Lipschitz continuity of hypotheses.

2 Problem Setting

In this section, we provide notations to describe a problem setting for the binary classification treated in this paper. Let $\mathcal{X}$ and $\mathcal{Y}$ be a measurable feature space and the set of binary labels $\{-1, 1\}$, respectively. We denote by ρ 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 $(X, Y) \sim \rho$. The ultimate goal in binary classification problems is to find a classifier $g : \mathcal{X} \to \mathbb{R}$ such that $\mathrm{sgn}(g(x))$ correctly classifies its label. In other words, we want to obtain the Bayes classifier derived from $g(x) = \mathbb{E}[Y|x] = 2\rho(1|x) - 1$ that minimizes the expected classification error $\mathcal{R}(g)$ defined below over all measurable functions:

$$\mathcal{R}(g) \stackrel{def}{=} \mathbb{E}_{(X,Y)}[I(\mathrm{sgn}(g(X)), Y)], \tag{1}$$

where the expectation is taken with respect to $(X, Y) \sim \rho$. Here, I is the 0-1 error function:

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

However, since minimizing $\mathcal{R}(g)$ is intractable due to its

non-continuity and non-convexity, we approximate the problem with a convex surrogate loss function $l(\zeta, y)$ for the 0-1 error function and minimize the expected risk defined by this surrogate loss function over a given hypothesis class of functions from $\mathcal{X}$ to $\mathbb{R}$. In general, a loss function l has a form: $l(\zeta, y) = \phi(y\zeta)$ where ϕ is a non-negative convex function from $\mathbb{R}$ to $\mathbb{R}$. Typical examples of such functions are $\phi(v) = \log(1 + \exp(-v))$ for logistic regression and $\phi(v) = \exp(-v)$ for Adaboost. We sometimes denote $z = (x, y)$ and $Z = (X, Y)$ for simplicity. In this paper, we consider a reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_k, \langle , \rangle_{\mathcal{H}_k})$ associated with a real-valued kernel function k on $\mathcal{X}$ as a hypothesis class, and denote by $\| \cdot \|_{\mathcal{H}_k}$ the norm induced by the inner product $\langle , \rangle_{\mathcal{H}_k}$. As a result, the problem to be solved takes the following form:

$$\min_{g \in \mathcal{H}_k} \left\{ \mathcal{L}_\lambda(g) \overset{def}{=} \mathbb{E}_Z[l(g(X), Y)] + \frac{\lambda}{2}\|g\|^2_{\mathcal{H}_k} \right\}. \quad (2)$$

where the last term is the L_2-regularization in $\mathcal{H}_k$ with a regularization parameter $\lambda > 0$. The purpose of the regularization in this paper is to accelerate and stabilize the stochastic gradient descent to solve this expected risk minimization problem. We also denote $\mathcal{L}(g) = \mathcal{L}_0(g)$. Remark that although stochastic gradient descent is used to solve the problem (2), the main interest in this paper is the convergence rate of the expected classification error (1).

3 Stochastic Gradient Descent and its Averaging Variant in RKHS

Stochastic gradient descent and its averaging variant are the most popular methods for solving large-scale machine learning problems. In this paper, we analyze the convergence behavior of the expected classification error for these methods. To do so, we give specific form of (averaged) stochastic gradient descent based on Bottou et al. (2018); Lacoste-Julien et al. (2012) for solving the problem. We first recall the definition of a gradient of a function F on $\mathcal{H}_k$ at $g \in \mathcal{H}_k$; it is an element $\nabla F(g)$ of $\mathcal{H}_k$ satisfying the following equation: for $\forall h \in \mathcal{H}_k$,

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

For the expected risk $\mathcal{L}$, when $k(x, x)$ is bounded on $\mathcal{X}$, its gradient exists and takes the form $\mathbb{E}[\partial_\zeta l(g(X), Y)k(X, \cdot)]$ (ζ is the first variable of l), which is confirmed by the following equations:

$$\begin{aligned}&\mathbb{E}[l((g + h)(X), Y)] \\ &\quad = \mathbb{E}[l(g(X), Y) + \partial_\zeta l(g(X), Y)h(X) + o(|h(X)|)],\end{aligned}$$

$h(X) = \langle h, k(X, \cdot) \rangle_{\mathcal{H}_k}$, and $|h(X)| \leq \|h\|_{\mathcal{H}_k} \sqrt{k(X, X)}$. Thus, the stochastic gradients of $\mathcal{L}$ and $\mathcal{L}_\lambda$ are given by $\partial_\zeta l(g(X), Y)k(X, \cdot)$ and $\partial_\zeta l(g(X), Y)k(X, \cdot) + \lambda g$ for $(X, Y) \sim \rho$. We denote by $G_\lambda(g, Z)$ the latter stochastic gradient. Stochastic gradient descent is described in Algorithm 1. We can also use averaged stochastic gradient descent that returns a weighted average of obtained iterates g_t, rather than the last iterate g_{T+1}. We denote by $\overline{g}_{T+1} = \sum_{t=1}^{T+1} \alpha_t g_t$.

Algorithm 1 Stochastic Gradient Descent with the Averaging Option

Input: number of outer-iterations T, regularization parameter λ, learning rates $(\eta_t)_{t=1}^T$, averaging weights $(\alpha_t)_{t=1}^{T+1}$, initial function g_1
for $t = 1$ **to** 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 g_{T+1} or $\sum_{t=1}^{T+1} \alpha_t g_t$ (averaging option)

In this paper, we adopt the following decreasing learning rate and averaging weight:

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

where γ is an offset parameter for the time index. This learning rate is also used in Bottou et al. (2018). As for an averaging weight, it is a modified version of that introduced in Lacoste-Julien et al. (2012). We note that an averaged iterate $\overline{g}_t$ can be obtained in an iterative fashion as follows: $\overline{g}_1 = g_1$ and

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

Moreover, since this update does not require storing all internal iterates $(g_t)_{t=1}^{T+1}$, it is more efficient than taking the average of them as described in Algorithm 1.

4 Analyses

To ensure the exponential convergence of these methods, we make several assumptions and provide several notations. Recall that $\phi : \mathbb{R} \to \mathbb{R}$ is a non-negative convex function to define a loss function l. We define the "link function" h_* from $(0, 1)$ to $\mathbb{R}$ as follows if it is well defined; for $\forall \mu \in (0, 1)$,

$$h_*(\mu) \overset{def}{=} \arg\min_{h \in \mathbb{R}} \{\mu\phi(h) + (1 - \mu)\phi(-h)\}$$

and denote by $l_*(\mu)$ a corresponding value:

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

The link function h_* is well-defined for several loss functions; e.g., for logistic loss, it becomes $h_*(\mu) = \log(\mu/(1 - \mu))$. Since l_* is concave (Zhang, 2004), the Bregman divergence for l_* can be defined by

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

Let g_* be Bayes rule for $\mathcal{L}$ that minimizes the $\mathcal{L}$ over all measurable functions.

Assumption 1.

(A1) *ϕ (and also $l(\cdot, y)$) is differentiable and convex. There exists $M > 0$ such that $|\partial_\zeta l(\zeta, y)| \leq M$. $\mathcal{L}(g)$ is L-smooth, that is, there exists $L > 0$ such that for $\forall g, \forall h \in \mathcal{H}_k$,*

$$|\mathcal{L}(g+h) - \mathcal{L}(g) - \langle \nabla \mathcal{L}(g), h \rangle_{\mathcal{H}_k}| \leq \frac{L}{2}\|h\|^2_{\mathcal{H}_k}.$$

(A2) *Assume $\mathrm{supp}(\rho_\mathcal{X}) = \mathcal{X}$ and there exists $R > 0$ such that $k(x, x) \leq R^2$ for $\forall x \in \mathcal{X}$.*

(A3) *The strong low-noise condition holds; $\exists \delta \in (0, 1/2)$ such that $|\rho(1|x) - \frac{1}{2}| > \delta$, $\rho_\mathcal{X}$-almost surely.*

(A4) *$\rho(1|X)$ takes values in $(0, 1)$, $\rho_\mathcal{X}$-almost surely. h_* is well-defined, differentiable, monotonically increasing, and invertible over $(0, 1)$. Moreover, it follows that*

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

(A5) *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 $\mathcal{H}_k$.*

Assumption **(A1)** is common in the literature and valid for several loss functions, for instance, the logistic loss and the smoothed hinge loss. The boundedness of kernel function **(A2)** is also reasonable. Indeed, Gaussian kernel is bounded by 1 and continuous kernel functions are bounded when $\mathcal{X}$ is compact. This boundedness leads to an important relationship between norms $\|\cdot\|_{\mathcal{H}_k}$ and $\|\cdot\|_{L_\infty}$ (the sup norm over $\mathcal{X}$) as follows. Since $g(x) = \langle g, k(x, \cdot)\rangle_{\mathcal{H}_k}$ for arbitrary function $g \in \mathcal{H}_k$ and $k(x, \cdot) \in \mathcal{H}_k$ by the definition of kernel function, we get

$$\|g\|_{L_\infty} = \sup_{x \in \mathcal{X}} |g(x)| \leq \|g\|_{\mathcal{H}_k} \|k(x, \cdot)\|_{\mathcal{H}_k} \leq R\|g\|_{\mathcal{H}_k}.$$

The strong low-noise condition assumed in **(A3)** is also adopted in Koltchinskii and Beznosova (2005); Audibert and Tsybakov (2007) to show the exponential convergence property of empirical risk minimizers for regularized problems. More recently, Pillaud-Vivien et al. (2017) exhibited exponential convergence of stochastic gradient descent for solving regularized least-squares regression for classification problems by using this condition. We also note that this condition is the strongest version of more general low-noise conditions used in Tsybakov (2004); Bartlett et al. (2006), which also gives a faster convergence rate of generalization error of the empirical risk minimizer than $O(1/\sqrt{n})$, where n is the number of training data, but an exponential convergence is not achievable. For logistic loss, since $h_*(\mu) = \log(\mu/(1-\mu))$ as introduced above, conditions in **(A4)** are satisfied. In this setting, the Bregman divergence d_{l_*} corresponds to the Kullback-Leibler divergence, which is positive. It is known from Zhang (2004) that the excess risk (that is the difference between the expected risk and the minimum expected risk over all measurable functions) can be measured by Bregman divergence d_{l_*} when ϕ, h_* are differentiable and h_* is invertible:

$$\mathcal{L}(g) - \mathcal{L}(g_*) = \mathbb{E}_X[d_{l_*}(h_*^{-1}(g(X)), \rho(1|X))]. \quad (3)$$

Therefore, if d_{l_*} is positive, then Bayes rule $g_*(X)$ equals $h_*(\rho(1|X))$, $\rho_\mathcal{X}$-almost surely. Thus, the uniqueness of the Bayes rule for $\mathcal{L}$ assumed in **(A5)** is also verified for logistic loss. Although we here focus on the logistic loss, Assumption **(A1)**, **(A4)**, and **(A5)** are valid for other loss functions, such as squared loss and smoothed hinge loss. Furthermore, when imposing a bounded convex constraint on the domain of the problem and assuming $\mathrm{supp}(\rho_\mathcal{X})$ is bounded, Assumption **(A1)** can be relaxed to capture a more comprehensive class of loss functions, including exponential loss, which also satisfies **(A4)** and **(A5)**. Even in this case, our analysis presented in the paper can be extended by considering projected stochastic gradient descent in an obvious way.

To describe our results, we introduce the following notation.

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

This provides a lower bound on $|g(X)|$, i.e. $|g(X)| \geq m(\delta)$ almost surely. For instance, for the logistic loss, $m(\delta) = \log((1+2\delta)/(1-2\delta))$ which converges to ∞ as $\delta \to 1/2$, resulting in better dependence on the low noise parameter in terms of the convergence rate. Note that if h_*^{-1} is L'-Lipschitz continuous, then $m(\delta) \geq \delta/L'$ for $\forall \delta \in (0, 1/2)$, e.g., $L' = 1/4$ for the logistic loss and $L' = 1/2$ for the squared loss.

4.1 Main Results

Here, we describe our main results where stochastic gradient descent and averaged stochastic gradient descent converge to the Bayes rule for the expected classification error with exponential convergence rates under sufficiently small $\lambda > 0$ guaranteeing sufficient closeness between g_λ (minimizer of $\mathcal{L}_\lambda$ in $\mathcal{H}_k$) and g_*. The existence of such a λ will be shown later under Assumptions **(A2)**–**(A5)**.

Theorem 1 is the main result for stochastic gradient descent. Since the rate of $O(1/\sqrt{T})$ is optimal without the (strong) low-noise condition, it is rather surprising that a significantly fast rate such as an exponential convergence can be achievable.

Theorem 1 (Exponential convergence rate for SGD). *Suppose Assumptions* **(A1)**–**(A5)** *hold. There exists a sufficiently small $\lambda > 0$ such that the following statement holds.*

Consider Algorithm 1 without the averaging option and with $\eta_t = 2/\lambda(\gamma+t)$, where γ is a positive value such that $\eta_1 \leq \min\{1/(L+\lambda), 1/2\lambda\}$. Let $\sigma^2 > 0$ be an upper-bound on the variance of stochastic gradient $\partial_\zeta l(g(X), Y)k(X, \cdot)$. We assume $\|g_1\|_{\mathcal{H}_k} \leq \left(2\eta_1 + \frac{1}{\lambda}\right) MR$. We set

$$\nu \overset{def}{=} \max\left\{\frac{2}{\lambda^2}(L+\lambda)\sigma^2, (1+\gamma)(\mathcal{L}_\lambda(g_1) - \mathcal{L}_\lambda(g_\lambda))\right\}.$$

Then, for $T \geq \frac{32R^2\nu}{m^2(\delta)\lambda} - \gamma$, we have

$$\mathbb{E}[\mathcal{R}(g_{T+1}) - \mathcal{R}([\mathbb{E}[Y|x]])] \leq 2\exp\left(-\frac{m^2(\delta)\lambda^2(\gamma+T)}{2^9 \cdot 9M^2R^4}\right) \tag{4}$$

Note that the bound (4) is valid only when T is larger than the threshold of the time given in the theorem. A similar threshold appeared in convergence results obtained in Pillaud-Vivien et al. (2017) with better dependence on δ, when $\delta \to 0$, than ours. However, we remark that our analysis generalizes their results to reasonable smooth convex loss functions which is more natural than the squared loss for classification problems. Moreover, since the low-noise parameter δ is a given fixed parameter, the dependency on δ is insignificant especially when $m(\delta)$ is rather large. For instance, recall that $m(\delta) \to \infty$ as $\delta \to 1/2$ for the logistic loss, resulting in the faster convergence rate. We moreover note that the threshold for the beginning time of exponential convergence is independent from a required precision for the excess classification error. Therefore, the number of iterations T needed to make the right hand side of (4) smaller than a given precision exceeds the threshold, if the precision is sufficiently small. Furthermore, even when T is smaller than the threshold, the convergence rate $O(1/T^q)$ ($q \in [0, 1]$) of the expected classification error can be obtained by the common analysis of the expected risk function and the consistency of the convex loss function.

We next give a main convergence result for averaged stochastic gradient descent which significantly reduces a time threshold compared to the vanilla stochastic gradient descent.

Theorem 2 (Exponential convergence rate for averaged SGD). *Suppose Assumptions* **(A1)–(A5)** *hold. There exists a sufficiently small $\lambda > 0$ such that the following statement holds. Consider Algorithm 1 with the averaging option, $\eta_t = 2/\lambda(\gamma+t)$, and $\alpha_t = 2(\gamma+t-1)/(2\gamma+T)(T+1)$, where γ is a positive value such that $\eta_1 \leq \min\{1/L, 1/2\lambda\}$. We assume $\|g_1\|_{\mathcal{H}_k} \leq \left(2\eta_1 + \frac{1}{\lambda}\right) MR$. Then, for sufficiently large T such that*

$$\max\left\{\frac{36M^2R^2}{\lambda^2(2\gamma+T)}, \frac{\gamma(\gamma-1)\|g_1 - g_\lambda\|^2_{\mathcal{H}_k}}{(2\gamma+T)(T+1)}\right\} \leq \frac{m^2(\delta)}{32R^2},$$

we have the following:

$$\mathbb{E}[\mathcal{R}(g_{T+1}) - \mathcal{R}([\mathbb{E}[Y|x]])] \leq 2\exp\left(-\frac{m^2(\delta)\lambda^2(2\gamma+T)}{2^{10} \cdot 9M^2R^4}\right), \tag{5}$$

Remark Although we assume **(A1)**, Lipschitz smoothness of $\mathcal{L}$ is not required in the proof of this theorem. That is, the parameter L does not affect the performance of averaged stochastic gradient descent.

We notice that two convergence rates (4) and (5) are comparable, but the threshold of averaged stochastic gradient descent for the beginning time of exponential convergence has much better dependence on λ than stochastic gradient descent, that is, the averaging technique accelerates the convergence in the early phase for small λ as shown in Rakhlin et al. (2012); Lacoste-Julien et al. (2012). As a result, threshold on T becomes not important. From the convergence rate (5), the required number of iterations to obtain ϵ-accuracy is

$$O\left(\frac{1}{m^2(\delta)\lambda^2}\log\left(\frac{1}{\epsilon}\right)\right). \tag{6}$$

We find clearly that this required iterations T naturally exceeds the time threshold for a rather small ϵ when we ignore the second term in the maximum in the threshold which is often inactive. In addition, this averaging scheme gives the sublinear convergence rate $O(1/T^q)$ ($q \in [0, 1]$), even when T is smaller than the threshold, as explained in the case of Theorem 1. Finally, we note that since $m(\delta) \geq \delta$ for the squared loss, a convergence rate (5) is exactly the same as that obtained in Pillaud-Vivien et al. (2017) when an additional assumption on the decreasing rate of eigenvalues of the covariance operator is not made. In other words, we succeed in generalizing the result in Pillaud-Vivien et al. (2017) without degradation under general settings.

As a corollary, we here derive a convergence rate for case of the logistic loss and Gaussian kernel, which can be obtained by setting $m(\delta) = \log((1+2\delta)/(1-2\delta))$ and $M, R = 1$.

Corollary 1. *Consider the logistic loss and Gaussian kernel. Suppose the same assumptions in Theorem 2 hold. Then, there exists a sufficiently small $\lambda > 0$ such that the following convergence rate of $\mathbb{E}[\mathcal{R}(g_{T+1}) - \mathcal{R}([\mathbb{E}[Y|x]])]$ holds.*

$$2\exp\left(-\frac{\lambda^2(2\gamma+T)}{2^{10} \cdot 9}\log^2\left(\frac{1+2\delta}{1-2\delta}\right)\right). \tag{7}$$

4.2 Proof Idea

We here explain the proof idea for convergence theorems. All missing proofs are found in the Appendix. The proof is mainly composed of three parts. We first show the Bayes optimality of g_λ for a small $\lambda > 0$ and specify the size of its neighborhood in $\mathcal{H}_k$ to ensure the optimality.

Proposition 1. *Suppose* **(A2)–(A5)** *in Assumption 1 hold. Then, there exists $\lambda > 0$ such that an arbitrary $g \in \mathcal{H}_k$ satisfying $\|g - g_\lambda\|_{\mathcal{H}_k} \leq m(\delta)/2R$ is the Bayes classifier of $\mathcal{R}(g)$. That is, $\mathcal{R}(\mathbb{E}[Y|x]) = \mathcal{R}(g)$.*

Remark This proposition shows the existence of λ to provide the Bayes classifier. In the expected risk minimization

problem, such an appropriate value of λ represents the inherent difficulty of the problem and that is controlled by the choice of kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. For the infinite dimensional problems, specifying the value of λ is somewhat difficult beforehand, indeed, it was not provided even for the squared loss (Pillaud-Vivien et al., 2017). However, we can specify λ for finite dimensional problems as follows. We assume that there exist positive values Δ, v_Δ such that $\mathcal{L}(g) \geq \mathcal{L}(g_*) + v_\Delta$ for arbitrary g satisfying $\|g - g_*\|_{\mathcal{H}_k} \geq \Delta$. This condition can be derived from the local strong convexity at g_* which is often assumed for the logistic loss (c.f. Bach and Moulines (2013)). Then, we can easily show that the minimizer g_λ satisfies $\|g_\lambda - g_*\|_{\mathcal{H}_k} < \Delta$ when $\|g_*\|_{\mathcal{H}_k} \leq 2v_\Delta/\lambda$. Therefore, λ should be chosen to satisfy the condition $\|g_*\|_{\mathcal{H}_k} \leq 2v_\Delta/\lambda$ for a target accuracy $\Delta = m(\delta)/2R$ as seen in the proof. In short, an appropriate λ depends on $\|g_*\|_{\mathcal{H}_k}$ and the local strong convexity v_Δ at g_*. As for the kernel k, it should be chosen for making them better conditioned under a bounded constraint $k(x, x) \leq R$. As a result, the required sample size (number of iterations) depends on $\|g_*\|_{\mathcal{H}_k}$ and the convexity v_Δ via the value of λ (cf. Theorem 1 and 2), but such a dependence is quite natural as seen in the theory of kernel ridge regression (Caponnetto and De Vito, 2007).

We notice that from Proposition 1, the goal of classification problems is achieved by finding a function included in the neighborhood of g_λ providing the Bayes rule for $\mathcal{R}$, of which the existence is shown in the proposition. Since g_λ is the minimizer of $\mathcal{L}_\lambda$ in $\mathcal{H}_k$, it is expected that a sequence of iterates obtained by a stochastic optimization method such as stochastic gradient descent to solve the problem converges to g_λ with high probability. To derive the probability and convergence rate to obtain the Bayes rule, we verify the convergence of an expected estimator and its variance. For the variance, we utilize the following proposition to bound it.

Proposition 2 (Pinelis (1994))**.** *Let $D_1, \ldots, D_T$ be a martingale difference sequence taking values in $\mathcal{H}_k$. We assume that there exists a constant $c_T > 0$ such that $\sum_{t=1}^T \|D_t\|_\infty^2 \leq c_T^2$, where $\|D_t\|_\infty$ is the essential supremum of $\|D_t\|_{\mathcal{H}_k}$. Then, for $\forall \epsilon > 0$,*

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

Let $\hat{g}_{T+1}$ stand for an output iterate of stochastic gradient descent or averaged stochastic gradient descent with T-iterations. Let $Z_1, \ldots, Z_T$ be i.i.d. random variables following ρ. Since $D_t = \mathbb{E}[\hat{g}_{T+1}|Z_1, \ldots, Z_t] - \mathbb{E}[\hat{g}_{T+1}|Z_1, \ldots, Z_{t-1}]$ ($t \in \{1, \ldots, T\}$) is a martingale difference sequence, Proposition 2 can be applied to bound the norm of the sum of D_t over $t \in \{1, \ldots, T\}$. Since $\sum_{t=1}^T D_t = \hat{g}_{T+1} - \mathbb{E}[\hat{g}_{T+1}]$, we see that for $\forall \delta > 0$, with the probability at least $1 - 2\exp(-m^2(\delta)/32R^2c_T^2)$,

$$\|\hat{g}_{T+1} - \mathbb{E}[\hat{g}_{T+1}]\|_{\mathcal{H}_k} < \frac{m(\delta)}{4R}. \tag{8}$$

Thus, by combining Proposition 1 and the inequality (8), we conclude that if an expected function $\mathbb{E}[\hat{g}_{T+1}]$ is in the neighborhood of the radius $\delta/4L'R$ around g_λ, then $\hat{g}_{T+1}$ is the Bayes optimal for $\mathcal{R}$ with the probability at least $1 - 2\exp(-m^2(\delta)/32R^2c_T^2)$. That is,

$$\mathcal{R}(\hat{g}_{T+1}) = \mathcal{R}(\mathbb{E}[Y|x]).$$

In other words, from the definition of the expected classification error $\mathcal{R}$, if $\|\mathbb{E}[\hat{g}_{T+1}] - g_\lambda\|_{\mathcal{H}_k} < m(\delta)/4R$, then

$$\mathbb{E}[\mathcal{R}(\hat{g}_{T+1}) - \mathcal{R}(\mathbb{E}[Y|x])] \leq 2\exp\left(-\frac{m^2(\delta)}{32R^2c_T^2}\right). \tag{9}$$

Therefore, by confirming the convergence of $\mathbb{E}[\hat{g}_{T+1}]$ to g_λ and specifying the convergence rate $O(1/T^q)$ ($q > 0$) of c_T to zero, we can conclude the exponential convergence of the expected classification error from the inequality (9). Thus, the remaining problem is to verify these convergences for Algorithm 1. As for the expected iterate $\mathbb{E}[\hat{g}_{T+1}]$, its convergence can be shown by naturally extending proofs (Bottou et al., 2018; Lacoste-Julien et al., 2012) for stochastic gradient descent in Euclidean space. For c_T, we can show its convergence by utilizing an argument (Hardt et al., 2016) to show the stability of stochastic gradient descent for strongly convex problems. Such a combination of the martingale bound and the stability analysis of stochastic gradient descent has also been adopted in Liu et al. (2017) for another purpose.

Auxiliary results for main theorems We now exhibit auxiliary results for deriving the exponential convergence rate of stochastic gradient descent (Algorithm 1 without averaging). To do so, we here present convergence rates of quantities $\mathbb{E}[g_{T+1}]$ and c_T. The rate of the former is given in the following proposition.

Proposition 3. *Suppose Assumption* **(A1)** *holds. Consider Algorithm 1 without averaging and with the same learning rates as in Theorem 1. We assume that $\eta_1 \leq 1/(L + \lambda)$ and $\sigma^2 > 0$ is an upper-bound on the variance of stochastic gradient $\partial_\zeta l(g(X), Y)k(X, \cdot)$. We set*

$$\nu \stackrel{def}{=} \max\left\{\frac{2}{\lambda^2}(L + \lambda)\sigma^2, (1 + \gamma)(\mathcal{L}_\lambda(g_1) - \mathcal{L}_\lambda(g_\lambda))\right\}.$$

Then, we have

$$\|\mathbb{E}[g_T] - g_\lambda\|_{\mathcal{H}_k}^2 \leq \frac{2\nu}{\lambda(\gamma + T)}.$$

This convergence can be shown in a standard way in the stochastic optimization literature.

On the other hand, a bound for the value of c_T can be derived in the following manner. Let $Z'_t \sim \rho$ be a random variable independent from $Z_1, \ldots, Z_T$ and let g^t_{T+1} be an output of stochastic gradient descent (Algorithm 1 without averaging) depending on $(Z_1, \ldots, Z_{t-1}, Z'_t, Z_{t+1}, \ldots, Z_T)$. By setting $D_t = \mathbb{E}[g_{T+1}|Z_1, \ldots, Z_t] - \mathbb{E}[g_{T+1}|Z_1, \ldots, Z_{t-1}]$, we find

$$\|D_t\|_\infty \leq \mathbb{E}[\|g_{T+1} - g^t_{T+1}\|_\infty | Z_1, \ldots, Z_t],$$

where we recall that $\|\cdot\|_\infty$ is the essential supremum of $\|\cdot\|_{\mathcal{H}_k}$. Therefore, c_T can be estimated by bounding $\|g_{T+1} - g^t_{T+1}\|_\infty$ uniformly with respect to random variables. Such a bound would be obtained by the stability property (Hardt et al., 2016), that is, the small deviation that results from replacing one example for stochastic gradient descent. As a result, this argument leads to the following proposition:

Proposition 4. *Suppose Assumptions* **(A1)** *and* **(A2)** *hold. Consider Algorithm 1 without averaging and with the same learning rates as in Theorem 1. We assume that* $\eta_1 \leq \min\{1/L, 1/2\lambda\}$ *and* $\|g_1\|_{\mathcal{H}_k} \leq \left(2\eta_1 + \frac{1}{\lambda}\right) MR$. *Then, it follows that*

$$\sum_{t=1}^{T} \|D_t\|^2_\infty \leq \frac{144 M^2 R^2}{\lambda^2(\gamma + T)},$$

where D_t *is a martingale difference, as defined previously.*

By combining these two propositions in the way explained earlier, we can prove the exponential convergence (Theorem 1) of the expected classification error for stochastic gradient descent.

We can also show Theorem 2, i.e., the exponential convergence of averaged stochastic gradient descent by specifying the rate of $\mathbb{E}[\bar{g}_{T+1}]$ to g_λ and c_T to zero for the algorithm. Although the averaging method brings more preferable results as seen in Theorem 2, we defer auxiliary results for this theorem to the Appendix for simplicity.

Comparison with another concentration inequality We emphasize that our proof technique can provide a much faster convergence rate than that derived from another concentration inequality (Kakade and Tewari, 2009). Indeed, the following convergence rate of the objective gap was shown in Kakade and Tewari (2009) with probability at least $1 - \log(T)q$,

$$O\left(\frac{\log T}{\sqrt{T}}\right) + \frac{\sqrt{\log T}}{T}\sqrt{\log\left(\frac{1}{q}\right)} + \frac{\log(1/q)}{T}.$$

Therefore, q should be $\exp(-o(T))$ to guarantee the convergence. As a result, a convergence rate of expected classification error is $O(\log(T)\exp(-o(T)))$ which is much slower than our rates and a threshold on T also has a worse trade-off with respect to the choice of q.

5 Experiments

In this section, we conduct numerical experiments on synthetic datasets to verify our theoretical analyses. The random Fourier feature (Rahimi and Recht, 2007) is the most popular and widely used technique to approximate shift invariant kernels k with the dot-product: $\iota(x)^\top \iota(x')$ $(\forall x, \forall x' \in \mathcal{X})$ through a non-linear embedding ι from $\mathcal{X}$ to a low-dimensional Euclidean space $\mathbb{R}^D$. When we use such a kernel defined by ι, stochastic gradient descent and its averaging variant are reduced to those for linear models in an Euclidean space. Moreover, since we assume that the Bayes rule g_* is in $\mathcal{H}_k$, it is reasonable for the numerical verification to consider linear models and linear separable datasets in an Euclidean space in experiments.

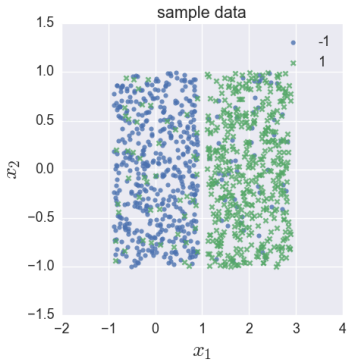

Figure 1: A subsample of the data used in the experiment with $\delta = 0.4$.

We here explain the experimental setting. For the loss function, we use logistic loss. For datasets, we use linear separable two dimensional synthetic datasets as shown in Figure 1 which is subsampled from a dataset. The support of datasets is composed of two part: $[-1 + r, 1 - r] \times [-1, 1]$ and $[1 + r, 3 - r] \times [-1, 1]$ in $\mathbb{R}^2$, where $r = 0.1$. We consider fixed conditional probabilities on each component, namely, for $\delta \in (0, 0.5)$, we use $\rho(Y = 1|x) = 0.5 - \delta$ for the left part and we use $\rho(Y = 1|x) = 0.5 + \delta$ for the right part of the support. As for the low-noise parameter δ, we test values from $\{0.1, 0.25, 0.4\}$ and as for the regularization parameter λ, we test values from $\{0.1, 0.01, 0.001, 0.0001\}$. Test datasets containing $100,000$ points are sampled from this distribution for each δ. We run stochastic gradient descent and averaged stochastic gradient descent 5-times with $20,000$-iterations and we report the best results, on training accuracies, with respect to λ for each δ. Before running these methods, we additionally use 1000-iterations for tuning hyperparameter γ which is an offset for time index as the optimization proceeds well. As for the regularization parameter λ, the value of 0.01 is chosen for $\delta = 0.1, 0.25$ and the value of 0.0001 is chosen for $\delta = 0.4$.

Experimental results are depicted in Figure 2. The top row shows mean curves of loss functions and the middle row shows mean curves of classification errors with standard

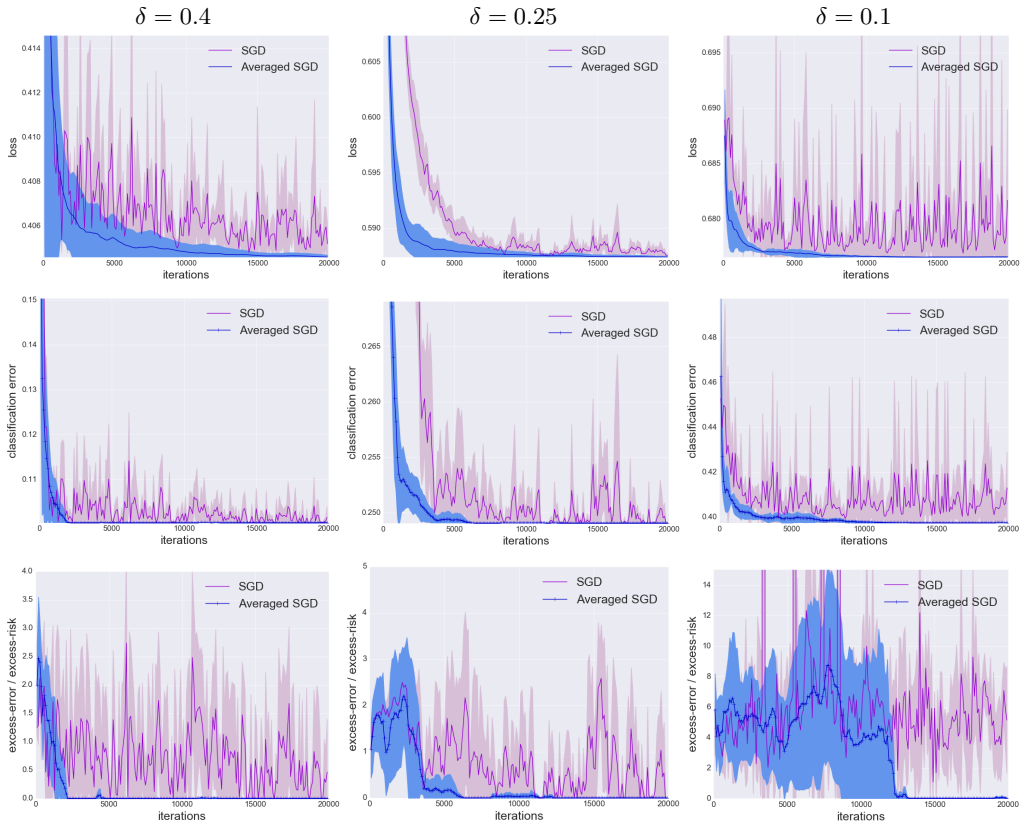

Figure 2: Learning curves by (averaged) stochastic gradient descent for the binary logistic regression. Figure depicts values of loss functions (top), classification errors (middle), and ratios (bottom): excess errors divided by excess risks, for each δ.

deviations obtained by stochastic gradient descent (purple line) and averaged stochastic gradient descent (blue line). As seen in Figure 2, the bigger low-noise parameter δ is, the faster the convergence speed of the classification error becomes. Especially, for the case $\delta = 0.4$ of the smallest noise, much faster convergence of the classification error than the loss is observed. Indeed, Bayes rule for $\mathcal{R}$ is achieved in earlier iterations. This phenomenon is also confirmed in the bottom row in Figure 2 that depicts curves of ratios of excess classification errors to excess risks, for each δ. The fast decreasing of these curves indicate the fast convergence of classification errors.

6 Conclusion

In this paper, we have shown the exponential convergence property of the expected classification error under a strong low-noise condition, rather than the expected risk for (averaged) stochastic gradient descent. The main contribution of this work, compared to existing work, is generalizing the loss function to more general differentiable loss functions, such as logistic loss, smoothed hinge loss, and exponential loss. As a result, the acceleration of the method has been shown for typical binary classification problems. Finally, our analysis has been verified experimentally. However, some problems are left for future work. First, we will investigate whether the class of loss functions can be further extended to non-differential functions such as the hinge loss function. The second is to exclude the assumption that the Bayes rule for the expected risk is included in the given hypothesis class. Finally, we will explore the convergence speed of more sophisticated methods, such as stochastic accelerated methods and stochastic variance reduced methods (Schmidt et al., 2017; Johnson and Zhang, 2013; Defazio et al., 2014; Nitanda, 2014; Allen-Zhu, 2017; Murata and Suzuki, 2017; Frostig et al., 2015) under the strong low-noise condition. Although these methods have been studied extensively for the empirical risk minimization problems, their performance for the expected risk and the expected classification error are still unclear.

Acknowledgement TS was partially supported by MEXT Kakenhi (26280009, 15H05707 and 18H03201), Japan Digital Design and JST-CREST.

References

Agarwal, A., Wainwright, M. J., Bartlett, P. L., and Ravikumar, P. K. (2009). Information-theoretic lower bounds

on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems 22*, pages 1–9.

Allen-Zhu, Z. (2017). Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of Annual ACM SIGACT Symposium on Theory of Computing 49*, pages 1200–1205. ACM.

Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633.

Bach, F. and Moulines, E. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems 24*, pages 451–459.

Bach, F. and Moulines, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems 26*, pages 773–781.

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.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357.

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

Cesa-Bianchi, N., Conconi, A., and Gentile, C. (2004). On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057.

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 27*, pages 1646–1654.

Frostig, R., Ge, R., Kakade, S. M., and Sidford, A. (2015). Competing with the empirical risk minimizer in a single pass. In *Proceedings of Conference on Learning Theory 28*, pages 728–763.

Ghadimi, S. and Lan, G. (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368.

Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning 33*, pages 1225–1234.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, pages 315–323.

Kakade, S. M. and Tewari, A. (2009). On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems 22*, pages 801–808.

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

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*.

Liu, T., Lugosi, G., Neu, G., and Tao, D. (2017). Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning 34*, pages 2159–2167.

Murata, T. and Suzuki, T. (2017). Doubly accelerated stochastic variance reduced dual averaging method for regularized empirical risk minimization. In *Advances in Neural Information Processing Systems 30*, pages 608–617.

Nemirovski, A. S., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

Nemirovskii, A. and Yudin, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. John Wiley.

Nitanda, A. (2014). Stochastic proximal gradient descent with acceleration techniques. In *Advances in Neural Information Processing Systems 27*, pages 1574–1582.

Pillaud-Vivien, L., Rudi, A., and Bach, F. (2017). Exponential convergence of testing error for stochastic gradient methods. *arXiv preprint arXiv:1712.04755*.

Pinelis, I. (1994). Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706.

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

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184.

Rakhlin, A., Shamir, O., and Sridharan, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of International Conference on Machine Learning 29*, pages 1571–1578.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Rosasco, L., Vito, E. D., Caponnetto, A., Piana, M., and Verri, A. (2004). Are loss functions all the same? *Neural Computation*, 16(5):1063–1076.

Ruppert, D. (1988). Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

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

Smale, S. and Yao, Y. (2006). Online learning algorithms. *Foundations of computational mathematics*, 6(2):145–170.

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

Ying, Y. and Zhou, D.-X. (2006). Online regularized classification algorithms. *IEEE Transactions on Information Theory*, 52(11):4775–4788.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex ris minimization. *The Annals of Statistics*, 32(1):56–134.