Journal of Machine Learning Research 22 (2021) 1-45 Submitted 5/19; Revised 12/20; Published 2/21

Stochastic Proximal AUC Maximization

Yunwen Lei Y.LEI@BHAM.AC.UK
School of Computer Science
University of Birmingham
Birmingham B15 2TT, United Kingdom

Yiming Ying* YYING@ALBANY.EDU
Department of Mathematics and Statistics
State University of New York at Albany
Albany, USA

Editor: Massimiliano Pontil

Abstract

In this paper we consider the problem of maximizing the Area under the ROC curve (AUC) which is a widely used performance metric in imbalanced classification and anomaly detection. Due to the pairwise nonlinearity of the objective function, classical SGD algorithms do not apply to the task of AUC maximization. We propose a novel stochastic proximal algorithm for AUC maximization which is scalable to large scale streaming data. Our algorithm can accommodate general penalty terms and is easy to implement with favorable $\mathcal{O}(d)$ space and per-iteration time complexities. We establish a high-probability convergence rate $\mathcal{O}(1/\sqrt{T})$ for the general convex setting, and improve it to a fast convergence rate $\mathcal{O}(1/T)$ for the cases of strongly convex regularizers and no regularization term (without strong convexity). Our proof does not need the uniform boundedness assumption on the loss function or the iterates which is more fidelity to the practice. Finally, we perform extensive experiments over various benchmark data sets from real-world application domains which show the superior performance of our algorithm over the existing AUC maximization algorithms.

Keywords: AUC maximization, imbalanced classification, stochastic gradient descent, proximal operator

1. Introduction

Area under the ROC curve (AUC) (Hanley and McNeil, 1982) measures the probability for a randomly drawn positive instance to have a higher decision value than a randomly sampled negative instance. It is a widely used metric for measuring the performance of machine learning algorithms in imbalanced classification and anomaly detection (Bradley, 1997; Cortes and Mohri, 2004; Fawcett, 2006; Narasimhan and Agarwal, 2017; Maurer and Pontil, 2020). In particular, minimization of the rank loss in bipartite ranking is equivalent to maximizing the AUC criterion (Agarwal et al., 2005; Güvenir and Kurtcephe, 2013; Kotlowski et al., 2011). At the same time, we are experiencing the fundamental change of the sheer size of commonly generated datasets where *streaming data* is continuously arriving in a real time manner. Hence, it is of practical importance to develop efficient optimization

∗. Corresponding author

algorithms for maximizing the AUC score which is scalable to large-scale streaming datasets for real-time predictions.

Stochastic (proximal) gradient descent (SGD), also known as stochastic approximation or incremental gradient, has become the workhorse in machine learning (Bach and Moulines, 2013; Bottou and Cun, 2004; Orabona, 2014; Rakhlin et al., 2012; Rosasco et al., 2014; Srebro and Tewari, 2010; Denevi et al., 2019). It can be regarded as online learning (Cesa-Bianchi and Lugosi, 2006; Hazan, 2016; Shalev-Shwartz, 2012; Orabona, 2019) in the stochastic setting where the individual data point is assumed to be drawn randomly from a (unknown) distribution. These algorithms are iterative and incremental in nature and process each new sample (input) with a computationally cheap update, making them amenable for streaming data analysis. The working mechanism behind classical SGD algorithms is to perform gradient descent using unbiased (random) samples of the true gradient. In the sense, the objective function is required to be *linear* in the sampling distribution. For example, in binary classification, let ρ be a probability measure (sampling distribution) defined on input/output space $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{\pm 1\}$. The linearity with respect to the sampling distribution ρ in this case means that the objective function (true risk) is the expectation of a *pointwise* loss function $\ell : \mathbb{R}^d \times \mathcal{X} \times \mathcal{Y} \to [0, \infty)$, i.e.

$$R(\mathbf{w}) = \mathbb{E}[\ell(\mathbf{w}, x, y)] = \iint_{\mathcal{X}\times\mathcal{Y}} \ell(\mathbf{w}, x, y) d\rho(x, y).$$

This linearity plays a pivotal role in studying the convergence of SGD and deriving many of its appealing properties.

In contrast, the problem of AUC maximization involves the expectation of a *pairwise* loss function which depends on pairs of data points. Consequently, the objective function in AUC maximization is *pairwise nonlinear* with respect to the sampling distribution ρ. To be more precise, recall (Hanley and McNeil, 1982; Clémençon et al., 2008) that the AUC score of a function $h_{\mathbf{w}}(x) = \mathbf{w}^\top x$ is defined by

$$\mathrm{AUC}(\mathbf{w}) = \Pr\{\mathbf{w}^\top x \geq \mathbf{w}^\top x' | y = +1, y' = -1\} = \mathbb{E}\big[\mathbb{I}_{[\mathbf{w}^\top x \geq \mathbf{w}^\top x']} | y = +1, y' = -1\big], \quad (1.1)$$

where $\mathbb{E}[\cdot]$ is with respect to (x, y) and (x', y') independently drawn from ρ. Since the indicator function $\mathbb{I}[\cdot]$ is discontinuous, one often resorts to a convex surrogate loss $\ell : \mathbb{R} \mapsto \mathbb{R}^+$ and some common choices are the square loss $\ell(a) = (1 - a)^2$, the hinge loss $\ell(a) = \max\{0, 1 - a\}$ (Zhao et al., 2011) and the exponential loss $\ell(a) = \exp(-a)$ (Rudin and Schapire, 2009). In this paper, we consider the square loss since it is statistically consistent with AUC (Gao and Zhou, 2015) and its specific structure allows us to reformulate the pairwise learning problem as a pointwise learning problem (Ying et al., 2016b; Liu et al., 2018; Natole et al., 2018). As a comparison, AUC maximization based on other loss functions requires to compare a pair of examples in updating models (Zhao et al., 2011; Kar et al., 2013; Wang et al., 2012b), which causes expensive space and per-iteration complexity. Then, we have

$$\begin{aligned} p(1-p)\Big[1 - \mathrm{AUC}(\mathbf{w})\Big] \leq f(\mathbf{w}) :=& p(1-p)\mathbb{E}[(1 - \mathbf{w}^\top(x - x'))^2 | y = 1, y' = -1] \\ =& \mathbb{E}\big[(1 - \mathbf{w}^\top(x - x'))^2 \mathbb{I}_{[y=1, y'=-1]}\big], \end{aligned} \quad (1.2)$$

where $p = \Pr(y = 1)$. As in Ying et al. (2016b); Natole et al. (2018); Liu et al. (2018), the idea of introducing the factor $p(1-p)$ is to replace the conditional expectation in the AUC score with the expectation, which is more convenient to deal with. Now the regularization framework for maximizing the AUC score can be formulated as follows

$$\min_{\mathbf{w}\in\mathbb{R}^d} \Big\{\phi(\mathbf{w}) := f(\mathbf{w}) + \Omega(\mathbf{w})\Big\}, \tag{1.3}$$

where $\Omega : \mathbb{R}^d \mapsto \mathbb{R}^+$ is a convex regularizer. This pairwise nonlinearity in the sampling distribution makes the direct deployment of standard SGD infeasible.

1.1 Related Work

There are considerable efforts on developing optimization algorithms for AUC maximization, which can roughly be divided into three categories.

The first category is batch learning algorithms for AUC maximization with focus on the empirical risk minimization (Cortes and Mohri, 2004) which use the training data at once. For instance, the early work (Joachims, 2005; Herschtal and Raskutti, 2004) proposed to use the cutting plane method and gradient descent algorithm, respectively. Zhang et al. (2012) developed an appealing algorithmic framework for optimizing the multivariate performance measures (Joachims, 2005) including the AUC score and precision-recall break-even point. The algorithms there used the smoothing techniques (Nesterov, 2007) and the Nesterov's accelerated gradient algorithm (Nesterov, 1983). Support Vector Algorithms were proposed to maximize the partial area under the ROC curve between any two false positive rates, which is interesting in several applications, e.g., ranking, biometric screening and medicine (Narasimhan and Agarwal, 2017). Such batch learning algorithms generally require $\mathcal{O}\big(\min\big(\frac{1}{\varepsilon}, \frac{1}{\sqrt{\lambda\varepsilon}}\big)\big)$ iterations to achieve an accuracy of ϵ, but have a high per-iteration cost of $\mathcal{O}(nd)$. Here, λ, n, and d are the regularization parameter, the number of samples, and the dimension of the data, respectively. Such algorithms train the model on the whole training data which are not suitable for analyzing massive streaming data that arrives continuously.

The second category of work (Kar et al., 2013; Wang et al., 2012b; Ying and Zhou, 2016) extended the classical online gradient descent (OGD) (Zinkevich, 2003; Hazan, 2016; Shalev-Shwartz, 2012) to the setting of pairwise learning and hence is applicable to the problem of AUC maximization. Regret bounds were established there which can be converted to generalization bounds in the stochastic setting as shown by Kar et al. (2013); Wang et al. (2012b). Such algorithms, however, need to compare the latest arriving data with previous data which require to store the historic data. This leads to expensive space and per-iteration complexities $\mathcal{O}(td)$ at the t-th iteration which is not feasible for streaming data. For the specific square loss, Gao et al. (2013) developed an one-pass AUC maximization method by updating the covariance matrices of the training data, which has $\mathcal{O}(d^2)$ space and per-iteration time complexity which could be problematic for high-dimensional data.

The third category of work (Ying et al., 2016a; Liu et al., 2018; Natole et al., 2018) considered the expected risk and used primal-dual SGD algorithms. In particular, Ying et al. (2016a); Natole et al. (2018) formulated AUC maximization (1.3) as a saddle point problem as follows

$$\min_{\mathbf{w},a,b\in\mathbb{R}} \max_{\alpha\in\mathbb{R}} \mathbb{E}_z\big[F(\mathbf{w}, a, b, \alpha; z)\big] + \Omega(\mathbf{w}), \tag{1.4}$$

where $F(\mathbf{w}, a, b, \alpha; z) = p(1-p) + (1-p)(\mathbf{w}^\top x - a)^2 \mathbb{I}_{[y=1]} + p(\mathbf{w}^\top x - b)^2 \mathbb{I}_{[y=-1]} + 2(1+\alpha)\mathbf{w}^\top x \big(p\mathbb{I}_{[y=-1]} - (1-p)\mathbb{I}_{[y=1]}\big) - p(1-p)\alpha^2$. Then, they proposed to perform SGD on both the primal variables $\mathbf{w}, a$ and b, and the dual variable α. This algorithm has per-iteration and space cost of $\mathcal{O}(d)$, making them amenable for streaming data analysis. It enjoys a moderate convergence rate $\mathcal{O}(1/\sqrt{T})$. The most recent work by Liu et al. (2018) also used this saddle point formulation and developed a novel multi-stage scheme for running primal-dual stochastic gradient algorithms which enjoy a fast convergence of $\widetilde{\mathcal{O}}(1/T)$[1] for non-strongly-convex objective functions. Both algorithms in Ying et al. (2016a); Liu et al. (2018) require a critical assumption of uniform boundedness for model parameters. i.e. $\|\mathbf{w}\| \leq R$ which might be difficult to adjust in practice. Natole et al. (2018) developed a stochastic proximal algorithm for AUC maximization with a convergence rate $\widetilde{\mathcal{O}}(1/T)$ for strongly convex objective function. The potential limitation of this method is that it assumes the conditional expectations $\mathbb{E}[x'|y'=1]$ and $\mathbb{E}[x'|y'=-1]$ are known a priori which is hard to satisfy in practice.

There are some other related work. For instance, Palaniappan and Bach (2016) developed an appealing stochastic primal-dual algorithm for saddle point problems with convergence rate of $\mathcal{O}(\frac{1}{T})$ which, as a by-product, can be applied to AUC maximization with the square loss. However, their saddle point formulation focused on the empirical risk minimization and cannot be applied to the population risk in our case. In addition, the primal-dual algorithm there requires strong convexity on both the primal and dual variables, and the algorithm has per-iteration complexity $\mathcal{O}(n+d)$ where n is the total number of training samples and d is the dimension of the data.

Our work falls in the regime where the aim is to minimize an expected-valued objective function which is nonlinear with respect to the sampling distribution. This research area is attracting more and more attention in optimization and machine learning with important applications to reinforcement learning and robust learning. For example, Wang et al. (2016, 2017) proposed a stochastic compositional gradient descent (SCGD) for solving the problem

$$\min_{\mathbf{w}\in\Omega} \mathbb{E}[f_v(\mathbb{E}(g_w(\mathbf{w})|\mathbf{v}))], \tag{1.5}$$

where Ω is a closed convex set of $\mathbb{R}^n$, $f_v : \mathbb{R}^m \mapsto \mathbb{R}$ and $g_w : \mathbb{R}^n \mapsto \mathbb{R}^m$ are functions parametrized by the random variables w and v. However, it is not clear how to formulate the problem of AUC maximization as (1.5). In addition, the SCGD algorithms proposed in (Wang et al., 2017, 2016) require that both the gradients of f_v and g_w are bounded which is not the case for our setting since we use the square loss. As we show soon in the next section, we explore the intrinsic structure of AUC maximization to show our proposed algorithms are guaranteed to converge with high probability without boundedness assumptions. Moreover, it can achieve a fast convergence rate of $\widetilde{\mathcal{O}}(\frac{1}{T})$ without strong convexity.

1.2 Main Contributions

In this paper, we propose novel SGD algorithms for AUC maximization which does not need the boundedness assumptions and can achieve a fast convergence rate without strong convexity. Our key idea is the new decomposition technique (see Proposition 1) which

1. We use the notation $\widetilde{\mathcal{O}}$ to hide polynomial of logarithms.

Algorithm	storage/per-iteration	bound type	rate	penalty
OAM (Zhao et al., 2011)	$\mathcal{O}(Bd)$	regret	$\mathcal{O}(1/\sqrt{T})$	ℓ_2
OPAUC (Gao et al., 2013)	$\mathcal{O}(d^2)$	regret	$\mathcal{O}(1/\sqrt{T})$	ℓ_2
SOLAM (Ying et al., 2016a)	$\mathcal{O}(d)$	w.h.p.	$\mathcal{O}(1/\sqrt{T})$	ℓ_2- constraint
FSAUC (Liu et al., 2018)	$\mathcal{O}(d)$	w.h.p.	$\widetilde{\mathcal{O}}(1/T)$	ℓ_1- constraint
SPAM (Natole et al., 2018)	$\mathcal{O}(d)$	expectation	$\widetilde{\mathcal{O}}(1/T)$	strongly convex
SPAUC (this work)	$\mathcal{O}(d)$	w.h.p.	$\widetilde{\mathcal{O}}(1/\sqrt{T})$	convex regularizer
SPAUC (this work)	$\mathcal{O}(d)$	w.h.p.	$\widetilde{\mathcal{O}}(1/T)$	strongly convex or no regularizer

Table 1: Comparison of different AUC maximization methods. The notation B refers to the buffer size in Zhao et al. (2011). For the bound type, "regret" refers to regret bounds, "expectation" refers to convergence rates in expectation and "w.h.p." refers to convergence rates with high probability. If the bound type is "regret", we use the rate $\mathcal{O}(1/\sqrt{T})$ to mean regret bounds $\mathcal{O}(\sqrt{T})$ for a consistent comparison.

directly works with the objective function motivated by the saddle point formulation (Ying et al., 2016a; Natole et al., 2018). From this new decomposition, we are able to design approximately unbiased estimators for the true gradient $\nabla f(\mathbf{w})$. Our algorithms do not need to store the previous data points in contrast to the approaches in (Wang et al., 2012b; Kar et al., 2013; Zhao et al., 2011) or accessing true conditional expectations as in (Natole et al., 2018). A comparison of our algorithm with other methods is summarized in Table 1.

From the side of technical novelty, we develop techniques to control the norm of iterates with high probability for stochastic learning based on biased estimators of $\nabla f(\mathbf{w})$, and hence there is no boundedness assumptions on the iterates. Our major contributions can be summarized as follows.

- We propose a novel stochastic proximal algorithm for AUC maximization which accommodates general convex regularizers with favorable $\mathcal{O}(d)$ space and per-iteration time complexities. Our algorithm is gradient-based and hence is simple and easy to implement which does not need the multi-stage design (Liu et al., 2018) and bounded assumption on model parameters (Liu et al., 2018; Ying et al., 2016a).

- We establish a convergence rate $\widetilde{O}(1/\sqrt{T})$ with high probability for our algorithm with T iterations, and improve it to a fast convergence $\widetilde{\mathcal{O}}(1/T)$ for both cases of no regularization term (non-strong convexity) and strongly convex regularizers.

- We perform a comprehensive empirical comparison against five state-of-the-art AUC maximization algorithms over sixteen benchmark data sets from real-world application domains. Experimental results show that our algorithm can achieve superior performance with a consistent and significant reduction in running time.

Organization of the paper. The remainder of this paper is organized as follows. We state the algorithm with motivation in Section 2. Theoretical and experimental results are presented

in Section 3 and Section 4, respectively. We give some proofs in Section 5 and defer others to the Appendix. We conclude the paper in 6.

2. Proposed Algorithm

Our objective is to develop efficient SGD-type algorithms for AUC maximization scalable to large scale streaming data. In particular, we aim to design an (approximately) unbiased estimator for the true gradient $\nabla f(\mathbf{w})$ with per-iteration cost $\mathcal{O}(d)$ to perform SGD-type algorithms. In particular, our new design is mainly motivated by the saddle point formulation in (Ying et al., 2016a; Natole et al., 2018).

To illustrate the main idea, let

$$\begin{aligned}\widetilde{F}(\mathbf{w};z) = p(1-p) + (1-p)\big(\mathbf{w}^\top\big(x - \mathbb{E}[x'|y'=1]\big)\big)^2\mathbb{I}_{[y=1]} \\ + p\big(\mathbf{w}^\top\big(x - \mathbb{E}[x'|y'=-1]\big)\big)^2\mathbb{I}_{[y=-1]} + 2p(1-p)\mathbf{w}^\top\big(\mathbb{E}[x'|y'=-1] - \mathbb{E}[x'|y'=1]\big) \\ + p(1-p)\big(\mathbf{w}^\top\big(\mathbb{E}[x'|y'=-1] - \mathbb{E}[x'|y'=1]\big)\big)^2. \quad (2.1)\end{aligned}$$

We will give an intuitive explanation of $\widetilde{F}$ in Remark 2. It was shown in Ying et al. (2016a); Natole et al. (2018) that the saddle point formulation (1.4) implies that $f(\mathbf{w}) = \min_{\mathbf{w},a,b}\max_\alpha \mathbb{E}[F(\mathbf{w},a,b,\alpha;z)]$. In particular, for any fixed $\mathbf{w}$ the optima a, b, α have a closed-form solution of $a(\mathbf{w}), b(\mathbf{w})$ and $\alpha(\mathbf{w})$ which are given by

$$a(\mathbf{w}) = \mathbf{w}^\top\mathbb{E}[x'|y'=1],\ b(\mathbf{w}) = \mathbf{w}^\top\mathbb{E}[x'|y'=-1],\ \alpha(\mathbf{w}) = b(\mathbf{w}) - a(\mathbf{w}). \quad (2.2)$$

Indeed, let $F_1(\mathbf{w};z) = F(\mathbf{w}, a(\mathbf{w}), b(\mathbf{w}), \alpha(\mathbf{w}); z)$ and then

$$\begin{aligned}F_1(\mathbf{w};z) = (1-p)\big(\mathbf{w}^\top\big(x - \mathbb{E}[x'|y'=1]\big)\big)^2\mathbb{I}_{[y=1]} + p\big(\mathbf{w}^\top\big(x - \mathbb{E}[x'|y'=-1]\big)\big)^2\mathbb{I}_{[y=-1]} \\ + 2\big(1 + \mathbf{w}^\top\big(\mathbb{E}[x'|y'=-1] - \mathbb{E}[x'|y'=1]\big)\big)\mathbf{w}^\top x\big(p\mathbb{I}_{[y=-1]} - (1-p)\mathbb{I}_{[y=1]}\big) \\ + p(1-p) - p(1-p)\big(\mathbf{w}^\top\big(\mathbb{E}[x'|y'=-1] - \mathbb{E}[x'|y'=1]\big)\big)^2.\end{aligned}$$

Note that $\mathbf{w}^\top\mathbb{E}\big[x\big(p\mathbb{I}_{[y=-1]} - (1-p)\mathbb{I}_{[y=1]}\big)\big] = p(1-p)\mathbf{w}^\top\big(\mathbb{E}[x'|y'=-1] - \mathbb{E}[x'|y'=1]\big)$. After organizing the terms, one can easily see that $\mathbb{E}[\widetilde{F}(\mathbf{w};z)] = \mathbb{E}[F_1(\mathbf{w};z)] = f(\mathbf{w})$ for any $\mathbf{w}$. Consequently, one can see that both $\nabla\widetilde{F}(\mathbf{w};z)$ and $\nabla F_1(\mathbf{w},z)$ are both unbiased estimators of $\nabla f(\mathbf{w})$, i.e. $\mathbb{E}[\nabla F_1(\mathbf{w};z)] = \mathbb{E}[\nabla\widetilde{F}(\mathbf{w};z)] = \nabla f(\mathbf{w})$. The work of Natole et al. (2018) proposed to use $\nabla F(\mathbf{w}, a(\mathbf{w}), b(\mathbf{w}), \alpha(\mathbf{w}); z)$ as an unbiased gradient and the convergence analysis was proved in expectation. It is easy to see that $F_1(\mathbf{w};z)$ is not convex, i.e. the Hessian of $F_1(\mathbf{w};z)$ is not positive-semi-definite (PSD). The non-convexity of $F_1(\mathbf{w};z)$ presents daunting difficulties to bound the iterates and deriving the convergence of the algorithm in high-probability using concentration inequalities. In contrast, the new design of $\widetilde{F}(\mathbf{w};z)$ is convex with respect to $\mathbf{w}$ which will enable us to prove convergence in high probability.

In a nutshell, we have the following important proposition. Motivated by the saddle-point formulation in Ying et al. (2016a); Natole et al. (2018) as mentioned above, we also

give an alternative but self-contained proof by writing the objective function as

$$\begin{aligned}(1-\mathbf{w}^\top(x-x'))^2 &= \big([1+\alpha(\mathbf{w})]+[\mathbf{w}^\top x'-b(\mathbf{w})]-[\mathbf{w}^\top x-a(\mathbf{w})]\big)^2\\ &= \big(\big[1+\mathbf{w}^\top(\mathbb{E}[\tilde{x}|\tilde{y}=-1]-\mathbb{E}[\tilde{x}|\tilde{y}=1])\big]\\ &\quad+\big[\mathbf{w}^\top(x'-\mathbb{E}[\tilde{x}|\tilde{y}=-1])-\mathbf{w}^\top(x-\mathbb{E}[\tilde{x}|\tilde{y}=1])\big]\big)^2. \end{aligned}\tag{2.3}$$

Proposition 1 *For any* $\mathbf{w}$*, we have*

$$\mathbb{E}\big[\widetilde{F}(\mathbf{w};z)\big]=f(\mathbf{w})\quad \textit{and}\quad \mathbb{E}\big[\widetilde{F}'(\mathbf{w};z)\big]=\nabla f(\mathbf{w}),\tag{2.4}$$

where we use the abbreviation $\widetilde{F}'(\mathbf{w};z):=\frac{\partial\widetilde{F}(\mathbf{w};z)}{\partial\mathbf{w}}$*. Furthermore, for any* z *the function* $\widetilde{F}(\mathbf{w};z)$ *is a convex function of* $\mathbf{w}$*.*

Proof As indicated by (2.3), we write $(1-\mathbf{w}^\top(x-x'))^2$ as three terms:

$$\begin{aligned}&\Big(\big[1+\mathbf{w}^\top(\mathbb{E}[\tilde{x}|\tilde{y}=-1]-\mathbb{E}[\tilde{x}|\tilde{y}=1])\big]+\big[\mathbf{w}^\top(x'-\mathbb{E}[\tilde{x}|\tilde{y}=-1])-\mathbf{w}^\top(x-\mathbb{E}[\tilde{x}|\tilde{y}=1])\big]\Big)^2\\ &=\big\{\big[1+\mathbf{w}^\top(\mathbb{E}[\tilde{x}|\tilde{y}=-1]-\mathbb{E}[\tilde{x}|\tilde{y}=1])\big]^2\big\}+\big\{\big[\mathbf{w}^\top(x'-\mathbb{E}[\tilde{x}|\tilde{y}=-1])-\mathbf{w}^\top(x-\mathbb{E}[\tilde{x}|\tilde{y}=1])\big]^2\big\}\\ &\quad+\big\{2\big[1+\mathbf{w}^\top(\mathbb{E}[\tilde{x}|\tilde{y}=-1]-\mathbb{E}[\tilde{x}|\tilde{y}=1])\big]\big[\mathbf{w}^\top(x'-\mathbb{E}[\tilde{x}|\tilde{y}=-1])-\mathbf{w}^\top(x-\mathbb{E}[\tilde{x}|\tilde{y}=1])\big]\big\}\\ &=\mathbf{I}+\mathbf{II}+\mathbf{III}.\end{aligned}$$

It suffices to estimate the above terms one by one. To this end, the first term is deterministic, and hence

$$\begin{aligned}\mathbb{E}[\,\mathbf{I}\,|y=1,y'=-1]&=2\mathbf{w}^\top\big(\mathbb{E}[\tilde{x}|\tilde{y}=-1]-\mathbb{E}[\tilde{x}|\tilde{y}=1]\big)\\ &\quad+\big(\mathbf{w}^\top\big(\mathbb{E}[\tilde{x}|\tilde{y}=-1]-\mathbb{E}[\tilde{x}|\tilde{y}=1]\big)\big)^2+1.\end{aligned}\tag{2.5}$$

For the second term, noticing that $\mathbb{E}\big[\mathbf{w}^\top(x'-\mathbb{E}[\tilde{x}|\tilde{y}=-1])\mathbf{w}^\top(x-\mathbb{E}[\tilde{x}|\tilde{y}=1])\,|y=1,y'=-1\big]=\mathbb{E}\big[\mathbf{w}^\top(x'-\mathbb{E}[\tilde{x}|\tilde{y}=-1])\,|y'=-1\big]\mathbb{E}\big[\mathbf{w}^\top(x-\mathbb{E}[\tilde{x}|\tilde{y}=1])\,|y=1\big]=0$ as (x,y) and (x',y') are independent, we have

$$\begin{aligned}&\mathbb{E}[\,\mathbf{II}\,|y=1,y'=-1]\\ &=\mathbb{E}\big[\big(\mathbf{w}^\top(x'-\mathbb{E}[\tilde{x}|\tilde{y}=-1])-\mathbf{w}^\top(x-\mathbb{E}[\tilde{x}|\tilde{y}=1])\big)^2|y=1,y'=-1\big]\\ &=\mathbb{E}\big[\big(\mathbf{w}^\top(x'-\mathbb{E}[\tilde{x}|\tilde{y}=-1])\big)^2|y=1,y'=-1\big]+\mathbb{E}\big[\big(\mathbf{w}^\top(x-\mathbb{E}[\tilde{x}|\tilde{y}=1])\big)^2|y=1,y'=-1\big]\\ &=\mathbb{E}\big[\big(\mathbf{w}^\top(x'-\mathbb{E}[\tilde{x}|\tilde{y}=-1])\big)^2|y'=-1\big]+\mathbb{E}\big[\big(\mathbf{w}^\top(x-\mathbb{E}[\tilde{x}|\tilde{y}=1])\big)^2|y=1\big]\\ &=\frac{1}{1-p}\mathbb{E}\big[\big(\mathbf{w}^\top(x'-\mathbb{E}[\tilde{x}|\tilde{y}=-1])\big)^2\mathbb{I}_{[y'=-1]}\big]+\frac{1}{p}\mathbb{E}\big[\big(\mathbf{w}^\top(x-\mathbb{E}[\tilde{x}|\tilde{y}=1])\big)^2\mathbb{I}_{[y=1]}\big].\end{aligned}\tag{2.6}$$

For the third term,

$$\begin{aligned}\mathbb{E}[\,\mathbf{III}\,|y=1,y'=-1]&=2\big[1+\mathbf{w}^\top(\mathbb{E}[\tilde{x}|\tilde{y}=-1]-\mathbb{E}[\tilde{x}|\tilde{y}=1])\big]\\ &\quad\times\mathbb{E}\big[\mathbf{w}^\top(x'-\mathbb{E}[\tilde{x}|\tilde{y}=-1])-\mathbf{w}^\top(x-\mathbb{E}[\tilde{x}|\tilde{y}=1])|y=1,y'=-1\big]=0.\end{aligned}\tag{2.7}$$

Combining equations (2.5),(2.6), and (2.7) together, we have

$$\begin{aligned}
f(\mathbf{w}) &= p(1-p)\mathbb{E}[(1-\mathbf{w}^\top(x-x'))^2|y=1,y'=-1] \\
&= 2p(1-p)\mathbf{w}^\top\big(\mathbb{E}[\tilde{x}|\tilde{y}=-1]-\mathbb{E}[\tilde{x}|\tilde{y}=1]\big) \\
&+ p(1-p)\big(\mathbf{w}^\top\big(\mathbb{E}[\tilde{x}|\tilde{y}=-1]-\mathbb{E}[\tilde{x}|\tilde{y}=1]\big)\big)^2 + p(1-p) \\
&+ p\mathbb{E}\Big[\big(\mathbf{w}^\top(x'-\mathbb{E}[\tilde{x}|\tilde{y}=-1])\big)^2\mathbb{I}_{[y'=-1]}\Big] + (1-p)\mathbb{E}\Big[\big(\mathbf{w}^\top(x-\mathbb{E}[\tilde{x}|\tilde{y}=1])\big)^2\mathbb{I}_{[y=1]}\Big],
\end{aligned}$$

which implies that $f(\mathbf{w}) = \mathbb{E}[\widetilde{F}(\mathbf{w};z)]$.

The fact of $\mathbb{E}\big[\widetilde{F}'(\mathbf{w};z)\big] = \nabla f(\mathbf{w})$ follows directly from the Leibniz's integral rule that the derivative and the integral can be interchangeable as F is a quadratic function and the input x is from a bounded domain.

For the last statement, notice that

$$\begin{aligned}
\nabla^2\widetilde{F}(\mathbf{w};z) = {} & 2(1-p)\big(x-\mathbb{E}[\tilde{x}|\tilde{y}=1]\big)\big(x-\mathbb{E}[\tilde{x}|\tilde{y}=1]\big)^\top\mathbb{I}_{[y=1]} \\
& + 2p(x-\mathbb{E}[\tilde{x}|\tilde{y}=-1])(x-\mathbb{E}[\tilde{x}|\tilde{y}=-1])^\top\mathbb{I}_{[y=-1]} \\
& + 2p(1-p)\big(\mathbb{E}[\tilde{x}|\tilde{y}=-1]-\mathbb{E}[\tilde{x}|\tilde{y}=1]\big)\big(\mathbb{E}[\tilde{x}|\tilde{y}=-1]-\mathbb{E}[\tilde{x}|\tilde{y}=1]\big)^\top.
\end{aligned}$$

It is clear that $\nabla^2\widetilde{F}(\mathbf{w};z)$ is positive semi-definite, and hence $\widetilde{F}(\mathbf{w};z)$ is a convex function of $\mathbf{w}$ for any z. This completes the proof of the proposition. ■

Remark 2 *We summarize the key idea in the proof. Let* $\widetilde{\mathbb{E}}[\cdot] := \mathbb{E}[\cdot|y=1,y'=-1]$. *Then the proof essentially shows*

$$\begin{aligned}
\widetilde{\mathbb{E}}\big[\big(1-\mathbf{w}^\top(x-x')\big)^2\big] &= \widetilde{\mathbb{E}}\Big[\Big(\big(1-\mathbf{w}^\top\widetilde{\mathbb{E}}[x-x']\big)+\mathbf{w}^\top(\widetilde{\mathbb{E}}[x]-x)+\mathbf{w}^\top(x'-\widetilde{\mathbb{E}}[x'])\Big)^2\Big] \\
&= \widetilde{\mathbb{E}}\Big[\big(1-\mathbf{w}^\top\widetilde{\mathbb{E}}[x-x']\big)^2\Big] + \widetilde{\mathbb{E}}\Big[\big(\mathbf{w}^\top(\widetilde{\mathbb{E}}[x]-x)\big)^2\Big] + \widetilde{\mathbb{E}}\Big[\big(\mathbf{w}^\top(\widetilde{\mathbb{E}}[x']-x')\big)^2\Big]. \quad (2.8)
\end{aligned}$$

Note that $\widetilde{\mathbb{E}}\Big[\big(\mathbf{w}^\top(\widetilde{\mathbb{E}}[x]-x)\big)^2\Big]$ *and* $\widetilde{\mathbb{E}}\Big[\big(\mathbf{w}^\top(\widetilde{\mathbb{E}}[x']-x')\big)^2\Big]$ *are conditional variance of* $\mathbf{w}^\top x$ *in the positive class and negative class, respectively. Therefore, the AUC score can be considered as the summation of two conditional variances and a term depending on* $\mathbb{E}[x|y=1]-\mathbb{E}[x'|y'=-1]$. *This formulation motivates us to introduce* $\widetilde{F}$ *in* (2.1). *Indeed, it is clear that*

$$\begin{aligned}
\frac{1}{p(1-p)}\widetilde{F}(\mathbf{w};z) = {} & \big(1-\mathbf{w}^\top\big(\mathbb{E}[x'|y'=1]-\mathbb{E}[x'|y'=-1]\big)\big)^2 \\
& + \frac{1}{p}\big(\mathbf{w}^\top\big(x-\mathbb{E}[x'|y'=1]\big)\big)^2\mathbb{I}_{[y=1]} + \frac{1}{1-p}\big(\mathbf{w}^\top\big(x-\mathbb{E}[x'|y'=-1]\big)\big)^2\mathbb{I}_{[y=-1]}.
\end{aligned}$$

Note that $\frac{1}{p}\big(\mathbf{w}^\top\big(x-\mathbb{E}[x'|y'=1]\big)\big)^2\mathbb{I}_{[y=1]}$ *is an unbiased estimator of the conditional variance* $\mathbb{E}[(\mathbf{w}^\top(x-\mathbb{E}[x|y=1]))^2|y=1]$, *and* $\frac{1}{1-p}\big(\mathbf{w}^\top\big(x-\mathbb{E}[x'|y'=-1]\big)\big)^2\mathbb{I}_{[y=-1]}$ *is an unbiased estimator of the conditional variance* $\mathbb{E}[(\mathbf{w}^\top(x-\mathbb{E}[x|y=-1]))^2|y=-1]$. *Therefore,* $\widetilde{F}$ *is derived from* (2.8) *by replacing the conditional variances with their unbiased estimators.*

Proposition 1 indicates to use $\widetilde{F}'(\mathbf{w}; z)$ as an unbiased estimator for the gradient $\nabla f(\mathbf{w})$. However, the function $\widetilde{F}$ requires the unknown information $p, \mathbb{E}[x'|y'=1]$ and $\mathbb{E}[x'|y'=-1]$, which is unknown in practice. We propose to replace them by their empirical counterpart defined as follows

$$p_t = \frac{\sum_{i=0}^{t-1} \mathbb{I}_{[y_i=1]}}{t}, \quad u_t = \frac{\sum_{i=0}^{t-1} x_i \mathbb{I}_{[y_i=1]}}{\sum_{i=0}^{t-1} \mathbb{I}_{[y_i=1]}}, \quad v_t = \frac{\sum_{i=0}^{t-1} x_i \mathbb{I}_{[y_i=-1]}}{\sum_{i=0}^{t-1} \mathbb{I}_{[y_i=-1]}}, \tag{2.9}$$

where we reserve an example (x_0, y_0) drawn independently from ρ. The resulting estimator for F at time t then becomes

$$\begin{aligned}\hat{F}_t(\mathbf{w}; z) = (1-p_t)\big(\mathbf{w}^\top\big(x-u_t\big)\big)^2\mathbb{I}_{[y=1]} + p_t\big(\mathbf{w}^\top\big(x-v_t\big)\big)^2\mathbb{I}_{[y=-1]} \\ + 2p_t(1-p_t)\mathbf{w}^\top\big(v_t-u_t\big) + p_t(1-p_t)\big(\mathbf{w}^\top\big(v_t-u_t\big)\big)^2 + p_t(1-p_t).\end{aligned}$$

It is easy to verify by computing its Hessian that $\hat{F}_t(\mathbf{w}; z)$ is convex with respect to $\mathbf{w}$. Its gradient can be directly computed as follows

$$\begin{aligned}\hat{F}_t'(\mathbf{w}; z) := \frac{\partial \hat{F}_t(\mathbf{w}; z)}{\partial \mathbf{w}} = 2(1-p_t)\big(x-u_t\big)\big(x-u_t\big)^\top\mathbf{w}\mathbb{I}_{[y=1]} + 2p_t\big(x-v_t\big)\big(x-v_t\big)^\top\mathbf{w}\mathbb{I}_{[y=-1]} \\ + 2p_t(1-p_t)\big(v_t-u_t\big) + 2p_t(1-p_t)\big(v_t-u_t\big)\big(v_t-u_t\big)^\top\mathbf{w}.\end{aligned} \tag{2.10}$$

Note the stochastic gradient $\hat{F}_t'(\mathbf{w}; z)$ can be efficiently computed with an arithmetic cost $\mathcal{O}(d)$ and we do not need to store covariance matrices.

Algorithm 1: Stochastic Proximal AUC Maximization (SPAUC)

Input: $\{\eta_t\}_t, \Omega, \mathbf{w}_1$ and T.
Output: an approximate solution of (1.3)

initialize: $n_+ \leftarrow 0, n_- \leftarrow 0, s_+ \leftarrow 0, s_- \leftarrow 0$
for $t = 1, 2, \ldots, T$ **do**
  $n_+ \leftarrow n_+ + \mathbb{I}_{[y_t=1]}$
  $n_- \leftarrow n_- + \mathbb{I}_{[y_t=-1]}$
  $s_+ \leftarrow s_+ + x_t\mathbb{I}_{[y_t=1]}$
  $s_- \leftarrow s_- + x_t\mathbb{I}_{[y_t=-1]}$
  $u_t \leftarrow \frac{s_+}{n_+}, v_t \leftarrow \frac{s_-}{n_-}$
  calculate the stochastic gradient $\hat{F}_t'(\mathbf{w}; z_t)$ according to (2.10)
  update $\mathbf{w}_{t+1}$ according to (2.11)

Algorithm: We propose to solve this regularization problem (1.3) by the following *Stochastic proximal AUC maximization* (SPAUC) algorithm with $\mathbf{w}_1 = 0$ and for any $t \geq 1$,

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \eta_t\big\langle \mathbf{w} - \mathbf{w}_t, \hat{F}_t'(\mathbf{w}_t; z_t)\big\rangle + \eta_t\Omega(\mathbf{w}) + \frac{1}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2, \tag{2.11}$$

where $\{\eta_t\}_t$ is a sequence of positive step sizes and z_t is drawn independently from ρ at the t-th iteration. At the t-th iteration, SPAUC builds a temporary objective function

consisting of three components: a first order approximation of $f(\mathbf{w})$ based on the stochastic gradient $\hat{F}_t'(\mathbf{w}; z)$, a regularizer kept intact to preserve a composite structure and a term $\frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|_2^2$ to make sure the upcoming iterate $\mathbf{w}_{t+1}$ not far away from the current iterate. The pseudo-code of SPAUC is given in Algorithm 1.

We comment that Algorithm 1 differs from the standard stochastic proximal algorithm (Duchi and Singer, 2009; Rosasco et al., 2014; Lei and Tang, 2018) in that $\hat{F}_t'$ is a biased estimator of ∇f due to the use of p_t, u_t and v_t. High-probability convergence rates of standard stochastic proximal algorithms without boundedness assumptions were recently established in Lei and Tang (2018). We extend the analysis in Lei and Tang (2018) by building novel techniques to handle the bias of $\hat{F}_t'$. Although we only consider the development of AUC maximization here, the developed techniques may apply to general stochastic proximal algorithms based on approximately unbiased gradient estimators.

3. Main Convergence Results

In this section, we present theoretical convergence rates with high probability for SPAUC. We consider two types of objective functions of the form (1.3): AUC maximization with a convex ϕ and AUC maximization with ϕ satisfying a quadratic functional growth. Let $S^* = \{\mathbf{w} \in \mathbb{R}^d : \phi(\mathbf{w}) = \min_{\tilde{\mathbf{w}}} \phi(\tilde{\mathbf{w}})\}$ be the set of minimizers. For any $\mathbf{w}$, we denote by $\mathbf{w}^* = \arg\min_{\tilde{\mathbf{w}} \in S^*} \|\mathbf{w} - \tilde{\mathbf{w}}\|_2$ the projection of $\mathbf{w}$ on to S^*, where for any $p \geq 1$ and $\mathbf{w} = (w_1, \ldots, w_d)^\top$, we denote $\|\mathbf{w}\|_p = \left[\sum_{j=1}^d |w_j|^p\right]^{\frac{1}{p}}$. We always assume $\|\mathbf{w}_1^*\|_2 < \infty$.

3.1 General Convergence Rates

In this subsection, we present convergence rates for the general regularization framework for AUC maximization. To this aim, we need to impose a so-called self-bounding property on the regularizers, meaning the subgradients can be bounded in terms of function values. We denote by $\Omega'(\mathbf{w})$ a subgradient of Ω at $\mathbf{w}$ and assume $\Omega(0)$=0.

Assumption 1 (Self-bounding property) *There exist constants $A_1, A_2 \geq 0$ such that the convex regularizer Ω satisfies*

$$\|\Omega'(\mathbf{w})\|_2^2 \leq A_1\Omega(\mathbf{w}) + A_2, \quad \textit{for all } \mathbf{w} \in \mathbb{R}^d. \tag{3.1}$$

Based on A_1 and $\kappa := \max\{1, \sup_{x \in \mathcal{X}} \|x\|_2\}$, we introduce a constant

$$C_1 = \max\{A_1, 16\kappa^2\}.$$

This self-bounding assumption above is very mild as many regularizers satisfy self-bounding property, including all smooth regularizers and all Lipschitz regularizers. For example, if $\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_2^2$, then (3.1) holds with $A_1 = 4\lambda$ and $A_2 = 0$. If $\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$, then (3.1) holds with $A_1 = 0$ and $A_2 = \lambda^2$. It is reasonable to assume a small regularization parameters in practice (e.g., $\lambda \leq 1$), in which case we can take universal constants A_1 and A_2 for the above mentioned regularizers.

Our theoretical analysis requires to estimate $\|\mathbf{w}_t\|_2$, which is achieved by the following lemma to be proved in Section 5.3. Essentially, it shows that $\|\mathbf{w}_t\|_2$ is bounded (ignoring logarithmic factors) if we consider step sizes satisfying (3.2). This result shows that the complexity of $\mathbf{w}_t$ is well controlled even if the iterates are updated in an unbounded domain.

Theorem 3 *Let $\{\mathbf{w}_t\}_t$ be produced by (2.11) with $\eta_t \leq (2C_1)^{-1}$ and $\eta_{t+1} \leq \eta_t$. We suppose Assumptions 1 holds,*

$$\sum_{t=1}^{\infty} \eta_t \sqrt{\log t}/\sqrt{t} < \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty. \tag{3.2}$$

Then for any $\delta \in (0,1)$, there exists a constant C_2 independent of T (explicitly given in the proof and of the order of $\|\mathbf{w}_1^*\|_2^2$) such that the following inequality holds with probability at least $1-\delta$

$$\max_{1\leq \tilde{t}\leq T} \|\mathbf{w}_{\tilde{t}} - \mathbf{w}_1^*\|_2^2 \leq C_2 \log(2T/\delta). \tag{3.3}$$

We are now ready to present convergence rates for SPAUC applied to general AUC objectives. In Theorem 4 we present general convergence rates in terms of step sizes satisfying (3.2), which are then instantiated in Corollary 5 by specifying step sizes. The convergence rate $\mathcal{O}(T^{-\frac{1}{2}} \log^{\frac{3+\beta}{2}} \frac{T}{\delta})$ is optimal up to a logarithmic factor for stochastic algorithms applied to general convex optimization problems (Agarwal et al., 2009).

Theorem 4 *Let the conditions of Theorem 3 hold. Then, for any $\delta \in (0,1)$ there exists a constant C_3 independent of T such that the following inequality holds with probability at least $1-\delta$*

$$\phi(\bar{\mathbf{w}}_T^{(1)}) - \inf_{\mathbf{w}} \phi(\mathbf{w}) \leq C_3 \big(\sum_{t=1}^{T} \eta_t\big)^{-1} \max\Big\{\sum_{t=1}^{T} \eta_t^2, \sum_{t=1}^{T} \eta_t/\sqrt{t}\Big\} \log^{\frac{3}{2}} \frac{T}{\delta},$$

where $\bar{\mathbf{w}}_T^{(1)} = \sum_{t=1}^{T} \eta_t \mathbf{w}_t / \sum_{t=1}^{T} \eta_t$ is a weighted average of the first T iterates.

Corollary 5 *Let $\{\mathbf{w}_t\}_t$ be produced by (2.11) and $\delta \in (0,1)$. Suppose Assumptions 1 holds and $\eta_1 \leq (2C_1)^{-1}$.*

(1) If we choose $\eta_t = \eta_1 t^{-\theta}$ with $\theta > 1/2$, then with probability at least $1-\delta$ we have $\phi(\bar{\mathbf{w}}_T^{(1)}) - \inf_{\mathbf{w}} \phi(\mathbf{w}) = O\big(T^{\theta-1} \log^{\frac{3}{2}} \frac{T}{\delta}\big)$;

(2) If we choose $\eta_t = \eta_1 (t \log^{\beta}(et))^{-\frac{1}{2}}$ with $\beta > 2$, then with probability at least $1-\delta$ we have $\phi(\bar{\mathbf{w}}_T^{(1)}) - \inf_{\mathbf{w}} \phi(\mathbf{w}) = O\big(T^{-\frac{1}{2}} \log^{\frac{3+\beta}{2}} \frac{T}{\delta}\big)$.

The proofs for Theorem 4 and Corollary 5 can be found in Section 5.4.

3.2 Fast Convergence Rates

In this subsection, we show that a faster convergence rate is possible for SPAUC if a quadratic functional growth condition is imposed to the objective function (Anitescu, 2000; Necoara et al., 2018).

Assumption 2 (Quadratic functional growth) *We assume the existence of $\sigma_\phi > 0$ such that*

$$\phi(\mathbf{w}) - \phi(\mathbf{w}^*) \geq \sigma_\phi \|\mathbf{w} - \mathbf{w}^*\|_2^2, \quad \text{for all } \mathbf{w} \in \mathbb{R}^d. \tag{3.4}$$

The quadratic functional growth assumption (3.4) means that the objective function grows faster than the squared distance between any feasible point and the optimal set (Necoara et al., 2018). This condition is milder than assuming a strong convexity of ϕ (Necoara et al., 2018). Indeed, it holds if $\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_2^2$. It also holds if we consider no regularization, i.e., $\Omega(\mathbf{w}) = 0$ as shown in the next proposition. We give the proof for completeness.

Proposition 6 *The function $\phi(\mathbf{w}) = f(\mathbf{w})$ satisfies Assumption 2.*

Proof Indeed, the objective function can be written as

$$f(\mathbf{w}) = p(1-p)\big(\|A\mathbf{w}\|_2^2 + \mathbf{w}^\top c + 1\big)$$

with $A \in \mathbb{R}^{d\times d}$ being a symmetric matrix satisfying $A^2 = \mathbb{E}[(x-x')(x-x')^\top | y = 1, y' = -1]$ and $c = -2\mathbb{E}[x - x' | y = 1, y' = -1]$. Analyzing analogously to the proof of Theorem 9 in Necoara et al. (2018), one can show that $S^* = \big\{\mathbf{w} : A\mathbf{w} = g^*\big\}$ for some $g^* \in \mathbb{R}^d$. By the definition of $\mathbf{w}^*$ we know that $\mathbf{w} - \mathbf{w}^*$ is orthogonal to the kernel of A^2 and therefore $\lambda_{\min}(A^2)\|\mathbf{w} - \mathbf{w}^*\|_2^2 \le \|A\mathbf{w} - A\mathbf{w}^*\|_2^2$, where $\lambda_{\min}(A^2)$ denotes the smallest non-zero eigenvalue of A^2. Furthermore, we know

$$\begin{aligned} &p^{-1}(1-p)^{-1}\big(f(\mathbf{w}) - f(\mathbf{w}^*)\big) = \|A\mathbf{w}\|_2^2 - \|A\mathbf{w}^*\|_2^2 + (\mathbf{w} - \mathbf{w}^*)^\top c \\ &= \|A\mathbf{w} - A\mathbf{w}^*\|_2^2 + 2\big(A\mathbf{w} - A\mathbf{w}^*\big)^\top A\mathbf{w}^* + (\mathbf{w} - \mathbf{w}^*)^\top c = \|A\mathbf{w} - A\mathbf{w}^*\|_2^2, \end{aligned}$$

where the last identity is due to the optimality condition $2A^\top A\mathbf{w}^* + c = 0$. It then follows that $f(\mathbf{w}) - f(\mathbf{w}^*) \ge p(1-p)\lambda_{\min}(A^2)\|\mathbf{w} - \mathbf{w}^*\|_2^2$. The proof is complete. ■

Under Assumption 2, we show with high probability that the suboptimality measured by both the parameter distance and function values decay with the rate $\widetilde{\mathcal{O}}(T^{-1})$, which is optimal up to a logarithmic factor (Agarwal et al., 2009). Let $\sigma_\Omega \ge 0$ be a constant satisfying

$$\Omega(\mathbf{w}) - \Omega(\tilde{\mathbf{w}}) - \langle \mathbf{w} - \tilde{\mathbf{w}}, \Omega'(\tilde{\mathbf{w}})\rangle \ge 2^{-1}\sigma_\Omega\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2, \quad \forall \mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^d.$$

Note σ_Ω can be zero and therefore our results apply to non-strongly-convex regularizers, e.g., $\Omega(\mathbf{w}) = 0$ for all $\mathbf{w}$. Without loss of generality, we assume $\sigma_f := \sigma_\phi - \sigma_\Omega \ge 0$.

Theorem 7 *Let $\delta \in (0,1)$. Suppose Assumption 1 and Assumption 2 hold. Let $\{\mathbf{w}_t\}_t$ be the sequence produced by (2.11) with $\eta_t = \frac{2}{\sigma_\phi t + 2\sigma_f + \sigma_\phi t_1}$, where $t_1 \ge 32C_1\sigma_\phi^{-1}\log\frac{2T}{\delta}$. Then, the following inequality holds with probability at least $1-\delta$ for $t = 1, \ldots, T$ ($T > 2$)*

$$\|\mathbf{w}_t - \mathbf{w}_t^*\|_2^2 = \widetilde{\mathcal{O}}(1/t) \quad \textit{and} \quad \phi(\bar{\mathbf{w}}_t^{(2)}) - \inf_{\mathbf{w}} \phi(\mathbf{w}) = \widetilde{\mathcal{O}}(1/t), \tag{3.5}$$

where $\bar{\mathbf{w}}_t^{(2)}$ is a weighted average of iterates defined by

$$\bar{\mathbf{w}}_t^{(2)} = \big(\sum_{k=1}^t (k + t_1 + 1)\big)^{-1} \sum_{k=1}^t (k + t_1 + 1)\mathbf{w}_k.$$

The proof of Theorem 7 is postponed to Section C.

The following two corollaries follow directly from Theorem 7 by noting the quadratic functional growth property of the associated objective functions and the self-bounding property of the regularizers. We omit the proof here for brevity.

Corollary 8 *Let $\delta \in (0,1)$. Let $\{\mathbf{w}_t\}_t$ be produced by (2.11) with $\eta_t = \frac{2}{\sigma_\phi t + 2\sigma_f + \sigma_\phi t_1}$ and $\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_2^2/2$, where $t_1 \geq 32C_1\sigma_\phi^{-1}\log\frac{2T}{\delta}$. Then, (3.5) holds with probability $1-\delta$.*

Corollary 9 *Let $\delta \in (0,1)$. Let $\{\mathbf{w}_t\}_t$ be produced by (2.11) with $\eta_t = \frac{2}{\sigma_\phi t + 2\sigma_f + \sigma_\phi t_1}$ and $\Omega(\mathbf{w}) = 0$, where $t_1 \geq 32C_1\sigma_\phi^{-1}\log\frac{2T}{\delta}$. Then, (3.5) holds with probability $1-\delta$.*

Remark 10 *Our excess risk bounds are established for the square loss. There are some existing studies on connecting the bounds measured by convex surrogate and the bounds measured by AUC scores (Agarwal, 2013; Gao and Zhou, 2015). However, it is not clear to us how to combine these results to derive excess risk bounds in terms of AUC. We mention some difficulties in this direction as follows. First, the loss function $\widetilde{F}(\mathbf{w}; z)$ in (2.1) is defined on the model parameter $\mathbf{w}$ instead of the predicted output $\mathbf{w}^\top x$ (note there is $\mathbf{w}^\top\mathbb{E}[x'|y'=1]$ in $\widetilde{F}$), and therefore it is not clear to us how to define conditional risks as in Agarwal (2013); Gao and Zhou (2015). Furthermore, we only consider linear models and Proposition 1 only holds for linear models. It would be very interesting to study excess risk bounds for the AUC score.*

4. Experiments

In this section, we present experimental results to show the effectiveness of the proposed algorithm in achieving a satisfactory AUC with a fast convergence speed. We first describe the baseline methods used in our experimental comparison as well as the associated parameter setting in Section 4.1. Datasets used in the experiments and detailed experimental results are presented in Section 4.2 and Section 4.3, respectively.

4.1 Baseline Methods

We compare SPAUC to several state-of-the-art online AUC maximization algorithms. The algorithms we consider include

- the stochastic proximal AUC maximization (SPAUC) (2.11) with either no regularizers $\Omega(\mathbf{w}) = 0$, an ℓ_1 regularizer $\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$ or an ℓ_2 regularizer $\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_2^2$;
- the stochastic proximal AUC maximization (SPAM) (Natole et al., 2018) with $\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_2^2$;
- the stochastic online AUC maximization (SOLAM) (Ying et al., 2016a) based on a saddle problem formulation;
- the one-pass AUC maximization (OPAUC) (Gao et al., 2013) which uses the first and second-order statistics of training data to compute gradients;

- the online AUC maximization based on the hinge loss function (OAM_gra) (Zhao et al., 2011);
- the fast stochastic AUC maximization (FSAUC) (Liu et al., 2018) which applies a multi-stage stochastic optimization technique to a saddle problem formulation.

datasets	# inst	# feat	datasets	# inst	# feat	datasets	# inst	# feat	datasets	# inst	# feat
diabetes	768	8	ijcnn1	141691	22	german	1000	24	satimage	6435	36
acoustic	78823	50	covtype	581012	54	a9a	32561	123	connect	67557	126
usps	9298	256	w8a	49749	300	mnist	60000	780	gisette	7000	5000
real-sim	72309	20958	protein_h	145751	74	malware	71709	122	webspam_u	350000	254

Table 2: Description of the datasets used in the experiments.

The performance of these algorithms depends on some parameters, which, as described below, we tune with the five-fold cross-validation. For SPAUC, SPAM and SOLAM, we consider step sizes of the form $\eta_t = 2/(\mu t + 1)$ and validate the parameter μ over the interval $10^{\{-7,-6.5,\ldots,-2.5\}}$. Both SPAM and SPAUC with the ℓ_1/ℓ_2 regularizer require another regularization parameter to tune, which is validated over the interval $10^{\{-5,-4,\ldots,0\}}$. SOLAM involves the constraint on $\mathbf{w}$, i.e. $\mathbf{w}$ belonging to ℓ_2-ball with radius R in $\mathbb{R}^d$, for which we tune over the interval $10^{\{-1,0,\ldots,5\}}$. For OAM_gra, we need to tune a parameter to weight the comparison between released examples and bulk, which is validated over the interval $10^{\{-3,-2.5,\ldots,1.5\}}$. As recommended in Zhao et al. (2011), we fix the buffer size to 100. For OPAUC, we need to tune both the constant step size and the regularization parameter λ, which are validated over the interval $10^{\{-3.5,-3,\ldots,1\}}$ and $10^{\{-5,-4,\ldots,0\}}$, respectively. The multi-stage scheme in FSAUC specifies how the step size decreases along the implementation of the algorithm and leave the initial step size as a free parameter to tune, which we validate over the interval $10^{\{-2.5,-2,\ldots,2\}}$. Furthermore, each iteration of FSAUC requires a projection onto an ℓ_1-ball of radius of R, which we tune over the interval $10^{\{-1,0,\ldots,5\}}$. It should be noticed that both SPAUC with no regularizers and OAM_gra only have a single parameter to tune, while all other algorithms have two parameters to tune. To speed up the training process, if the algorithm has two parameters p_1, p_2 to tune, we first construct all the possible pairs (p_1, p_2) by enumerating all possible candidate values of p_1 and p_2, out of which we randomly sample 15 pairs without replacement to tune. We repeat the experiments 20 times and report the average of experimental results.

4.2 Datasets

We perform our experiments on several real-world datasets. We consider two types of datasets: the UCI benchmark dataset and the dataset in the domain of anomaly detection. The task of anomaly detection is to identify rare items, events or observations which raise suspicions by differing significantly from the majority of the data. As such, this is suitable to test the performance of AUC maximization methods since the class there is intrinsically and highly imbalanced. We consider three datasets in the domain of anomaly detection: protein_h, webspam_u and malware. In particular, webspam_u is a subset used in the Pascal Large Scale Learning Challenge (Wang et al., 2012a) to detect malicious web pages,

protein_h is a dataset in bioinformatics used to predict which proteins are homologous to a query sentence (non-homologous sequences are labeled as anomalies) (Caruana et al., 2004), and malware was collected in the Android Malware Genome Project used to detect mobile malware app (Jiang and Zhou, 2012). The remaining UCI datasets can be downloaded from the LIBSVM webpage (Chang and Lin, 2011). For each dataset, we use 80% of data for training and the remaining 20% for testing. We transform datasets with multiple class labels into datasets with binary class labels by grouping the first half of class labels into positive labels, and grouping the remaining class labels into negative labels. We run each algorithm until 15 passes of the training data are reached, and report the AUC values on the test dataset. The information of the dataset is summarized in Table 2 where we list the UCI datasets according to the dimensionality while datasets for anomaly detection are listed at the end.

4.3 Experimental Results

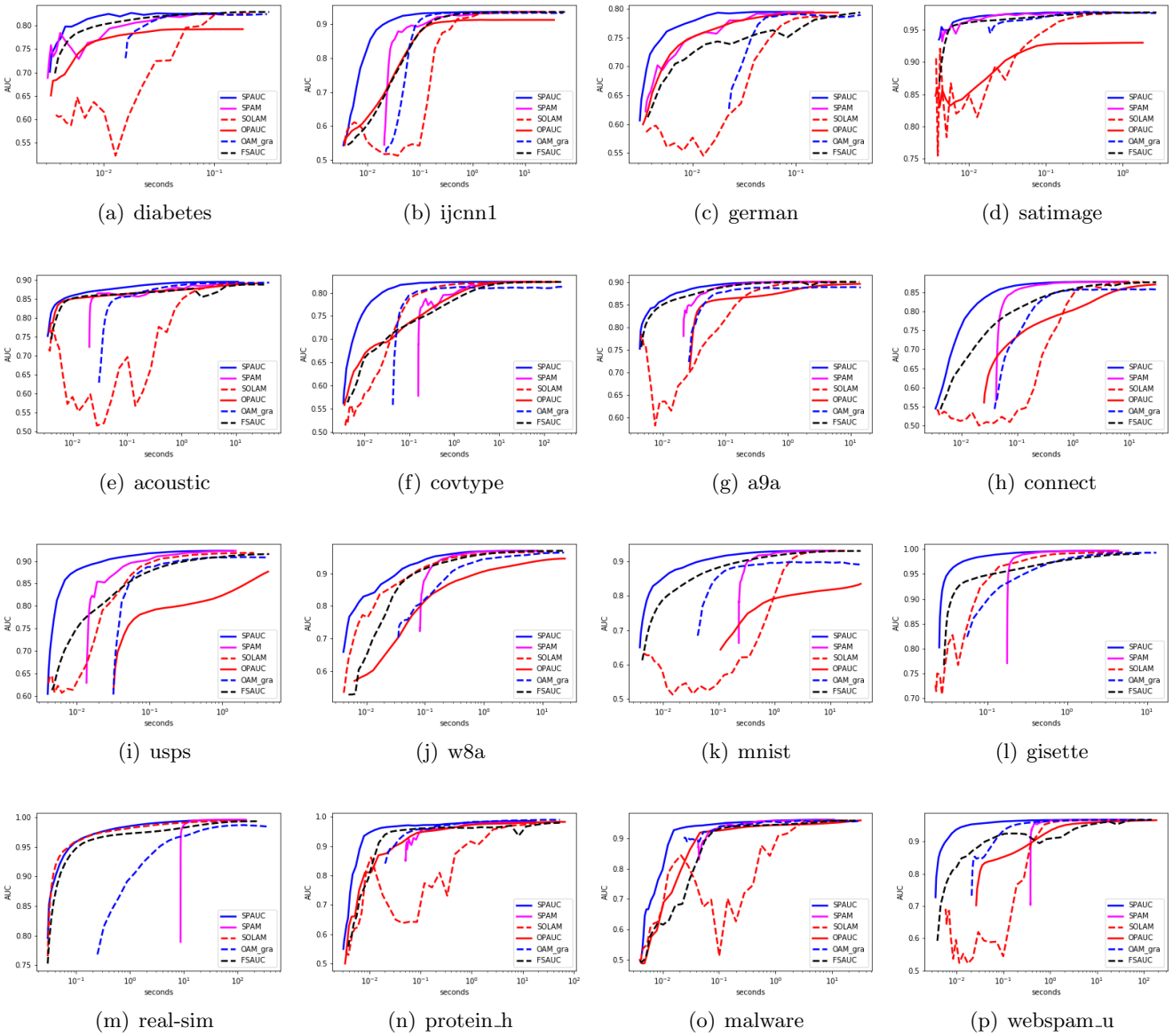
In this section, we present the experimental results and discuss the comparisons of our algorithm against other ones. In Figure 1, we plot the AUC values of the constructed models on the test data versus execution time in seconds for SPAUC (without regularization), SPAM, SOLAM, OPAUC, OAM_gra and FSAUC. It is observed that SPAUC attains a faster training speed than all baseline methods.

In particular, the curve of SOLAM fluctuates rapidly, especially in the early stage of the optimization, which is perhaps due to the requirement of updating both primal and dual variables. OAM_gra behaves more robustly, which, however, requires a high computation burden due to the requirement in updating a buffer and comparing the current example and examples in the buffer per iteration. As one can see from Figure 1, SPAUC converges faster than FSAUC on most of the datasets. The underlying reason could be two-fold. Firstly, FSAUC requires a projection onto the intersection of an ℓ_1-ball and ℓ_2-ball which requires an alternating projection step. Secondly, FSAUC requires to update both primal and dual variables, which further increases the computational cost per iteration. OPAUC has a low training speed due to the requirement in handling a covariance matrix, which is especially unfavorable for high-dimensional datasets. For example, OPAUC has the slowest training speed on USPS for which the dimensionality is 256. We do not run OPAUC on datasets with dimensionality larger than 1000 due to the heavy dependency of its time complexity on the dimensionality. The implementation of SPAM requires an accurate information of $p, \mathbb{E}[x'|y'=1]$ and $\mathbb{E}[x'|y'=-1]$, which we approximate with

$$\hat{p} = \frac{\sum_{i=1}^n \mathbb{I}_{[y_i=1]}}{n}, \quad \hat{u} = \frac{\sum_{i=1}^n x_i \mathbb{I}_{[y_i=1]}}{\sum_{i=1}^n \mathbb{I}_{[y_i=1]}}, \quad \hat{v} = \frac{\sum_{i=1}^n x_i \mathbb{I}_{[y_i=-1]}}{\sum_{i=1}^n \mathbb{I}_{[y_i=-1]}}. \tag{4.1}$$

It is observed that the AUC curve for SPAM attains a sharp increase at the beginning of the curve and then moderately increases. The underlying reason is that we include the computational cost of calculating $\hat{p}, \hat{u}$ and $\hat{v}$ in the curve, which requires to go through the whole training set.

In Table 3, we also report detailed AUCs as well as the execution time per pass, both in the form of mean $\pm$ standard deviation. We can see from Table 3 that SPAUC achieves accuracies comparable to the state-of-the-art methods over all datasets. SPAUC (without



(a) diabetes (b) ijcnn1 (c) german (d) satimage

(e) acoustic (f) covtype (g) a9a (h) connect

(i) usps (j) w8a (k) mnist (l) gisette

(m) real-sim (n) protein_h (o) malware (p) webspam_u

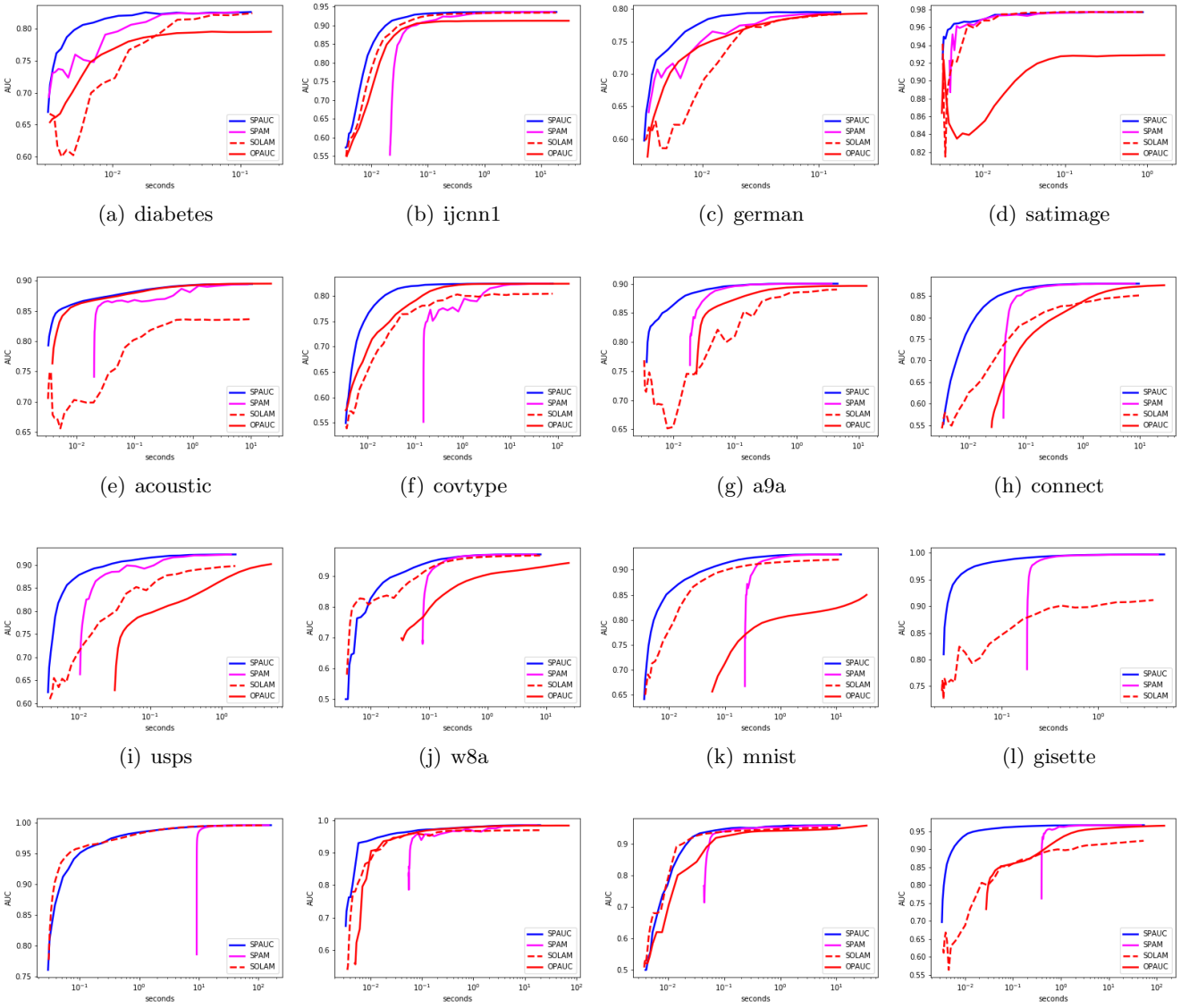
Figure 1: AUC versus time curves (in seconds) for SPAUC (without regularization), SPAM, SOLAM and OPAUC, OAM_gra and FSAUC.

regularization) and SPAM require comparable running time per iteration since both algorithms require no projections and no updates on the dual variables. An advantage of SPAUC with no regularization over SPAM is that SPAUC can deal with streaming data in a truly online fashion, while SPAM needs to know the conditional expectations in (4.1) and hence is not an online learning algorithm. Furthermore, the fast convergence of SPAM requires the objective function to be strongly convex (Natole et al., 2018), which introduces an additional regularization parameter to tune. Other baseline methods require longer per-

datasets		SPAUC	SPAM	SOLAM	OPAUC	OAM_gra	FSAUC
diabetes	AUC	0.8266±0.0284	0.8246±0.0303	0.8264±0.0308	0.7926±0.0462	0.8247±0.0266	0.8293±0.0375
	Time	0.0075±0.0013	0.0071±0.0002	0.0141±0.0015	0.0121±0.0011	0.0201±0.0014	0.0210±0.0011
ijcnn1	AUC	0.9361±0.0019	0.9358±0.0018	0.9362±0.0019	0.9127±0.0021	0.9331±0.0031	0.9361±0.0015
	Time	1.2881±0.0076	1.3080±0.0500	2.4498±0.0682	2.3807±0.0934	3.5811±0.1253	3.8352±0.0709
german	AUC	0.7938±0.0246	0.7943±0.0255	0.7879±0.0326	0.7932±0.0313	0.7890±0.0278	0.7933±0.0262
	Time	0.0094±0.0011	0.0099±0.0015	0.0179±0.0010	0.0169±0.0007	0.0281±0.0026	0.0278±0.0015
satimage	AUC	0.9772±0.0029	0.9769±0.0040	0.9765±0.0028	0.9300±0.0066	0.9760±0.0029	0.9770±0.0041
	Time	0.0609±0.0038	0.0589±0.0010	0.1181±0.0102	0.1212±0.0024	0.1802±0.0114	0.1826±0.0123
acoustic	AUC	0.8952±0.0026	0.8910±0.0028	0.8911±0.0032	0.8911±0.0026	0.8929±0.0028	0.8877±0.0076
	Time	0.7281±0.0216	0.7304±0.0278	1.3972±0.0168	1.6672±0.0213	2.7367±0.2275	2.2063±0.1009
covtype	AUC	0.8236±0.0009	0.8235±0.0009	0.8228±0.0013	0.8233±0.0009	0.8134±0.0036	0.8233±0.0007
	Time	5.4320±0.2447	5.4988±0.0629	10.5169±0.3023	13.1290±0.6334	19.303±3.6072	16.064±0.2350
a9a	AUC	0.9000±0.0033	0.9003±0.0042	0.9003±0.0033	0.8957±0.0028	0.8879±0.0043	0.9002±0.0031
	Time	0.3123±0.0018	0.3120±0.0023	0.5862±0.0035	0.9417±0.0610	0.9686±0.1143	0.9273±0.0146
connect	AUC	0.8786±0.0031	0.8783±0.0023	0.8783±0.0032	0.8716±0.0027	0.8583±0.0035	0.8776±0.0036
	Time	0.6520±0.0082	0.6532±0.0053	1.2386±0.0142	1.9633±0.0864	2.0464±0.2030	2.0307±0.0376
usps	AUC	0.9225±0.0048	0.9226±0.0046	0.9182±0.0065	0.8765±0.0105	0.9079±0.0069	0.9154±0.0050
	Time	0.0947±0.0044	0.1026±0.0022	0.1821±0.0093	0.2851±0.0282	0.2638±0.0195	0.2949±0.0045
w8a	AUC	0.9694±0.0035	0.9692±0.0040	0.9663±0.0041	0.9454±0.0057	0.9640±0.0044	0.9695±0.0036
	Time	0.5414±0.0069	0.5401±0.0262	0.9725±0.0119	1.6080±0.1558	1.5541±0.1302	1.5782±0.0228
mnist	AUC	0.9306±0.0020	0.9302±0.0017	0.9304±0.0027	0.8345±0.0086	0.8908±0.0047	0.9302±0.0015
	Time	0.8272±0.0241	0.8409±0.0168	1.3983±0.0592	2.3366±0.2359	2.2333±0.2533	2.3376±0.0418
gisette	AUC	0.9970±0.0011	0.9969±0.0011	0.9940±0.0014	-	0.9931±0.0017	0.9908±0.0024
	Time	0.2899±0.0224	0.2846±0.0208	0.3291±0.0253	-	0.8719±0.0807	0.5778±0.0423
real-sim	AUC	0.9955±0.0004	0.9959±0.0002	0.9936±0.0005	-	0.9842±0.0021	0.9934±0.0006
	Time	8.6884±0.2815	9.5146±0.3872	8.9692±0.3466	-	25.505±0.7707	16.132±0.4540
protein_h	AUC	0.9858±0.0029	0.9806±0.0030	0.9807±0.0054	0.9825±0.0040	0.9895±0.0017	0.9793±0.0036
	Time	1.1807±0.0293	1.1943±0.0296	2.2331±0.0853	4.4396±0.4447	3.5537±0.7592	3.5484±0.1478
malware	AUC	0.9606±0.0122	0.9595±0.0126	0.9589±0.0143	0.9587±0.0129	0.9566±0.0152	0.9581±0.0114
	Time	0.7291±0.0212	0.7296±0.0183	1.2822±0.0473	2.1483±0.2648	1.9903±0.1021	1.9604±0.0637
webspam_u	AUC	0.9673±0.0008	0.9664±0.0007	0.9668±0.0005	0.9659±0.0006	0.9670±0.0012	0.9671±0.0006
	Time	3.7163±0.2121	3.3759±0.1412	6.0562±0.3276	12.3849±0.9478	9.9146±0.1668	9.7933±0.5240

Table 3: Comparison of the testing AUC values and running time per pass (mean±std.).

pass running time due to the same reasons we mentioned above for explaining the AUC curve in Figure 1. It can be seen that OAM_gra requires longer per-pass running time than OPAUC if the dimensionality is relatively small, while the reverse is the case for datasets with a relatively large dimensionality. This is consistent with the dependency of the time complexity on the dimensionality for these two methods, i.e., linear versus quadratic.



(a) diabetes (b) ijcnn1 (c) german (d) satimage

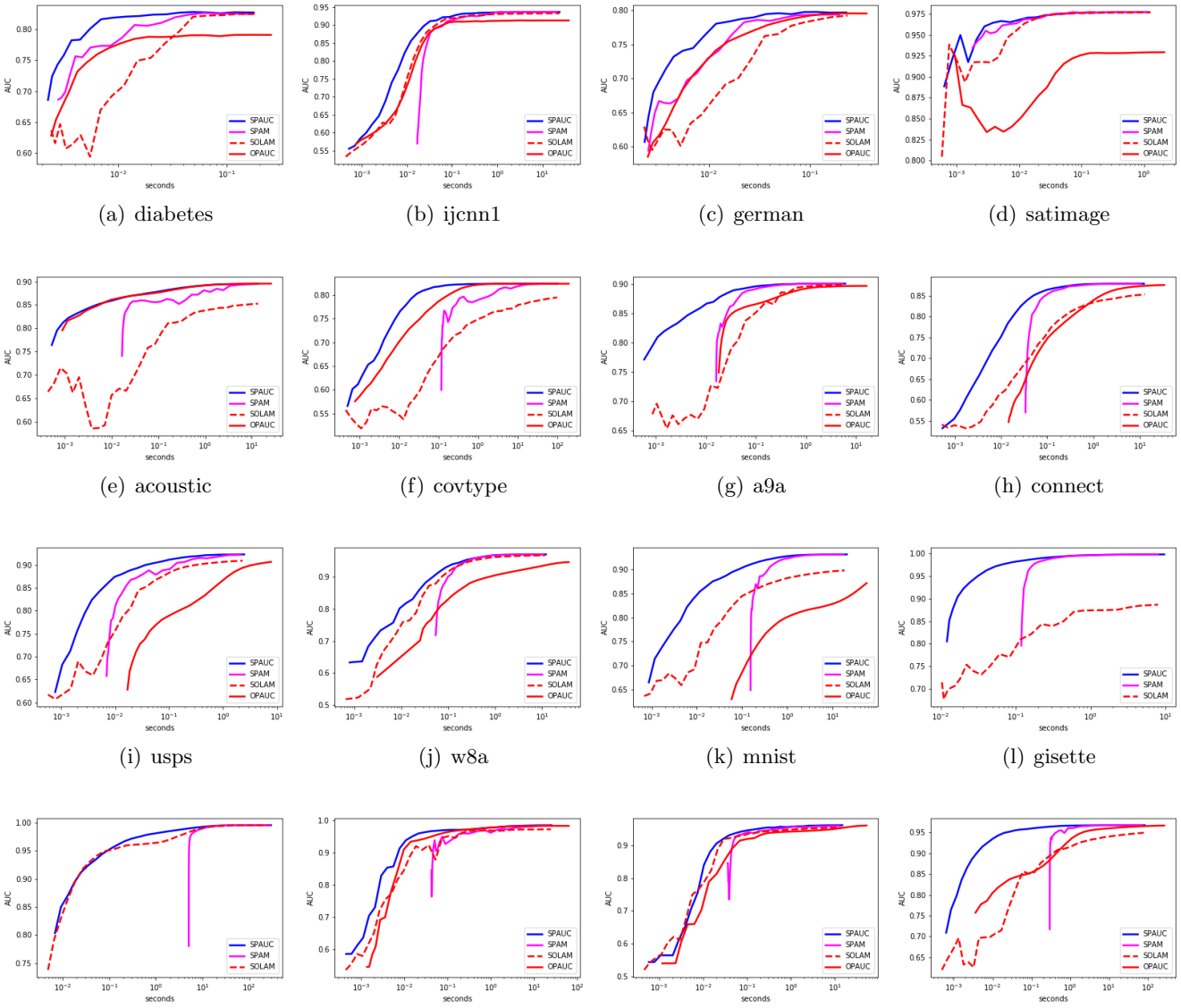
(e) acoustic (f) covtype (g) a9a (h) connect

(i) usps (j) w8a (k) mnist (l) gisette

(m) real-sim (n) protein_h (o) malware (p) webspam_u

Figure 2: AUC versus time curves (in seconds) for SPAUC, SPAM, SOLAM and OPAUC for objective functions with ℓ_2 regularizer $\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_2^2, \lambda = 10^{-6}$.

To show that SPAUC also works well with regularization, we consider (2.11) with either $\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_2^2$ or $\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$ in our experiments. We compare our method with SPAM, SOLAM and OPAUC. For the comparison ℓ_2 regularization, we modify the original SOLAM in Ying et al. (2016a) by replacing the ℓ_2-constraint with an ℓ_2-regularizer. For the comparison with ℓ_1 regularization, we modify the original implementation of SPAM, SOLAM and OPAUC to handle the ℓ_1 regularizer. Therefore, these methods optimize the same objective function. We fix the regularization parameter and tune the step-size



(a) diabetes (b) ijcnn1 (c) german (d) satimage

(e) acoustic (f) covtype (g) a9a (h) connect

(i) usps (j) w8a (k) mnist (l) gisette

(m) real-sim (n) protein_h (o) malware (p) webspam_u

Figure 3: AUC versus time curves (in seconds) for SPAUC, SPAM, SOLAM and OPAUC for objective functions with ℓ_1 regularizer $\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_1, \lambda = 10^{-6}$.

parameter μ by 5-fold cross validation. In Figure 2, we plot the AUC values as a function of execution time (in seconds) for SPAUC, SPAM, SOLAM and OPAUC in the setting of ℓ_2 regularization with $\lambda = 10^{-6}$. In Figure 3, we plot the AUC values as a function of execution time (in seconds) for SPAUC, SPAM, SOLAM and OPAUC in the setting of ℓ_1 regularization with $\lambda = 10^{-6}$. It can be seen that SPAUC attains a faster convergence speed as compared to the baseline methods. The same phenomenon also occurs for other choice

of regularization parameters, e.g., $\lambda = 10^{-2}$ and $\lambda = 10^{-4}$. We omit these results to save space.

5. Proofs

In this section, we present proofs for theoretical properties of SPAUC. We first give a *road-map* to clarify the basic idea. An essential ingredient of our proof is to show the almost boundedness of iterates. To this end, we first establish the self-bounding property of $\hat{F}_t$ (Lemma 11) and the one-step progress inequality (Lemma 12), based on which we establish a crude bound of $\mathbf{w}_t$ (Corollary 13). Then, we give high-probability bounds on the bias of using $\hat{F}_t'(\mathbf{w}_t; z_{i_t})$ as a gradient estimate (Lemma 14). We then use these results and a Bernstein-type inequality to tackle a martingale difference sequence in the one-step progress inequality, yielding a high-probability bound on the iterates (Theorem 3). This bound on $\{\mathbf{w}_t\}$ is further used to prove the convergence rate for general objective functions and objective functions with a quadratic functional growth.

5.1 Preliminary Lemmas

We present here some preliminary lemmas. The following lemma shows that an approximation of $p, \mathbb{E}[x'|y' = 1]$ and $\mathbb{E}[x'|y' = -1]$ by (2.9) still preserves the convexity. It also establishes the self-bounding property of $\hat{F}_t(\mathbf{w}; z)$.

Lemma 11 *For any* $\mathbf{w}$ *and* z*, we have*

$$\|\hat{F}_t'(\mathbf{w}; z)\|_2^2 \leq 16\kappa^2 \hat{F}_t(\mathbf{w}; z) \quad \textit{and} \quad \hat{F}_t(\mathbf{w}; z) \geq 0. \tag{5.1}$$

Furthermore, for any z *the function* $\hat{F}_t(\mathbf{w}; z)$ *is a convex function of* $\mathbf{w}$*.*

Proof The inequality $\hat{F}_t(\mathbf{w}; z) \geq 0$ follows directly from the Schwartz's inequality:

$$\hat{F}_t(\mathbf{w}; z) \geq 2p_t(1-p_t)\mathbf{w}^\top \big(v_t - u_t\big) + p_t(1-p_t)\big(\mathbf{w}^\top \big(v_t - u_t\big)\big)^2 + p_t(1-p_t) \geq 0.$$

For any $\mathbf{w}$ and $\tilde{\mathbf{w}}$, we have

$$\begin{aligned}\big\|\hat{F}_t'(\mathbf{w}; z) - \hat{F}_t'(\tilde{\mathbf{w}}; z)\big\|_2 \leq{}& 2(1-p_t)\big\|(x-u_t)(x-u_t)^\top(\mathbf{w}-\tilde{\mathbf{w}})\big\|_2 \mathbb{I}_{[y=1]} \\ &+ 2p_t\big\|(x-v_t)(x-v_t)^\top(\mathbf{w}-\tilde{\mathbf{w}})\big\|_2 \mathbb{I}_{[y=-1]} + \\ &2p_t(1-p_t)\big\|\big(v_t-u_t\big)\big(v_t-u_t\big)^\top(\mathbf{w}-\tilde{\mathbf{w}})\big\|_2 \leq 8\kappa^2\|\mathbf{w}-\tilde{\mathbf{w}}\|_2,\end{aligned}$$

where in the last inequality we have used the definition of κ.

Therefore, it follows from the self-bounding property of non-negative smooth functions (Lemma 17) that $\|\hat{F}_t'(\mathbf{w}; z)\|_2^2 \leq 16\kappa^2 \hat{F}_t(\mathbf{w}; z)$. This establishes (5.1).

It is clear that the Hessian matrix of $\hat{F}_t(\mathbf{w}; z)$ is

$$2(1-p_t)\big(x-u_t\big)\big(x-u_t\big)^\top \mathbb{I}_{[y=1]} + 2p_t\big(x-v_t\big)\big(x-v_t\big)^\top \mathbb{I}_{[y=-1]} + 2p_t(1-p_t)\big(v_t-u_t\big)\big(v_t-u_t\big)^\top,$$

which is a semi-positive definite matrix. Therefore, $\hat{F}_t(\cdot; z)$ is a convex function for any z. The proof is complete. ■

Our theoretical analysis roots its foundation on the following one-step progress inequality measuring how the iterate would change after a single iteration of (2.11). The proof is standard and is deferred to Appendix B.

Lemma 12 (One-step Progress Inequality) *Let $\{\mathbf{w}_t\}_t$ be produced by (2.11). If Assumption 1 holds, then for any $\mathbf{w} \in \mathbb{R}^d$ we have*

$$\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 - \|\mathbf{w} - \mathbf{w}_t\|_2^2 \leq 2\eta_t\langle \mathbf{w} - \mathbf{w}_t, \hat{F}_t'(\mathbf{w}_t; z_t)\rangle + 2\eta_t\big(\Omega(\mathbf{w}) - \Omega(\mathbf{w}_t)\big) \\ - \eta_t\sigma_\Omega\|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2 + 2\eta_t^2\big(C_1\hat{F}_t(\mathbf{w}_t; z_t) + C_1\Omega(\mathbf{w}_t) + A_2\big). \quad (5.2)$$

Based on Lemma 12, we can derive several useful inequalities collected in the following corollary. Eq. (5.3) provides a general bound on the norm of iterates in terms of step sizes. Eqs. (5.4) and (5.5) show how the accumulation of function values can be controlled by step sizes, which, according to Lemma 11 and Assumption 1, in turn give useful estimates on $\sum_{k=1}^t \eta_k^2\big(\|\hat{F}_k'(\mathbf{w}_k, z_k)\|_2^2 + \|\Omega'(\mathbf{w}_k)\|_2^2\big)$ and $\sum_{k=1}^t \big(\|\hat{F}_k'(\mathbf{w}_k, z_k)\|_2^2 + \|\Omega'(\mathbf{w}_k)\|_2^2\big)$ required to handle in convergence analysis. Since similar ideas have been used in Lei and Tang (2018), we defer the proof of Corollary 13 to Appendix B.

Corollary 13 *Let $\{\mathbf{w}_t\}_t$ be produced by (2.11). Suppose $\eta_t \leq (2C_1)^{-1}$ and Assumption 1 holds. Let $C_4 := C_1^{-1}A_2 + 2^{-1}$. Then*

$$\|\mathbf{w}_{t+1}\|_2^2 \leq C_4 \sum_{k=1}^t \eta_k. \quad (5.3)$$

Furthermore, if $\eta_{t+1} \leq \eta_t$, then

$$\sum_{k=1}^t \eta_k^2\big(\hat{F}_k(\mathbf{w}_k; z_k) + \Omega(\mathbf{w}_k)\big) \leq C_4 \sum_{k=1}^t \eta_k^2 \quad (5.4)$$

and

$$\sum_{k=1}^t \big(\hat{F}_k(\mathbf{w}_k; z_k) + \Omega(\mathbf{w}_k)\big) \leq C_4 t + C_4\eta_t^{-1}\sum_{k=1}^t \eta_k. \quad (5.5)$$

5.2 Approximation of Stochastic Gradients

The implementation of SPAUC requires to approximate the unbiased stochastic gradient $\widetilde{F}'(\mathbf{w}_t; z_t)$ by replacing the involved $p, \mathbb{E}[x'|y' = 1], \mathbb{E}[x'|y' = -1]$ with their empirical counterparts. The following lemma gives a quantitative measure on the accuracy of this approximation.

Lemma 14 *Let $\delta \in (0, 1)$. For any $t \in \mathbb{N}$, the following inequality holds with probability at least $1 - \delta$*

$$\big\|\widetilde{F}'(\mathbf{w}_t; z_t) - \hat{F}_t'(\mathbf{w}_t; z_t)\big\|_2 \leq \frac{2\kappa^2\big(2 + \sqrt{2\log(3/\delta)}\big)}{\sqrt{t}}\Big(\big(24 + \frac{8}{p}\mathbb{I}_{[y_t=1]} + \frac{8}{1-p}\mathbb{I}_{[y_t=-1]}\big)\|\mathbf{w}_t\|_2 + 3\Big).$$

Before proving Lemma 14, we need to introduce the following preliminary lemma. For a matrix A, we denote by $\|A\|_{\text{op}}$ the operator norm of A, i.e., $\|A\|_{\text{op}} = \sup_{\|\mathbf{w}\|_2=1} \|A\mathbf{w}\|_2$. For any $u, v \in \mathbb{R}^d$, there holds

$$\|uv^\top\|_{\text{op}} \leq \|u\|_2\|v\|_2. \tag{5.6}$$

Lemma 15 *Let $\delta \in (0,1)$. For any $t \in \mathbb{N}$, with probability at least $1-\delta$ the following inequalities hold simultaneously*

$$|p - p_t| \leq \big(2 + \sqrt{2\log(3/\delta)}\big)/\sqrt{t}, \tag{5.7}$$

$$\|\mathbb{E}[x'|y'=1] - u_t\|_2 \leq \frac{2\kappa(2+\sqrt{2\log(3/\delta)})}{p\sqrt{t}}, \tag{5.8}$$

$$\|\mathbb{E}[x'|y'=-1] - v_t\|_2 \leq \frac{2\kappa(2+\sqrt{2\log(3/\delta)})}{(1-p)\sqrt{t}}, \tag{5.9}$$

$$\Big\|(1-p_t)(x-u_t)(x-u_t)^\top - (1-p)\big(x-\mathbb{E}[x'|y'=1]\big)\big(x-\mathbb{E}[x'|y'=1]\big)^\top\Big\|_{\text{op}} \leq \frac{8\kappa^2\big(2+\sqrt{2\log(3/\delta)}\big)}{p\sqrt{t}}, \tag{5.10}$$

$$\Big\|p_t(x-v_t)(x-v_t)^\top - p\big(x-\mathbb{E}[x'|y'=-1]\big)\big(x-\mathbb{E}[x'|y'=-1]\big)^\top\Big\|_{\text{op}} \leq \frac{8\kappa^2\big(2+\sqrt{2\log(3/\delta)}\big)}{(1-p)\sqrt{t}}, \tag{5.11}$$

$$\big\|p(1-p)\big(\mathbb{E}[x'|y'=-1] - \mathbb{E}[x'|y'=1]\big) - p_t(1-p_t)(v_t-u_t)\big\|_2 \leq 3\kappa\big(2+\sqrt{2\log(3/\delta)}\big)/\sqrt{t}, \tag{5.12}$$

$$\Big\|p(1-p)\big(\mathbb{E}[x'|y'=-1] - \mathbb{E}[x'|y'=1]\big)\big(\mathbb{E}[x'|y'=-1] - \mathbb{E}[x'|y'=1]\big)^\top - p_t(1-p_t)(v_t-u_t)(v_t-u_t)^\top\Big\|_{\text{op}} \leq \frac{24\kappa^2\big(2+\sqrt{2\log(3/\delta)}\big)}{\sqrt{t}}. \tag{5.13}$$

Proof According to Lemma 18, with probability at least $1-\delta$ the following three inequalities hold simultaneously

$$|p - p_t| \leq \frac{2+\sqrt{2\log(3/\delta)}}{\sqrt{t}},$$

$$\Big\|\mathbb{E}[x'\mathbb{I}_{[y'=1]}] - \frac{1}{t}\sum_{i=0}^{t-1} x_i\mathbb{I}_{[y_i=1]}\Big\|_2 \leq \frac{(2+\sqrt{2\log(3/\delta)})\kappa}{\sqrt{t}}, \tag{5.14}$$

$$\Big\|\mathbb{E}[x'\mathbb{I}_{[y'=-1]}] - \frac{1}{t}\sum_{i=0}^{t-1} x_i\mathbb{I}_{[y_i=-1]}\Big\|_2 \leq \frac{(2+\sqrt{2\log(3/\delta)})\kappa}{\sqrt{t}}. \tag{5.15}$$

We now prove (5.8). According to (2.9), we know

$$\begin{aligned}\|\mathbb{E}[x'|y'=1]-u_t\|_2 &= \frac{1}{p}\Big\|p\mathbb{E}[x'|y'=1]-p_tu_t+p_tu_t-pu_t\Big\|_2\\ &\le \frac{1}{p}\Big\|\mathbb{E}[x'\mathbb{I}_{[y'=1]}]-\frac{1}{t}\sum_{i=0}^{t-1}x_i\mathbb{I}_{[y_i=1]}\Big\|_2+\frac{\|u_t\|_2}{p}|p_t-p|.\end{aligned}$$

Then we can apply (5.7) and (5.14) to derive (5.8) with probability at least $1-\delta$.

Eq. (5.9) can be proved in a similar manner and we omit the proof for brevity.

We now show (5.10). It is clear that

$$\begin{aligned}(1-p_t)(x-u_t)(x-u_t)^\top-(1-p)\big(x-\mathbb{E}[x'|y'=1]\big)\big(x-\mathbb{E}[x'|y'=1]\big)^\top &= \big((1-p_t)-(1-p)\big)(x-u_t)(x-u_t)^\top\\ &+(1-p)(x-u_t)(x-u_t)^\top-(1-p)(x-u_t)\big(x-\mathbb{E}[x'|y'=1]\big)^\top\\ &+(1-p)(x-u_t)\big(x-\mathbb{E}[x'|y'=1]\big)^\top-(1-p)\big(x-\mathbb{E}[x'|y'=1]\big)\big(x-\mathbb{E}[x'|y'=1]\big)^\top,\end{aligned}$$

from which and (5.6) we derive

$$\begin{aligned}&\Big\|(1-p_t)(x-u_t)(x-u_t)^\top-(1-p)\big(x-\mathbb{E}[x'|y'=1]\big)\big(x-\mathbb{E}[x'|y'=1]\big)^\top\Big\|_{\rm op}\le|p-p_t|\Big\|\big(x-u_t\big)\big(x-u_t\big)^\top\Big\|_{\rm op}\\ &+(1-p)\Big\|(x-u_t)\big(\mathbb{E}[x'|y'=1]-u_t\big)^\top\Big\|_{\rm op}+(1-p)\Big\|\big(\mathbb{E}[x'|y'=1]-u_t\big)\big(x-\mathbb{E}[x'|y'=1]\big)^\top\Big\|_{\rm op}\\ &\le 4\kappa^2|p-p_t|+4\kappa(1-p)\|\mathbb{E}[x'|y'=1]-u_t\|_2.\end{aligned}$$

This together with (5.7) and (5.8) shows (5.10) with probability at least $1-\delta$.

Eq. (5.11) can be proved in a similar manner and we omit the proof for brevity.

We now prove (5.12). It is clear

$$\begin{aligned}p(1-p)\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x'|y'=1]\big)-p_t(1-p_t)(v_t-u_t) &= p\mathbb{E}[x'\mathbb{I}_{[y'=-1]}]-(1-p)\mathbb{E}[x'\mathbb{I}_{[y'=1]}]\\ &-\frac{p_t}{t}\Big(\sum_{i=0}^{t-1}x_i\mathbb{I}_{[y_i=-1]}\Big)+\frac{1-p_t}{t}\Big(\sum_{i=0}^{t-1}x_i\mathbb{I}_{[y_i=1]}\Big),\end{aligned}$$

from which we derive

$$\begin{aligned}&\big\|p(1-p)\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x'|y'=1]\big)-p_t(1-p_t)(v_t-u_t)\big\|_2\\ &\le\Big\|p\mathbb{E}[x'\mathbb{I}_{[y'=-1]}]-\frac{p_t}{t}\Big(\sum_{i=0}^{t-1}x_i\mathbb{I}_{[y_i=-1]}\Big)\Big\|_2+\Big\|(1-p)\mathbb{E}[x'\mathbb{I}_{[y'=1]}]-\frac{(1-p_t)}{t}\Big(\sum_{i=0}^{t-1}x_i\mathbb{I}_{[y_i=1]}\Big)\Big\|_2\\ &\le 2\kappa|p-p_t|+p_t\Big\|\mathbb{E}[x'\mathbb{I}_{[y'=-1]}]-\frac{1}{t}\Big(\sum_{i=0}^{t-1}x_i\mathbb{I}_{[y_i=-1]}\Big)\Big\|_2+(1-p_t)\Big\|\mathbb{E}[x'\mathbb{I}_{[y'=1]}]-\frac{1}{t}\Big(\sum_{i=0}^{t-1}x_i\mathbb{I}_{[y_i=1]}\Big)\Big\|_2,\end{aligned}$$

where we have used $p\mathbb{E}[x'\mathbb{I}_{[y'=-1]}]=(p-p_t)\mathbb{E}[x'\mathbb{I}_{[y'=-1]}]+p_t\mathbb{E}[x'\mathbb{I}_{[y'=-1]}]$. We can then apply (5.7), (5.14) and (5.15) to derive the bound (5.12) with probability $1-\delta$.

We now prove (5.13). It is clear

$$\begin{aligned}
&p(1-p)\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x'|y'=1]\big)\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x'|y'=1]\big)^\top - p_t(1-p_t)(v_t-u_t)(v_t-u_t)^\top\\
&\quad= p(1-p)\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x'|y'=1]\big)\Big(\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x'|y'=1]\big)^\top-\big(v_t-u_t\big)^\top\Big)\\
&\quad+ p(1-p)\Big(\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x'|y'=1]\big)-\big(v_t-u_t\big)\Big)\big(v_t-u_t\big)^\top\\
&\quad+ \big(p(1-p)-p_t(1-p_t)\big)\big(v_t-u_t\big)\big(v_t-u_t\big)^\top,
\end{aligned}$$

from which and (5.6) it follows that

$$\begin{aligned}
&\Big\|p(1-p)\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x'|y'=1]\big)\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x'|y'=1]\big)^\top-p_t(1-p_t)(v_t-u_t)(v_t-u_t)^\top\Big\|_{\rm op}\\
&\quad\le 4p(1-p)\kappa\Big\|\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x'|y'=1]\big)-\big(v_t-u_t\big)\Big\|_2+4\kappa^2|p-p_t||p+p_t-1|.
\end{aligned}$$

Furthermore, there holds that

$$\begin{aligned}
&p(1-p)\Big\|\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x'|y'=1]\big)-\big(v_t-u_t\big)\Big\|_2\\
&\le\Big\|p(1-p)\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x|y=1]\big)-p_t(1-p_t)(v_t-u_t)\Big\|_2+\big|p_t(1-p_t)-p(1-p)\big|\|v_t-u_t\|_2\\
&\le\Big\|p(1-p)\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x'|y'=1]\big)-p_t(1-p_t)(v_t-u_t)\Big\|_2+2\kappa|p-p_t||p+p_t-1|.
\end{aligned}$$

Combining the above two inequalities and (5.7), (5.12) together then imply the stated inequality (5.13) with probability $1-\delta$. The proof is complete. ■

Proof of Lemma 14 It follows from (2.1) that

$$\begin{aligned}
&\widetilde{F}'(\mathbf{w};z)=2(1-p)\big(x-\mathbb{E}[x'|y'=1]\big)\big(x-\mathbb{E}[x'|y'=1]\big)^\top\mathbf{w}\mathbb{I}_{[y=1]}+\\
&2p(x-\mathbb{E}[x'|y'=-1])(x-\mathbb{E}[x'|y'=-1])^\top\mathbf{w}\mathbb{I}_{[y=-1]}+2p(1-p)\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x'|y'=1]\big)\\
&\qquad+2p(1-p)\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x'|y'=1]\big)\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x'|y'=1]\big)^\top\mathbf{w}. \quad (5.16)
\end{aligned}$$

This together with (2.10) shows that

$$\begin{aligned}
&\big\|\widetilde{F}'(\mathbf{w}_t;z_t)-\hat{F}_t'(\mathbf{w}_t;z_t)\big\|_2\le 2\Big\|p(1-p)\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x'|y'=1]\big)-p_t(1-p_t)(v_t-u_t)\Big\|_2\\
&+2\Big\|(1-p)\big(x-\mathbb{E}[x'|y'=1]\big)\big(x-\mathbb{E}[x'|y'=1]\big)^\top-(1-p_t)(x-u_t)(x-u_t)^\top\Big\|_{\rm op}\|\mathbf{w}_t\|_2\mathbb{I}_{[y_t=1]}\\
&+2\Big\|p\big(x-\mathbb{E}[x'|y'=-1]\big)\big(x-\mathbb{E}[x'|y'=-1]\big)^\top-p_t(x-v_t)(x-v_t)^\top\Big\|_{\rm op}\|\mathbf{w}_t\|_2\mathbb{I}_{[y_t=-1]}\\
&+2\Big\|p(1-p)\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x|y=1]\big)\big(\mathbb{E}[x'|y'=-1]-\mathbb{E}[x|y=1]\big)^\top-p_t(1-p_t)(v_t-u_t)(v_t-u_t)^\top\Big\|_{\rm op}\|\mathbf{w}_t\|_2\\
&:=\mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{IV}.
\end{aligned}$$

We can apply (5.12), (5.10), (5.11), and (5.13) to develop upper bounds of I, II, III and IV with probability at least $1-\delta$, respectively. Also, note that the high probability is w.r.t.

$z_1, \ldots, z_{t-1}$, which is independent of z_t. We can plug these high-probability bounds back to the above inequality, and get the stated bound. The proof is complete. ■

5.3 Boundedness of Iterates

In this subsection, we prove Lemma 13 on the almost boundedness of iterates. To this aim, we first establish a recursive inequality showing how $\|\mathbf{w}_{t+1} - \mathbf{w}_1^*\|_2^2$ can be controlled by $\|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2$ for $k = 1, \ldots, t$. Our basic idea is to control $\|\mathbf{w}_{t+1} - \mathbf{w}_1^*\|_2^2$ by

$$\mathcal{O}(1)\Big(\sum_{k=1}^t \eta_k(\phi(\mathbf{w}_1^*) - \phi(\mathbf{w}_k)) + \sum_{k=1}^t \eta_k\big\langle \mathbf{w}_1^* - \mathbf{w}_k, \hat{F}_k'(\mathbf{w}_k; z_k) - \widetilde{F}'(\mathbf{w}_k; z_k)\big\rangle + \sum_{k=1}^t \xi_k\Big), \quad (5.17)$$

where $\{\xi_k\}_k$ is a martingale difference sequence defined in (5.23). We apply Lemma 14 to control $\sum_{k=1}^t \eta_k\langle \mathbf{w}_1^* - \mathbf{w}_k, \hat{F}_k'(\mathbf{w}_k; z_k) - \widetilde{F}'(\mathbf{w}_k; z_k)\rangle$, and apply Part (b) of Lemma 19 to show with high probability that $\sum_{k=1}^t \xi_k \leq \sum_{k=1}^t \eta_k(\phi(\mathbf{w}_k) - \phi(\mathbf{w}_1^*)) + \widetilde{C}\sum_{k=1}^t \eta_k^2\|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2$ for a constant $\widetilde{C} > 0$. The key observation is that the partial variance $\sum_{k=1}^t \eta_k(\phi(\mathbf{w}_k) - \phi(\mathbf{w}_1^*))$ can be cancelled out by the term $\sum_{k=1}^t \eta_k(\phi(\mathbf{w}_1^*) - \phi(\mathbf{w}_k))$ in (5.17).

Proposition 16 *Let $\{\mathbf{w}_t\}_t$ be produced by (2.11) with $\eta_t \leq (2C_1)^{-1}$ and $\eta_{t+1} \leq \eta_t$. We suppose Assumption 1 holds,*

$$C_5 = \sup_k \eta_k \sum_{j=1}^{k-1} \eta_j < \infty,\ C_6 = \eta_1 \sup_z \widetilde{F}(\mathbf{w}_1^*, z) + 2p(1-p)\Big(7\kappa^2 C_4 C_5 + \eta_1\big(1 + 2\kappa\|\mathbf{w}_1^*\|_2 + 2\kappa^2\|\mathbf{w}_1^*\|_2^2\big)\Big).$$

Then for any $\delta \in (0,1)$ and $\rho = \min\{1, (2C_1)^{-1}(\eta_1\|\mathbf{w}_1^*\|_2^2 + C_4C_5)^{-1}C_6\}$, the following inequality holds with probability at least $1-\delta$ simultaneously for all $t = 1, \ldots, T$

$$\|\mathbf{w}_{t+1} - \mathbf{w}_1^*\|_2^2 \leq \|\mathbf{w}_1^*\|_2^2 + \sum_{k=1}^t \frac{2C_{k,\delta}\eta_k(\|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2 + 1)}{\sqrt{k}} + \frac{\phi(\mathbf{w}_1^*)}{C_4C_5}\sum_{k=1}^t \eta_k^2\|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2 + \frac{2C_6\log(2T/\delta)}{\rho} + 2(C_1C_4 + A_2)\sum_{k=1}^t \eta_k^2,$$

where we introduce $C_p = 8\max\{p^{-1}, (1-p)^{-1}\} + 24$ and

$$C_{k,\delta} = 2\kappa^2\big(2 + \sqrt{2\log(12k^2/\delta)}\big)\max\big\{C_p + 1, 4^{-1}(C_p\|\mathbf{w}_1^*\|_2 + 3)^2\big\}.$$

Proof Taking $\mathbf{w} = \mathbf{w}_1^*$ in (5.2) gives

$$\|\mathbf{w}_{t+1} - \mathbf{w}_1^*\|_2^2 - \|\mathbf{w}_t - \mathbf{w}_1^*\|_2^2 \leq 2\eta_t\langle \mathbf{w}_1^* - \mathbf{w}_t, \hat{F}_t'(\mathbf{w}_t, z_t)\rangle + 2\eta_t(\Omega(\mathbf{w}_1^*) - \Omega(\mathbf{w}_t)) + 2\eta_t^2\big(C_1\hat{F}_t(\mathbf{w}_t; z_t) + C_1\Omega(\mathbf{w}_t) + A_2\big).$$

Taking a summation of the above inequality gives ($\mathbf{w}_1 = 0$)

$$\|\mathbf{w}_{t+1} - \mathbf{w}_1^*\|_2^2 - \|\mathbf{w}_1^*\|_2^2 = \sum_{k=1}^{t} \left[\|\mathbf{w}_{k+1} - \mathbf{w}_1^*\|_2^2 - \|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2\right] \leq 2\sum_{k=1}^{t} \eta_k \langle \mathbf{w}_1^* - \mathbf{w}_k, \hat{F}_k'(\mathbf{w}_k; z_k)\rangle$$
$$+ 2\sum_{k=1}^{t} \eta_k (\Omega(\mathbf{w}_1^*) - \Omega(\mathbf{w}_k)) + 2\sum_{k=1}^{t} \eta_k^2 \big(C_1 \hat{F}_k(\mathbf{w}_k; z_k) + C_1 \Omega(\mathbf{w}_k) + A_2\big)$$
$$\leq 2\sum_{k=1}^{t} \eta_k \langle \mathbf{w}_1^* - \mathbf{w}_k, \hat{F}_k'(\mathbf{w}_k; z_k)\rangle + 2\sum_{k=1}^{t} \eta_k (\Omega(\mathbf{w}_1^*) - \Omega(\mathbf{w}_k)) + 2(C_1 C_4 + A_2) \sum_{k=1}^{t} \eta_k^2, \tag{5.18}$$

where the last inequality is due to (5.4). We consider the following decomposition

$$\sum_{k=1}^{t} \eta_k \langle \mathbf{w}_1^* - \mathbf{w}_k, \hat{F}_k'(\mathbf{w}_k; z_k)\rangle = \sum_{k=1}^{t} \eta_k \big\langle \mathbf{w}_1^* - \mathbf{w}_k, \hat{F}_k'(\mathbf{w}_k; z_k) - \widetilde{F}'(\mathbf{w}_k; z_k)\big\rangle$$
$$+ \sum_{k=1}^{t} \eta_k \big\langle \mathbf{w}_1^* - \mathbf{w}_k, \widetilde{F}'(\mathbf{w}_k; z_k) - \nabla f(\mathbf{w}_k)\big\rangle + \sum_{k=1}^{t} \eta_k \langle \mathbf{w}_1^* - \mathbf{w}_k, \nabla f(\mathbf{w}_k)\rangle. \tag{5.19}$$

For any $k \in \mathbb{N}$, by Lemma 14 the following inequality holds with probability at least $1 - \delta/(4k^2)$

$$\big\|\widetilde{F}'(\mathbf{w}_k; z_k) - \hat{F}_k'(\mathbf{w}_k; z_k)\big\|_2 \leq \frac{2\kappa^2\big(2 + \sqrt{2\log(12k^2/\delta)}\big)}{\sqrt{k}} \big(C_p \|\mathbf{w}_k\|_2 + 3\big),$$

which together with union bounds and $\sum_{k=1}^{\infty} k^{-2} \leq 2$ gives the following inequality with probability $1 - \delta/2$ simultaneously for all $k = 1, \ldots, \infty$

$$\big\|\widetilde{F}'(\mathbf{w}_k; z_k) - \hat{F}_k'(\mathbf{w}_k; z_k)\big\|_2 \leq \frac{2\kappa^2\big(2 + \sqrt{2\log(12k^2/\delta)}\big)}{\sqrt{k}} \big(C_p \|\mathbf{w}_k - \mathbf{w}_1^*\|_2 + C_p \|\mathbf{w}_1^*\|_2 + 3\big). \tag{5.20}$$

It then follows that the following inequality holds with probability at least $1 - \delta/2$ simultaneously for all $t = 1, \ldots, \infty$

$$\sum_{k=1}^{t} \eta_k \big\langle \mathbf{w}_1^* - \mathbf{w}_k, \hat{F}_k'(\mathbf{w}_k; z_k) - \widetilde{F}'(\mathbf{w}; z_k)\big\rangle \leq \sum_{k=1}^{t} \eta_k \|\mathbf{w}_k - \mathbf{w}_1^*\|_2 \big\|\hat{F}_k'(\mathbf{w}_k; z_k) - \widetilde{F}'(\mathbf{w}_k; z_k)\big\|_2$$
$$\leq 2\kappa^2 \sum_{k=1}^{t} \eta_k \big(2 + \sqrt{2\log(12k^2/\delta)}\big) \frac{C_p \|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2 + (C_p \|\mathbf{w}_1^*\|_2 + 3)\|\mathbf{w}_k - \mathbf{w}_1^*\|_2}{\sqrt{k}}$$
$$\leq \sum_{k=1}^{t} \eta_k C_{k,\delta} (\|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2 + 1)/\sqrt{k}, \tag{5.21}$$

where in the last step we have used the Schwartz's inequality

$$\big(3 + C_p \|\mathbf{w}_1^*\|_2\big)\|\mathbf{w}_k - \mathbf{w}_1^*\|_2 \leq \|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2 + (3 + C_p \|\mathbf{w}_1^*\|_2)^2/4.$$

It follows from the convexity of f that

$$\sum_{k=1}^{t} \eta_k \langle \mathbf{w}_1^* - \mathbf{w}_k, \nabla f(\mathbf{w}_k) \rangle \leq \sum_{k=1}^{t} \eta_k \big(f(\mathbf{w}_1^*) - f(\mathbf{w}_k) \big). \tag{5.22}$$

We now control the last second term of (5.19) with an application of a concentration inequality for a martingale difference sequence. Introduce a sequence of random variables

$$\xi_k := \eta_k \big\langle \mathbf{w}_1^* - \mathbf{w}_k, \widetilde{F}'(\mathbf{w}_k; z_k) - \nabla f(\mathbf{w}_k) \big\rangle, \quad k \in \mathbb{N}. \tag{5.23}$$

It follows from Proposition 1 that $\mathbb{E}_{z_k}[\xi_k] = 0$ and therefore $\{\xi_k\}_k$ is a martingale difference sequence. Analogous to Lemma 11, we can show

$$\big\| \widetilde{F}'(\mathbf{w}_k; z_k) \big\|_2^2 \leq 16 \kappa^2 \widetilde{F}(\mathbf{w}_k, z_k). \tag{5.24}$$

Since $\mathbb{E}[(\xi - \mathbb{E}[\xi])^2] \leq \mathbb{E}[\xi^2]$ for any real-valued random variable ξ, it then follows that

$$\begin{aligned}
&\mathbb{E}_{z_k} \Big[\big| \big\langle \mathbf{w}_1^* - \mathbf{w}_k, \widetilde{F}'(\mathbf{w}_k; z_k) - \nabla f(\mathbf{w}_k) \big\rangle \big|^2 \Big] \leq \mathbb{E}_{z_k} \Big[\big| \big\langle \mathbf{w}_1^* - \mathbf{w}_k, \widetilde{F}'(\mathbf{w}_k; z_k) \big\rangle \big|^2 \Big] \\
&\leq \|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2 \mathbb{E}_{z_k} \Big[\big\| \widetilde{F}'(\mathbf{w}_k; z_k) \big\|_2^2 \Big] \leq \|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2 \mathbb{E}_{z_k} \big[C_1 \widetilde{F}(\mathbf{w}_k, z_k) \big] = C_1 f(\mathbf{w}_k) \|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2,
\end{aligned}$$

where we have used the definition of C_1 and Proposition 1. It then follows that

$$\begin{aligned}
&\sum_{k=1}^{t} \mathbb{E}_{z_k} \big[\big(\xi_k - \mathbb{E}_{z_k}[\xi_k] \big)^2 \big] = \sum_{k=1}^{t} \eta_k^2 \mathbb{E}_{z_k} \Big[\big| \big\langle \mathbf{w}_1^* - \mathbf{w}_k, \widetilde{F}'(\mathbf{w}_k; z_k) - \nabla f(\mathbf{w}_k) \big\rangle \big|^2 \Big] \\
&\leq \sum_{k=1}^{t} \eta_k^2 \|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2 \big(C_1 \phi(\mathbf{w}_k) - C_1 \phi(\mathbf{w}_1^*) \big) + \sum_{k=1}^{t} \eta_k^2 \|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2 C_1 \phi(\mathbf{w}_1^*).
\end{aligned}$$

By (5.3), $C_5 = \sup_k \eta_k \sum_{j=1}^{k-1} \eta_j < \infty$ and $f(\mathbf{w}) \leq \phi(\mathbf{w})$, we know

$$\eta_k^2 \|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2 \leq 2\eta_k^2 (\|\mathbf{w}_k\|_2^2 + \|\mathbf{w}_1^*\|_2^2) \leq 2\eta_k \Big(\eta_k \|\mathbf{w}_1^*\|_2^2 + C_4 \eta_k \sum_{j=1}^{k-1} \eta_j \Big) \leq 2\eta_k \big(\eta_1 \|\mathbf{w}_1^*\|_2^2 + C_4 C_5 \big).$$

Combining the above two inequalities together, we derive

$$\begin{aligned}
&\sum_{k=1}^{t} \mathbb{E}_{z_k} \big[\big(\xi_k - \mathbb{E}_{z_k}[\xi_k] \big)^2 \big] \leq \\
&\quad 2C_1 \big(\eta_1 \|\mathbf{w}_1^*\|_2^2 + C_4 C_5 \big) \sum_{k=1}^{t} \eta_k \big(\phi(\mathbf{w}_k) - \phi(\mathbf{w}_1^*) \big) \big) + C_1 \phi(\mathbf{w}_1^*) \sum_{k=1}^{t} \eta_k^2 \|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2.
\end{aligned} \tag{5.25}$$

According to the convexity of $\widetilde{F}$ established in Proposition 1, we know

$$\begin{aligned}
&\xi_k - \mathbb{E}_{z_k}[\xi_k] = \eta_k\langle \mathbf{w}_1^* - \mathbf{w}_k, \widetilde{F}'(\mathbf{w}_k; z_k)\rangle + \eta_k\langle \mathbf{w}_k - \mathbf{w}_1^*, \nabla f(\mathbf{w}_k)\rangle \\
&\le \eta_k\big[\widetilde{F}(\mathbf{w}_1^*; z_k) - \widetilde{F}(\mathbf{w}_k; z_k)\big] + \eta_k\big(\|\mathbf{w}_k\|_2 + \|\mathbf{w}_1^*\|_2\big)\big(4p(1-p)\kappa + 8p(1-p)\kappa^2\|\mathbf{w}_k\|_2\big) \\
&\le \eta_k\widetilde{F}(\mathbf{w}_1^*; z_k) + 4\eta_k p(1-p)\Big(\kappa\|\mathbf{w}_k\|_2 + \kappa\|\mathbf{w}_1^*\|_2 + 2\kappa^2\|\mathbf{w}_k\|_2^2 + 2\kappa^2\|\mathbf{w}_k\|_2\|\mathbf{w}_1^*\|_2\Big) \\
&\le \eta_k\widetilde{F}(\mathbf{w}_1^*; z_k) + 2\eta_k p(1-p)\Big(\kappa^2\|\mathbf{w}_k\|_2^2 + 1 + 2\kappa\|\mathbf{w}_1^*\|_2 + 4\kappa^2\|\mathbf{w}_k\|_2^2 + 2\kappa^2\|\mathbf{w}_k\|_2^2 + 2\kappa^2\|\mathbf{w}_1^*\|_2^2\Big) \\
&\le \eta_k\widetilde{F}(\mathbf{w}_1^*; z_k) + 2p(1-p)\Big(7\kappa^2\eta_k C_4\sum_{j=1}^{k-1}\eta_j + \eta_k\big(1 + 2\kappa\|\mathbf{w}_1^*\|_2 + 2\kappa^2\|\mathbf{w}_1^*\|_2^2\big)\Big) \le C_6,
\end{aligned}$$

where we have used the following inequality in the second inequality ($\nabla f(\mathbf{w}) = 2p(1-p)\mathbb{E}\big[(1-\mathbf{w}^\top(x-x'))(x-x')|y=1, y'=-1\big]$)

$$\|\nabla f(\mathbf{w})\|_2 \le 4p(1-p)\kappa + 8p(1-p)\kappa^2\|\mathbf{w}\|_2, \quad \forall \mathbf{w}\in\mathbb{R}^d, \tag{5.26}$$

(5.3) and $C_5 = \sup_k \eta_k\sum_{j=1}^{k-1}\eta_j < \infty$ in the last inequality. The above bounds on magnitudes and variances of ξ_k together with Part (b) of Lemma 19 (see the Appendix) imply the following inequality with probability $1-\delta/2$

$$\begin{aligned}
&\sum_{k=1}^t \xi_k \le \frac{\rho}{C_6}\bigg(2C_1(\eta_1\|\mathbf{w}_1^*\|_2^2 + C_4C_5)\sum_{k=1}^t \eta_k\big(\phi(\mathbf{w}_k) - \phi(\mathbf{w}_1^*)\big) + C_1\phi(\mathbf{w}_1^*)\sum_{k=1}^t \eta_k^2\|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2\bigg) \\
&\quad + \frac{C_6\log(2/\delta)}{\rho} \le \sum_{k=1}^t \eta_k\big(\phi(\mathbf{w}_k) - \phi(\mathbf{w}_1^*)\big) + \frac{\phi(\mathbf{w}_1^*)}{2C_4C_5}\sum_{k=1}^t \eta_k^2\|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2 + \frac{C_6\log(2/\delta)}{\rho},
\end{aligned} \tag{5.27}$$

where we have used the inequality $2C_1\rho(\eta_1\|\mathbf{w}_1^*\|_2^2 + C_4C_5) \le C_6$. Plugging (5.21), (5.22) and (5.27) into (5.19) gives the following inequality with probability $1-\delta$

$$\begin{aligned}
\sum_{k=1}^t \eta_k\langle \mathbf{w}_1^* - \mathbf{w}_k, \hat{F}_k'(\mathbf{w}_k; z_k)\rangle &\le \sum_{k=1}^t C_{k,\delta}\eta_k(\|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2 + 1)/\sqrt{k} + \sum_{k=1}^t \eta_k\big(f(\mathbf{w}_1^*) - f(\mathbf{w}_k)\big) \\
&\quad + \sum_{k=1}^t \eta_k\big(\phi(\mathbf{w}_k) - \phi(\mathbf{w}_1^*)\big) + \frac{\phi(\mathbf{w}_1^*)}{2C_4C_5}\sum_{k=1}^t \eta_k^2\|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2 + \frac{C_6\log(2/\delta)}{\rho}.
\end{aligned}$$

This together with (5.18) shows the following inequality with probability $1-\delta$

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}_1^*\|_2^2 \le \|\mathbf{w}_1^*\|_2^2 + \sum_{k=1}^t \frac{2C_{k,\delta}\eta_k(\|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2 + 1)}{\sqrt{k}} + \frac{\phi(\mathbf{w}_1^*)}{C_4C_5}\sum_{k=1}^t \eta_k^2\|\mathbf{w}_k - \mathbf{w}_1^*\|_2^2 \\
+ \frac{2C_6\log(2/\delta)}{\rho} + 2(C_1C_4 + A_2)\sum_{k=1}^t \eta_k^2.
\end{aligned}$$

Note (5.20) holds simultaneously for all $k = 1, \ldots, \infty$. To derive the stated inequality for all $t = 1, \ldots, T$, one needs to derive (5.27) simultaneously for all $k = 1, \ldots, T$. This can be

done by replacing $\log(2/\delta)$ in (5.27) with $\log(2T/\delta)$. The proof is complete. ■

According to the assumption $\sum_{k=1}^{\infty}\eta_k^2<\infty$ and $\sum_{k=1}^{\infty}\eta_k\sqrt{\log k}/\sqrt{k}<\infty$, Proposition 16 essentially implies that

$$\max_{1\le k\le t}\|\mathbf{w}_k-\mathbf{w}_1^*\|_2^2\le\frac{1}{2}\max_{1\le k\le t}\|\mathbf{w}_k-\mathbf{w}_1^*\|_2^2+\widetilde{C}\log\frac{1}{\delta}$$

for a $\widetilde{C}>0$, from which we can derive an almost boundedness of $\{\mathbf{w}_t\}_t$. We will rigorously show this in the following proof.

Proof of Theorem 3 Introduce the set

$$\Omega_T=\Bigg\{(z_1,\ldots,z_T):\|\mathbf{w}_{t+1}-\mathbf{w}_1^*\|_2^2\le\|\mathbf{w}_1^*\|_2^2+\sum_{k=1}^{t}\frac{2C_{k,\delta}\eta_k(\|\mathbf{w}_k-\mathbf{w}_1^*\|_2^2+1)}{\sqrt{k}}+$$
$$\frac{\phi(\mathbf{w}_1^*)}{C_4C_5}\sum_{k=1}^{t}\eta_k^2\|\mathbf{w}_k-\mathbf{w}_1^*\|_2^2+\frac{2C_6\log(2T/\delta)}{\rho}+2(C_1C_4+A_2)\sum_{k=1}^{t}\eta_k^2\quad\text{for all } t=1,\ldots,T\Bigg\},$$

where ρ is defined in Proposition 16. Proposition 16 shows that $\Pr(\Omega_T)\ge1-\delta$. Since $\sum_{t=1}^{\infty}\eta_t\sqrt{\log t}/\sqrt{t}<\infty$ and $\sum_{t=1}^{\infty}\eta_t^2<\infty$, we can find a $t_2\in\mathbb{N}$ such that

$$\sum_{k=t_2+1}^{\infty}\frac{2C_{k,\delta}\eta_k}{\sqrt{k}}<1/4\quad\text{and}\quad\sum_{k=t_2+1}^{\infty}\eta_k^2<\frac{C_4C_5}{4\phi(\mathbf{w}_1^*)}.\tag{5.28}$$

Conditioned on the event Ω_T, we derive the following inequality for all $t=1,\ldots,T$

$$\begin{aligned}
&\|\mathbf{w}_{t+1}-\mathbf{w}_1^*\|_2^2-\|\mathbf{w}_1^*\|_2^2\\
&\le\sum_{k=1}^{t_2}\frac{2C_{k,\delta}\eta_k\|\mathbf{w}_k-\mathbf{w}_1^*\|_2^2}{\sqrt{k}}+\max_{1\le\tilde{t}\le T}\|\mathbf{w}_{\tilde{t}}-\mathbf{w}_1^*\|_2^2\sum_{k=t_2+1}^{T}\frac{2C_{k,\delta}\eta_k}{\sqrt{k}}+\frac{\phi(\mathbf{w}_1^*)}{C_4C_5}\sum_{k=1}^{t_2}\eta_k^2\|\mathbf{w}_k-\mathbf{w}_1^*\|_2^2\\
&+\frac{\phi(\mathbf{w}_1^*)\max_{1\le\tilde{t}\le T}\|\mathbf{w}_{\tilde{t}}-\mathbf{w}_1^*\|_2^2}{C_4C_5}\sum_{k=t_2+1}^{T}\eta_k^2+\sum_{k=1}^{t}\frac{2C_{k,\delta}\eta_k}{\sqrt{k}}+\frac{2C_6\log(2T/\delta)}{\rho}+2(C_1C_4+A_2)\sum_{k=1}^{t}\eta_k^2\\
&\le\sum_{k=1}^{t_2}\frac{2C_{k,\delta}\eta_k\|\mathbf{w}_k-\mathbf{w}_1^*\|_2^2}{\sqrt{k}}+\frac{1}{4}\max_{1\le\tilde{t}\le T}\|\mathbf{w}_{\tilde{t}}-\mathbf{w}_1^*\|_2^2+\frac{\phi(\mathbf{w}_1^*)}{C_4C_5}\sum_{k=1}^{t_2}\eta_k^2\|\mathbf{w}_k-\mathbf{w}_1^*\|_2^2\\
&+\frac{1}{4}\max_{1\le\tilde{t}\le T}\|\mathbf{w}_{\tilde{t}}-\mathbf{w}_1^*\|_2^2+\sum_{k=1}^{t}\frac{2C_{k,\delta}\eta_k}{\sqrt{k}}+\frac{2C_6\log(2T/\delta)}{\rho}+2(C_1C_4+A_2)\sum_{k=1}^{t}\eta_k^2.
\end{aligned}$$

It then follows the following inequality under the event Ω_T

$$\begin{aligned}
\max_{1\le\tilde{t}\le T}\|\mathbf{w}_{\tilde{t}}-\mathbf{w}_1^*\|_2^2&\le\|\mathbf{w}_1^*\|_2^2+\sum_{k=1}^{t_2}\frac{2C_{k,\delta}\eta_k\|\mathbf{w}_k-\mathbf{w}_1^*\|_2^2}{\sqrt{k}}+\frac{1}{2}\max_{1\le\tilde{t}\le T}\|\mathbf{w}_{\tilde{t}}-\mathbf{w}_1^*\|_2^2\\
&+\frac{\phi(\mathbf{w}_1^*)}{C_4C_5}\sum_{k=1}^{t_2}\eta_k^2\|\mathbf{w}_k-\mathbf{w}_1^*\|_2^2+\sum_{k=1}^{T}\frac{2C_{k,\delta}\eta_k}{\sqrt{k}}+\frac{2C_6\log(2T/\delta)}{\rho}+2(C_1C_4+A_2)\sum_{k=1}^{T}\eta_k^2,
\end{aligned}$$

from which we derive the stated inequality with probability $1-\delta$ (notice $\|\mathbf{w}_k-\mathbf{w}_1^*\|_2^2 \leq 2(\|\mathbf{w}_1^*\|_2^2 + C_4\sum_{j=1}^{k-1}\eta_j)$)

$$C_2 = 2\|\mathbf{w}_1^*\|_2^2 + \sum_{k=1}^{t_2}\frac{8C_7\eta_k\big(\|\mathbf{w}_1^*\|_2^2 + C_4\sum_{j=1}^{k-1}\eta_j\big)}{\sqrt{k}} + \frac{4\phi(\mathbf{w}_1^*)}{C_4C_5}\sum_{k=1}^{t_2}\eta_k^2\big(\|\mathbf{w}_1^*\|_2^2 + C_4\sum_{j=1}^{k-1}\eta_j\big) + \sum_{k=1}^{\infty}\frac{4C_7\eta_k}{\sqrt{k}} + \frac{4C_6}{\rho} + 4(C_1C_4+A_2)\sum_{k=1}^{\infty}\eta_k^2,$$

where we introduce (notice $C_{k,\delta} \leq C_7\sqrt{\log(T/\delta)}$)

$$C_7 = 2\kappa^2\big(2+\sqrt{2\log 12+4}\big)\max\Big\{C_p+1, 4^{-1}(C_P\|\mathbf{w}_1^*\|_2+3)^2\Big\}.$$

The proof is complete. ■

5.4 Proofs for General Convergence Rates

In this subsection, we prove Theorem 4 on the probabilistic convergence rates by taking a deduction analogous to the proof of Proposition 16. The difference is to apply Part (a) of Lemma 19 together with the bound of $\|\mathbf{w}_t\|_2$ established in Theorem 3 to control $\sum_{k=1}^t \xi_k$ in (5.23).

Proof of Theorem 4 According to Lemma 14 followed with union bounds, we know the existence of $\Omega_T^{(1)}$ with $\Pr(\Omega_T^{(1)}) \geq 1-\delta/3$ such that the following inequality holds with probability $1-\delta/3$ simultaneously for all $t=1,\ldots,T$ conditioned on $\Omega_T^{(1)}$

$$\big\|\widetilde{F}'(\mathbf{w}_t;z_t) - \hat{F}_t'(\mathbf{w}_t;z_t)\big\|_2 \leq \frac{2\kappa^2\big(2+\sqrt{2\log(9T/\delta)}\big)}{\sqrt{t}}\big(C_p\|\mathbf{w}_t-\mathbf{w}_1^*\|_2 + 3 + C_p\|\mathbf{w}_1^*\|_2\big).$$

It then follows the following inequality conditioned on $\Omega_T^{(1)}$

$$\begin{aligned}&\sum_{t=1}^T \eta_t\big\langle \mathbf{w}_1^*-\mathbf{w}_t, \hat{F}_t'(\mathbf{w}_t;z_t) - \widetilde{F}'(\mathbf{w}_t;z_t)\big\rangle \mathbb{I}_{[\|\mathbf{w}_t-\mathbf{w}_1^*\|_2^2 \leq C_2\log(6T/\delta)]} \\ &\quad\leq \sum_{t=1}^T \eta_t\|\mathbf{w}_1^*-\mathbf{w}_t\|_2\Big\|\hat{F}_t'(\mathbf{w}_t;z_t) - \widetilde{F}'(\mathbf{w}_t;z_t)\Big\|_2 \mathbb{I}_{[\|\mathbf{w}_t-\mathbf{w}_1^*\|_2^2 \leq C_2\log(6T/\delta)]} \leq \widetilde{C}_{T,\delta}\sum_{t=1}^T\frac{\eta_t}{\sqrt{t}},\end{aligned} \tag{5.29}$$

where we introduce

$$\widetilde{C}_{T,\delta} = 2\kappa^2\sqrt{C_2}\big(2+\sqrt{2\log(9T/\delta)}\big)\big(C_p\sqrt{C_2}+3+C_p\|\mathbf{w}_1^*\|_2\big)\log(6T/\delta).$$

Introduce a sequence of random variables

$$\xi_t' = \eta_t\big\langle \mathbf{w}_1^*-\mathbf{w}_t, \widetilde{F}'(\mathbf{w}_t;z_t) - \nabla f(\mathbf{w}_t)\big\rangle \mathbb{I}_{[\|\mathbf{w}_t-\mathbf{w}_1^*\|_2^2 \leq C_2\log(6T/\delta)]}, \quad t=1,\ldots,T.$$

According to Schwartz's inequality, we derive

$$\begin{aligned}|\xi_t'| &\leq \eta_t\Big[\|\mathbf{w}_1^* - \mathbf{w}_t\|_2^2 + 4^{-1}\|\widetilde{F}'(\mathbf{w}_t; z_t) - \nabla f(\mathbf{w}_t)\|_2^2\Big]\mathbb{I}_{[\|\mathbf{w}_t - \mathbf{w}_1^*\|_2^2 \leq C_2 \log \frac{6T}{\delta}]} \\ &\leq \eta_t\Big[\|\mathbf{w}_t - \mathbf{w}_1^*\|_2^2 + 2^{-1}\|\widetilde{F}'(\mathbf{w}_t; z_t)\|_2^2 + 2^{-1}\|\nabla f(\mathbf{w}_t)\|_2^2\Big]\mathbb{I}_{[\|\mathbf{w}_t - \mathbf{w}_1^*\|_2^2 \leq C_2 \log \frac{6T}{\delta}]}.\end{aligned}$$

According to (5.16) and (5.26), it is clear that

$$\begin{aligned}\max\big\{\|\nabla f(\mathbf{w})\|_2, \|\widetilde{F}'(\mathbf{w}; z)\|_2\big\} &\leq 8\kappa^2\|\mathbf{w}\|_2 + \kappa \\ &\leq 8\kappa^2\|\mathbf{w} - \mathbf{w}_1^*\|_2 + 8\kappa^2\|\mathbf{w}_1^*\|_2 + \kappa.\end{aligned} \tag{5.30}$$

Therefore, there holds

$$|\xi_t'| \leq C_8\eta_t \log(6T/\delta), \quad \text{where } C_8 = C_2 + 2(8\kappa^2\|\mathbf{w}_1^*\|_2 + \kappa)^2 + 128\kappa^4 C_2.$$

It is clear that $\{\xi_t'\}$ is a martingale difference sequence and therefore we can apply Part (a) of Lemma 19 in the Appendix A to show the existence of $\Omega_T^{(2)}$ with $\Pr(\Omega_T^{(2)}) \geq 1 - \delta/3$ such that the following inequality holds conditioned on $\Omega_T^{(2)}$

$$\sum_{t=1}^T \xi_t' \leq C_8\sqrt{2\sum_{t=1}^T \eta_t^2 \log\frac{3}{\delta}}\log\frac{6T}{\delta}. \tag{5.31}$$

Theorem 3 implies the existence of $\Omega_T^{(3)}$ with $\Pr(\Omega_T^{(3)}) \geq 1 - \delta/3$ such that $\max_{1\leq \tilde{t}\leq T}\|\mathbf{w}_{\tilde{t}} - \mathbf{w}_1^*\|_2^2 \leq C_2\log(6T/\delta)$. According to (5.19), (5.22), (5.29) and (5.31), it is clear that the following inequality holds under the event $\Omega_T^{(1)}\cap\Omega_T^{(2)}\cap\Omega_T^{(3)}$ (note $\xi_t' = \eta_t\langle \mathbf{w}_1^* - \mathbf{w}_t, \widetilde{F}'(\mathbf{w}_t; z_t) - \nabla f(\mathbf{w}_t)\rangle$ in this case)

$$\sum_{t=1}^T \eta_t\langle \mathbf{w}_1^* - \mathbf{w}_t, \hat{F}_t'(\mathbf{w}_t; z_t)\rangle \leq \widetilde{C}_{T,\delta}\sum_{t=1}^T \frac{\eta_t}{\sqrt{t}} + C_8\log\frac{6T}{\delta}\sqrt{2\sum_{t=1}^T \eta_t^2\log\frac{3}{\delta}} + \sum_{t=1}^T \eta_t\big[f(\mathbf{w}_1^*) - f(\mathbf{w}_t)\big].$$

Plugging the above inequality back into (5.18) and noting $\Pr\big(\Omega_T^{(1)}\cap\Omega_T^{(2)}\cap\Omega_T^{(3)}\big) \geq 1 - \delta$, we derive the following inequality with probability at least $1 - \delta$

$$\begin{aligned}\|\mathbf{w}_{T+1} - \mathbf{w}_1^*\|_2^2 - \|\mathbf{w}_1^*\|_2^2 \leq 2\sum_{t=1}^T \eta_t\big(\phi(\mathbf{w}_1^*) - \phi(\mathbf{w}_t)\big) + 2(C_1C_4 + A_2)\sum_{t=1}^T \eta_t^2 \\ + 2\widetilde{C}_{T,\delta}\sum_{t=1}^T \frac{\eta_t}{\sqrt{t}} + 2C_8\log\frac{6T}{\delta}\sqrt{2\sum_{t=1}^T \eta_t^2\log\frac{3}{\delta}}.\end{aligned}$$

This combined with the convexity of ϕ establishes the stated inequality with probability $1 - \delta$. The proof is complete. ■

Proof of Corollary 5 We first prove Part (a). It is clear that the step sizes satisfy (3.2) and therefore Theorem 4 holds. Part (a) then follows from the standard inequality

$\sum_{t=1}^{T} t^{-\theta} \geq (1-\theta)^{-1}(T^{1-\theta}-1), \theta \in (0,1)$. We now turn to Part (b). It is clear that $\sum_{t=1}^{\infty} \eta_t \log^{\frac{1}{2}} t/\sqrt{t} \leq \eta_1 \sum_{t=1}^{\infty} \log^{\frac{1-\beta}{2}}(et)/t < \infty$ and $\sum_{t=1}^{\infty} \eta_t^2 < \infty$. Part (b) then follows from the inequality $\sum_{t=1}^{T} \left(t \log^{\beta}(et)\right)^{-\frac{1}{2}} \geq 2(\sqrt{T}-1)\log^{-\frac{\beta}{2}}(eT)$. The proof is complete. ■

5.5 Discussion of the Proof

We follow the arguments in Lei and Tang (2018) in our proofs. In this subsection, we give details on the similarity and difference between our proofs and those in Lei and Tang (2018). Both arguments first build a crude estimate $\|\mathbf{w}_t\|_2^2 = O\big(\sum_{k=1}^{t} \eta_k\big)$ and then refine it to an almost boundedness of iterates with high probability by martingale analysis. As in Lei and Tang (2018), we use the self-bounding property of loss functions to control a weighted summation of loss functions (5.4), which removes bounded gradient assumptions imposed in the literature. As in Lei and Tang (2018), we use the Bernstein inequality to control the martingale difference sequence $\{\xi_k\}$ defined in (5.23), and show that the conditional variance of this martingale can be offset by some negative terms in the one-step progress inequality.

A key difference is that we use approximately biased stochastic gradients in our algorithm, while the discussions (Lei and Tang, 2018) consider stochastic optimization with unbiased gradient estimates. Specifically, the approximate unbiased stochastic gradient is caused by replacing $p = P(y=1), \mathbb{E}[x|y=1]$ and $\mathbb{E}[x|y=-1]$ by its empirical counterparts at the present time t. To overcome this hindrance, we build high-probability bounds on $\left\|\widetilde{F}'(\mathbf{w}_t; z_t) - \hat{F}_t'(\mathbf{w}_t; z_t)\right\|_2$ (Lemma 14) by developing concentration inequalities for approximating conditional expectations and variances by their empirical counterparts (Lemma 15). We show that this approximate unbias does not affect the almost boundedness of iterates provided that the step size sequence satisfies an additional assumption $\sum_{t=1}^{\infty} \eta_t \sqrt{\log t}/t < \infty$. Another notable difference is that we show that the fast convergence rates can be derived for objective functions satisfying the quadratic functional growth, while the discussions in Lei and Tang (2018) require a stronger assumption on the strong convexity of objective functions.

As compared to the technical analysis, our main novelty is the development of a new stochastic optimization algorithm for AUC maximization. Previous reformulation of AUC maximization as a pointwise problem either introduces *additional dual variables* (Ying et al., 2016b; Liu et al., 2018) or uses a *non-convex* estimator of stochastic gradients (Natole et al., 2018). The former formulation requires to introduce an explicit boundedness constraint on $\mathbf{w}$ (Ying et al., 2016b; Liu et al., 2018), while the latter one requires the information of p, conditional expectation and is only guaranteed convergence rates in expectation (Natole et al., 2018). We develop a novel reformulation of AUC maximization as a pointwise problem which leads to a *convex estimator* of stochastic gradient. This key convexity is critical to handle the case when $\mathbf{w} \in \mathbb{R}^d$, as well as in both the algorithm design and the high-probability theoretical analysis.

6. Conclusion

In this paper, we presented a new stochastic gradient descent method for AUC maximization which can accommodate general penalty terms. Our algorithm can update the model parameter upon receiving individual data with favorable $\mathcal{O}(d)$ space and per-iteration time complexity, making it amenable for streaming data analysis. We established a high-probability convergence rate $\widetilde{\mathcal{O}}(1/\sqrt{T})$ for the general convex setting, and a fast convergence $\widetilde{\mathcal{O}}(1/T)$ for the cases of strongly convex regularizers and no regularization term (without strong convexity).

There are several directions for future work. Firstly, we focused on the square loss and it remains unclear to us on how to develop similar algorithms for general loss functions. Secondly, it would be very interesting to develop stochastic optimization algorithms for AUC maximization under nonlinear models. There are two possible approaches for developing nonlinear models for AUC maximization including the kernel trick and and deep neural networks. For the approach using the kernel trick, one could use the techniques of random feature (Rahimi and Recht, 2008) for RBF kernels and then apply the linear model in this paper. One can easily prove a similar saddle point formulation even for non-convex deep neural network, and develop stochastic primal-dual stochastic gradient decent algorithms (Nemirovski et al., 2009) for deep AUC maximization models. However, it is not clear on how to establish theoretical guarantees for the convergence of such algorithms as the objective function is generally non-convex.

Acknowledgments

The authors would like to thank the anonymous reviewers and the editor for their constructive comments and suggestions. This work is supported by National Science Foundation (NSF) grants IIS-1816227 and IIS-2008532.

Appendix A. Lemmas

In this section we provide some useful lemmas. Lemma 17 shows a self-bounding property for smooth and non-negative functions (Nesterov, 2013).

Lemma 17 *If $h : \mathbb{R}^d \to \mathbb{R}$ is non-negative and β-smooth, i.e., $\|\nabla h(\mathbf{w}) - \nabla h(\tilde{\mathbf{w}})\|_2 \leq \beta\|\mathbf{w} - \tilde{\mathbf{w}}\|_2$, then $\|\nabla h(\mathbf{w})\|_2^2 \leq 2\beta h(\mathbf{w})$ for all $\mathbf{w} \in \mathbb{R}^d$.*

Our discussion is also based on some concentration inequalities. Lemma 18 is the Hoeffding's inequality for vector-valued random variables (Boucheron et al., 2013).

Lemma 18 (Hoeffding's inequality) *Let $Z_1, \ldots, Z_n$ be a sequence of i.i.d. random variables taking values in $\mathbb{R}^d$ with $\|Z_i\|_2 \leq B$ for every i. Then, for any $0 < \delta < 1$, with probability $1 - \delta$ we have*

$$\Big\| \frac{1}{n} \sum_{i=1}^{n} \big[Z_i - \mathbb{E}[Z_i] \big] \Big\|_2 \leq \frac{B}{\sqrt{n}} \Big[2 + \sqrt{2 \log 1/\delta} \Big].$$

Part (a) of Lemma 19 is the Azuma-Hoeffding inequality for martingales with bounded increments (Hoeffding, 1963), and part (b) is a conditional Bernstein inequality using the conditional variance to quantify better the concentration behavior of martingales (Zhang, 2005).

Lemma 19 *Let $z_1, \ldots, z_n$ be a sequence of random variables such that z_k may depend on the previous random variables $z_1, \ldots, z_{k-1}$ for all $k = 1, \ldots, n$. Consider a sequence of functionals $\xi_k(z_1, \ldots, z_k), k = 1, \ldots, n$. Let $\sigma_n^2 = \sum_{k=1}^n \mathbb{E}_{z_k}\big[\big(\xi_k - \mathbb{E}_{z_k}[\xi_k]\big)^2\big]$ be the conditional variance and $\delta \in (0, 1)$.*

(a) Assume that $|\xi_k - \mathbb{E}_{z_k}[\xi_k]| \leq b_k$ for each k. With probability at least $1 - \delta$ we have

$$\sum_{k=1}^n \xi_k - \sum_{k=1}^n \mathbb{E}_{z_k}[\xi_k] \leq \Big(2\sum_{k=1}^n b_k^2 \log\frac{1}{\delta}\Big)^{\frac{1}{2}}. \tag{A.1}$$

(b) Assume that $\xi_k - \mathbb{E}_{z_k}[\xi_k] \leq b$ for each k and $\rho \in (0, 1]$. With probability at least $1 - \delta$ we have

$$\sum_{k=1}^n \xi_k - \sum_{k=1}^n \mathbb{E}_{z_k}[\xi_k] \leq \frac{\rho\sigma_n^2}{b} + \frac{b\log\frac{1}{\delta}}{\rho}. \tag{A.2}$$

Appendix B. Proof of Results in Section 5.1

Proof of Lemma 12 According to the first-order optimality condition in (2.11), we get

$$\eta_t \hat{F}_t'(\mathbf{w}_t; z_t) + \eta_t \Omega'(\mathbf{w}_{t+1}) + \mathbf{w}_{t+1} - \mathbf{w}_t = 0, \tag{B.1}$$

from which we derive

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 &= \langle \mathbf{w}_{t+1} - \mathbf{w}, \mathbf{w}_{t+1} - \mathbf{w}_t + \mathbf{w}_t - \mathbf{w}\rangle \\ &= -\eta_t\big\langle \mathbf{w}_{t+1} - \mathbf{w}, \hat{F}_t'(\mathbf{w}_t; z_t)\big\rangle + \eta_t\langle \mathbf{w} - \mathbf{w}_{t+1}, \Omega'(\mathbf{w}_{t+1})\rangle + \langle \mathbf{w}_{t+1} - \mathbf{w}, \mathbf{w}_t - \mathbf{w}\rangle.\end{aligned} \tag{B.2}$$

It follows from the definition of σ_Ω that

$$\begin{aligned}&\langle \mathbf{w} - \mathbf{w}_{t+1}, \Omega'(\mathbf{w}_{t+1})\rangle \leq \Omega(\mathbf{w}) - \Omega(\mathbf{w}_{t+1}) - 2^{-1}\sigma_\Omega\|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2 \\ &= \Omega(\mathbf{w}) - \Omega(\mathbf{w}_t) + \Omega(\mathbf{w}_t) - \Omega(\mathbf{w}_{t+1}) - 2^{-1}\sigma_\Omega\|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2 \\ &\leq \Omega(\mathbf{w}) - \Omega(\mathbf{w}_t) + \langle \mathbf{w}_t - \mathbf{w}_{t+1}, \Omega'(\mathbf{w}_t)\rangle - 2^{-1}\sigma_\Omega\big(\|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2 + \|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2^2\big).\end{aligned} \tag{B.3}$$

It can be directly checked that

$$\langle \mathbf{w}_{t+1} - \mathbf{w}, \mathbf{w}_t - \mathbf{w}\rangle = \frac{1}{2}\Big(\|\mathbf{w} - \mathbf{w}_t\|_2^2 + \|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2 - \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2\Big).$$

Plugging the above identity and (B.3) back into (B.2), we derive

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 \leq{}& \eta_t\big\langle \mathbf{w} - \mathbf{w}_t + \mathbf{w}_t - \mathbf{w}_{t+1}, \hat{F}_t'(\mathbf{w}_t; z_t)\big\rangle + \eta_t\Omega(\mathbf{w}) - \eta_t\Omega(\mathbf{w}_t) + \eta_t\langle \mathbf{w}_t - \mathbf{w}_{t+1}, \Omega'(\mathbf{w}_t)\rangle \\ &- 2^{-1}\eta_t\sigma_\Omega\|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2 + \frac{1}{2}\Big(\|\mathbf{w} - \mathbf{w}_t\|_2^2 + \|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2 - \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2\Big).\end{aligned} \tag{B.4}$$

According to the Schwartz's inequality, we know

$$\eta_t\big\langle \mathbf{w}_t-\mathbf{w}_{t+1}, \hat{F}_t'(\mathbf{w}_t;z_t)\big\rangle+\eta_t\langle \mathbf{w}_t-\mathbf{w}_{t+1}, \Omega'(\mathbf{w}_t)\rangle \leq \frac{1}{2}\|\mathbf{w}_t-\mathbf{w}_{t+1}\|_2^2+\eta_t^2\|\hat{F}_t'(\mathbf{w}_t;z_t)\|_2^2+\eta_t^2\|\Omega'(\mathbf{w}_t)\|_2^2.$$

Plugging the above inequality back into (B.4) gives

$$\begin{aligned}\|\mathbf{w}_{t+1}-\mathbf{w}\|_2^2-\|\mathbf{w}-\mathbf{w}_t\|_2^2 \leq 2\eta_t\big\langle \mathbf{w}-\mathbf{w}_t, \hat{F}_t'(\mathbf{w}_t;z_t)\big\rangle + 2\eta_t\big(\Omega(\mathbf{w})-\Omega(\mathbf{w}_t)\big)\\ -\eta_t\sigma_\Omega\|\mathbf{w}-\mathbf{w}_{t+1}\|_2^2+2\eta_t^2\|\hat{F}_t'(\mathbf{w}_t;z_t)\|_2^2+2\eta_t^2\|\Omega'(\mathbf{w}_t)\|_2^2. \quad \text{(B.5)}\end{aligned}$$

The stated bound then follows from Lemma 11, Assumption 1 and the definition of C_1. The proof is complete. ■

Proof of Corollary 13 Eq. (5.2) together with the convexity of $\hat{F}_t$ established in Lemma 11 implies

$$\begin{aligned}\|\mathbf{w}_{t+1}-\mathbf{w}\|_2^2-\|\mathbf{w}_t-\mathbf{w}\|_2^2 \leq 2\eta_t\big(\hat{F}_t(\mathbf{w};z_t)-\hat{F}_t(\mathbf{w}_t;z_t)\big)+2\eta_t\big(\Omega(\mathbf{w})-\Omega(\mathbf{w}_t)\big)\\ -\eta_t\sigma_\Omega\|\mathbf{w}-\mathbf{w}_{t+1}\|_2^2+2\eta_t^2\big(C_1\hat{F}_t(\mathbf{w}_t;z_t)+C_1\Omega(\mathbf{w}_t)+A_2\big). \quad \text{(B.6)}\end{aligned}$$

Taking $\mathbf{w}=0$ in (B.6) and using $\hat{F}_t(0;z_t)=p_t(1-p_t), \Omega(0)=0$, we get

$$\begin{aligned}&\|\mathbf{w}_{t+1}\|_2^2-\|\mathbf{w}_t\|_2^2\\ &\leq 2\eta_t\big(\hat{F}_t(0;z_t)-\hat{F}_t(\mathbf{w}_t;z_t)\big)+2\eta_t\big(\Omega(0)-\Omega(\mathbf{w}_t)\big)+2\eta_t^2\big(C_1\hat{F}_t(\mathbf{w}_t;z_t)+C_1\Omega(\mathbf{w}_t)+A_2\big)\\ &\leq 2\eta_t(C_1\eta_t-1)\big(\hat{F}_t(\mathbf{w}_t;z_t)+\Omega(\mathbf{w}_t)\big)+\eta_t/2+2\eta_t^2A_2 \quad \text{(B.7)}\\ &\leq \eta_t/2+C_1^{-1}A_2\eta_t,\end{aligned}$$

where the last inequality follows from $\hat{F}_t(\mathbf{w}_t;z_t)+\Omega(\mathbf{w}_t)\geq 0$ due to Lemma 11 and the assumption $0\leq\eta_t\leq(2C_1)^{-1}$. Taking a summation of the above inequality then shows

$$\|\mathbf{w}_{t+1}\|_2^2 \leq \big(C_1^{-1}A_2+2^{-1}\big)\sum_{k=1}^t\eta_k.$$

This establishes (5.3). Plugging the assumption $\eta_t\leq(2C_1)^{-1}$ into (B.7) gives

$$\eta_t\big(\hat{F}_t(\mathbf{w}_t;z_t)+\Omega(\mathbf{w}_t)\big)\leq\|\mathbf{w}_t\|_2^2-\|\mathbf{w}_{t+1}\|_2^2+\eta_t/2+C_1^{-1}A_2\eta_t.$$

Multiplying both sides by η_t, we derive

$$\begin{aligned}\eta_t^2\big(\hat{F}_t(\mathbf{w}_t;z_t)+\Omega(\mathbf{w}_t)\big)&\leq\eta_t\|\mathbf{w}_t\|_2^2-\eta_t\|\mathbf{w}_{t+1}\|_2^2+\eta_t^2/2+\eta_t^2C_1^{-1}A_2\\ &\leq\eta_t\|\mathbf{w}_t\|_2^2-\eta_{t+1}\|\mathbf{w}_{t+1}\|_2^2+\eta_t^2/2+\eta_t^2C_1^{-1}A_2,\end{aligned}$$

where we have used the assumption $\eta_{t+1}\leq\eta_t$. Taking a summation of the above inequality further yields

$$\sum_{k=1}^t\eta_k^2\big(\hat{F}_k(\mathbf{w}_k;z_k)+\Omega(\mathbf{w}_k)\big)\leq\big(C_1^{-1}A_2+2^{-1}\big)\sum_{k=1}^t\eta_k^2$$

We now turn to (5.5). Plugging the assumption $\eta_t \leq (2C_1)^{-1}$ into (B.7) and multiplying both sides by η_t^{-1}, we derive

$$\hat{F}'(\mathbf{w}_t; z_t) + \Omega(\mathbf{w}_t) \leq \eta_t^{-1}\big(\|\mathbf{w}_t\|_2^2 - \|\mathbf{w}_{t+1}\|_2^2\big) + 2^{-1} + C_1^{-1}A_2.$$

Taking a summation of the above inequality implies

$$\begin{aligned}
\sum_{k=1}^{t}\big(\hat{F}_k(\mathbf{w}_k; z_k) + \Omega(\mathbf{w}_k)\big) &\leq tC_4 + \sum_{k=1}^{t}\eta_k^{-1}\big(\|\mathbf{w}_k\|_2^2 - \|\mathbf{w}_{k+1}\|_2^2\big) \\
&\leq tC_4 + \sum_{k=2}^{t}\|\mathbf{w}_k\|_2^2(\eta_k^{-1} - \eta_{k-1}^{-1}) + \eta_1^{-1}\|\mathbf{w}_1\|_2^2 \\
&\leq tC_4 + \max_{1\leq\tilde{k}\leq t}\|\mathbf{w}_{\tilde{k}}\|_2^2\sum_{k=2}^{t}(\eta_k^{-1} - \eta_{k-1}^{-1}) \\
&\leq tC_4 + C_4\eta_t^{-1}\sum_{k=1}^{t}\eta_k,
\end{aligned}$$

where the last inequality is due to (5.3). The proof is complete. ■

Appendix C. Proofs for Fast Convergence Rates

In this subsection, we prove Theorem 7 on convergence rates for ϕ with a quadratic functional growth. To this aim, we need to introduce some lemmas. The following lemma provides probabilistic bounds for approximating $\widetilde{F}'(\mathbf{w}_k; z_k)$ with $\hat{F}_k'(\mathbf{w}_k; z_k)$ for $\{\mathbf{w}_k\}$ produced by (2.11) with specific step sizes.

Lemma 20 *Suppose Assumption 1 holds. Let $\{\mathbf{w}_t\}_t$ be the sequence produced by (2.11) with $\eta_t = \frac{2}{\sigma_\phi t + 2\sigma_f + \sigma_\phi t_1}$, where $t_1 \geq 4C_1\sigma_\phi^{-1}$. Then, for any $k \leq T$ the following inequality holds with probability $1-\delta$*

$$\big\|\widetilde{F}'(\mathbf{w}_k; z_k) - \hat{F}_k'(\mathbf{w}_k; z_k)\big\|_2 \leq C_\delta\sqrt{\log(eT)}/\sqrt{k}, \tag{C.1}$$

where $C_\delta := 2\kappa^2\big(2 + \sqrt{2\log(3/\delta)}\big)\big(32\sqrt{2C_4\sigma_\phi^{-1}} + 3\big)$.

Proof Since $t_1 \geq 4C_1\sigma_\phi^{-1}$ we know $\eta_t \leq (2C_1)^{-1}$ and therefore Corollary 13 holds. It follows from the definition of η_t that

$$\sum_{k=1}^{t}\eta_k \leq 2\sigma_\phi^{-1}\sum_{k=1}^{t}(k+t_1)^{-1} \leq 2\sigma_\phi^{-1}\log(et). \tag{C.2}$$

This together with (5.3) shows

$$\|\mathbf{w}_t\|_2^2 \leq 2C_4\sigma_\phi^{-1}\log(et). \tag{C.3}$$

For all $k = 1, \ldots, T$, we can then apply Lemma 14 to derive the following inequality with probability $1 - \delta$

$$\big\|\widetilde{F}'(\mathbf{w}_k; z_k) - \hat{F}'_k(\mathbf{w}_k; z_k)\big\|_2 \leq 2\kappa^2\big(2 + \sqrt{2\log(3/\delta)}\big)\big(32\sqrt{2C_4\sigma_\phi^{-1}} + 3\big)\sqrt{\log(eT)}/\sqrt{k}.$$

The proof is complete with the introduction of C_δ. ■

The following lemma plays a fundamental role in our analysis. It shows that both $\|\mathbf{w}_t - \mathbf{w}_t^*\|_2^2$ and a weighted summation of $\phi(\mathbf{w}_k) - \phi(\mathbf{w}_k^*)$ can be controlled by a summation of martingale difference sequences. It is established by taking a weighted summation of the one-step progress inequality (5.2).

Lemma 21 *Suppose Assumption 1 and Assumption 2 hold. Let $\{\mathbf{w}_t\}_t$ be the sequence produced by (2.11) with $\eta_t = \frac{2}{\sigma_\phi t + 2\sigma_f + \sigma_\phi t_1}$ with $t_1 \geq 4C_1\sigma_\phi^{-1}$. Let $\delta \in (0,1)$ and $C_9 = 16(C_1C_4 + A_2)$. Then the following inequality holds with probability $1-\delta$ for all $t = 1, 2, \ldots, T$*

$$\frac{\sum_{k=1}^t (k + t_1 + 1)(\phi(\mathbf{w}_k) - \phi(\mathbf{w}_k^*))}{(t + t_1 + 1)(t + t_1 + 2)\sigma_\phi} + \|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^*\|_2^2 \leq \frac{(t_1+1)(t_1+2)\|\mathbf{w}_1 - \mathbf{w}_1^*\|_2^2}{(t + t_1 + 1)(t + t_1 + 2)} + \frac{4\sum_{k=1}^t (k + t_1 + 1)\xi_k}{(t + t_1 + 1)(t + t_1 + 2)\sigma_\phi} + \frac{2\log^2(eT)(2C_{\delta/T}^2 + C_9)}{(t + t_1 + 2)\sigma_\phi^2}. \quad \text{(C.4)}$$

Proof It follows from (5.2) that

$$\begin{aligned}\|\mathbf{w}_{k+1} - \mathbf{w}\|_2^2 - \|\mathbf{w} - \mathbf{w}_k\|_2^2 \leq{}& 2\eta_k\langle \mathbf{w} - \mathbf{w}_k, \hat{F}'_k(\mathbf{w}_k; z_k) - \widetilde{F}'(\mathbf{w}_k; z_k)\rangle + \\ &+ 2\eta_k\langle \mathbf{w} - \mathbf{w}_k, \widetilde{F}'(\mathbf{w}_k; z_k) - \nabla f(\mathbf{w}_k)\rangle + 2\eta_k\langle \mathbf{w} - \mathbf{w}_k, \nabla f(\mathbf{w}_k)\rangle + 2\eta_k\big(\Omega(\mathbf{w}) - \Omega(\mathbf{w}_k)\big) \\ &- \eta_k\sigma_\Omega\|\mathbf{w} - \mathbf{w}_{k+1}\|_2^2 + 2\eta_k^2\big(C_1\hat{F}_k(\mathbf{w}_k; z_k) + C_1\Omega(\mathbf{w}_k) + A_2\big).\end{aligned}$$

Taking $\mathbf{w} = \mathbf{w}_k^*$ in the above inequality and introducing the sequence of random variables $\{\xi_k\}_k$ as

$$\xi_k = \langle \mathbf{w}_k^* - \mathbf{w}_k, \widetilde{F}'(\mathbf{w}_k; z_k) - \nabla f(\mathbf{w}_k)\rangle, \quad k = 1, 2, \ldots, \quad \text{(C.5)}$$

we derive

$$\begin{aligned}&(1+\eta_k\sigma_\Omega)\|\mathbf{w}_{k+1} - \mathbf{w}_k^*\|_2^2 \leq \|\mathbf{w}_k - \mathbf{w}_k^*\|_2^2 + 2^{-1}\eta_k\sigma_\phi\|\mathbf{w}_k^* - \mathbf{w}_k\|_2^2 + 2\eta_k\sigma_\phi^{-1}\big\|\hat{F}'_k(\mathbf{w}_k, z_k) - \widetilde{F}'(\mathbf{w}_k; z_k)\big\|_2^2 \\ &+ 2\eta_k\xi_k + 2^{-1}\eta_k[\phi(\mathbf{w}_k^*) - \phi(\mathbf{w}_k)] - 3\eta_k\sigma_\phi\|\mathbf{w}_k - \mathbf{w}_k^*\|_2^2/2 + 2\eta_k^2\big(C_1\hat{F}_k(\mathbf{w}_k; z_k) + C_1\Omega(\mathbf{w}_k) + A_2\big),\end{aligned}$$

where we have used Schwartz's inequality

$$2\big\langle \mathbf{w}_k^* - \mathbf{w}_k, \hat{F}'_k(\mathbf{w}_k; z_k) - \widetilde{F}'(\mathbf{w}_k; z_k)\big\rangle \leq \frac{\sigma_\phi}{2}\|\mathbf{w}_k^* - \mathbf{w}_k\|_2^2 + \frac{2}{\sigma_\phi}\big\|\hat{F}'_k(\mathbf{w}_k; z_k) - \widetilde{F}'(\mathbf{w}_k; z_k)\big\|_2^2$$

and the following inequality due to Assumption 2

$$\begin{aligned}2\langle \mathbf{w}_k^* - \mathbf{w}_k, \nabla f(\mathbf{w}_k)\rangle + 2\big(\Omega(\mathbf{w}_k^*) - \Omega(\mathbf{w}_k)\big) &\leq \Big(\frac{1}{2} + \frac{3}{2}\Big)\big(\phi(\mathbf{w}_k^*) - \phi(\mathbf{w}_k)\big) \\ &\leq \frac{1}{2}\Big(\phi(\mathbf{w}_k^*) - \phi(\mathbf{w}_k)\Big) - \frac{3}{2}\sigma_\phi\|\mathbf{w}_k^* - \mathbf{w}_k\|_2^2.\end{aligned}$$

It then follows from $\|\mathbf{w}_{k+1} - \mathbf{w}^*_{k+1}\|_2 \leq \|\mathbf{w}_{k+1} - \mathbf{w}^*_k\|_2$ that

$$\frac{\eta_k(\phi(\mathbf{w}_k) - \phi(\mathbf{w}^*_k))}{2(1+\eta_k\sigma_\Omega)} + \|\mathbf{w}_{k+1} - \mathbf{w}^*_{k+1}\|_2 \leq \frac{1-\eta_k\sigma_\phi}{1+\eta_k\sigma_\Omega}\|\mathbf{w}_k - \mathbf{w}^*_k\|_2^2 + \frac{2\eta_k\|\hat{F}'_k(\mathbf{w}_k; z_k) - \widetilde{F}'(\mathbf{w}_k; z_k)\|_2^2}{(1+\eta_k\sigma_\Omega)\sigma_\phi} + \frac{2\eta_k\xi_k}{1+\eta_k\sigma_\Omega} + \frac{2\eta_k^2\big(C_1\hat{F}_k(\mathbf{w}_k; z_k) + C_1\Omega(\mathbf{w}_k) + A_2\big)}{1+\eta_k\sigma_\Omega}. \quad \text{(C.6)}$$

According to the step size choice $\eta_k = \frac{2}{\sigma_\phi k + 2\sigma_f + \sigma_\phi t_1}$ and $\sigma_\phi = \sigma_f + \sigma_\Omega$ we know

$$\frac{1-\sigma_\phi\eta_k}{1+\sigma_\Omega\eta_k} \leq \frac{1-\sigma_f\eta_k}{1+\sigma_\Omega\eta_k} = \frac{k+t_1}{k+t_1+2} \quad \text{and} \quad \frac{\eta_k}{1+\sigma_\Omega\eta_k} = \frac{2}{\sigma_\phi(k+t_1+2)}.$$

According to Lemma 20, we derive the following inequality with probability at least $1-\delta$ simultaneously for all $k = 1, \ldots, T$

$$\big\|\widetilde{F}'(\mathbf{w}_k; z_k) - \hat{F}'_k(\mathbf{w}_k; z_k)\big\|_2 \leq C_{\delta/T}\sqrt{\log(eT)}/\sqrt{k}.$$

Plugging the above two inequalities back into (C.6), we get the following inequality with probability $1-\delta$ for all $k = 1, \ldots, T$

$$\frac{\phi(\mathbf{w}_k) - \phi(\mathbf{w}^*_k)}{\sigma_\phi(k+t_1+2)} + \|\mathbf{w}_{k+1} - \mathbf{w}^*_{k+1}\|_2^2 \leq \frac{(k+t_1)\|\mathbf{w}_k - \mathbf{w}^*_k\|_2^2}{k+t_1+2} + \frac{4C^2_{\delta/T}\log(eT)}{\sigma_\phi^2 k(k+t_1+2)} + \frac{4\xi_k}{\sigma_\phi(k+t_1+2)} + \frac{4\eta_k\big(C_1\hat{F}_k(\mathbf{w}_k; z_k) + C_1\Omega(\mathbf{w}_k) + A_2\big)}{\sigma_\phi(k+t_1+2)}.$$

Multiplying both sides with $(k+t_1+2)(k+t_1+1)$ implies the following inequality with probability $1-\delta$ for all $k = 1, \ldots, T$

$$\frac{(k+t_1+1)(\phi(\mathbf{w}_k) - \phi(\mathbf{w}^*_k))}{\sigma_\phi} + (k+t_1+1)(k+t_1+2)\|\mathbf{w}_{k+1} - \mathbf{w}^*_{k+1}\|_2^2 \leq (k+t_1)(k+t_1+1)\|\mathbf{w}_k - \mathbf{w}^*_k\|_2^2 + \frac{4C^2_{\delta/T}\log(eT)(k+t_1+1)}{\sigma_\phi^2 k} + \frac{4(k+t_1+1)\xi_k}{\sigma_\phi} + \frac{4\eta_k(k+t_1+1)\big(C_1\hat{F}_k(\mathbf{w}_k; z_k) + C_1\Omega(\mathbf{w}_k) + A_2\big)}{\sigma_\phi}.$$

Taking a summation of the above inequality from $k=1$ to t shows the following inequality with probability $1-\delta$ for all $t = 1, \ldots, T$

$$\sigma_\phi^{-1}\sum_{k=1}^{t}(k+t_1+1)(\phi(\mathbf{w}_k) - \phi(\mathbf{w}^*_k)) + (t+t_1+1)(t+t_1+2)\|\mathbf{w}_{t+1} - \mathbf{w}^*_{t+1}\|_2^2 \leq (t_1+1)(t_1+2)\|\mathbf{w}_1 - \mathbf{w}^*_1\|_2^2 + \frac{4C^2_{\delta/T}\log(eT)}{\sigma_\phi^2}\sum_{k=1}^{t}\frac{k+t_1+1}{k} + 4\sigma_\phi^{-1}\sum_{k=1}^{t}(k+t_1+1)\xi_k + 16\sigma_\phi^{-2}\sum_{k=1}^{t}\big(C_1\hat{F}_k(\mathbf{w}_k; z_k) + C_1\Omega(\mathbf{w}_k) + A_2\big), \quad \text{(C.7)}$$

where we have used $\eta_k \leq 4/((k+t_1+1)\sigma_\phi)$. Since $t_1 \geq 4C_1\sigma_\phi^{-1}$ we know $\eta_t \leq (2C_1)^{-1}$ and therefore Corollary 13 holds. According to (C.2) and $\eta_t^{-1} \leq 2^{-1}\sigma_\phi(t+t_1+2)$, we know

$$\big(\sum_{k=1}^{t}\eta_k\big)\eta_t^{-1} \leq \big(2\sigma_\phi^{-1}\log(et)\big)\Big(2^{-1}\sigma_\phi(t+t_1+2)\Big) = (t+t_1+2)\log(et).$$

This together with (5.5) implies that

$$\sum_{k=1}^{t}\big(C_1\hat{F}_k(\mathbf{w}_k;z_k) + C_1\Omega(\mathbf{w}_k) + A_2\big)$$
$$\leq (C_1C_4+A_2)t + C_1C_4\big(\sum_{k=1}^{t}\eta_k\big)\eta_t^{-1} \leq (C_1C_4+A_2)\log(eT)(2t+t_1+2).$$

Plugging the above inequality into (C.7) and using $\sum_{k=1}^{t}k^{-1} \leq \log(eT)$ give the following inequality with probability $1-\delta$

$$\sigma_\phi^{-1}\sum_{k=1}^{t}(k+t_1+1)(\phi(\mathbf{w}_k)-\phi(\mathbf{w}_k^*)) + (t+t_1+1)(t+t_1+2)\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^*\|_2^2$$
$$\leq (t_1+1)(t_1+2)\|\mathbf{w}_1-\mathbf{w}_1^*\|_2^2 + \frac{4C_{\delta/T}^2\log(eT)}{\sigma_\phi^2}\big(t+(t_1+1)\log(eT)\big)$$
$$+ 4\sigma_\phi^{-1}\sum_{k=1}^{t}(k+t_1+1)\xi_k + C_9\sigma_\phi^{-2}\log(eT)(2t+t_1+2).$$

We can get the stated bound by dividing both sides by $(t+t_1+1)(t+t_1+2)$ and noting that

$$4C_{\delta/T}^2\log(eT)\big(t+(t_1+1)\log(eT)\big)+C_9\log(eT)(2t+t_1+2) \leq 2(t+t_1+1)\log^2(eT)\big(2C_{\delta/T}^2+C_9\big).$$

The proof is complete. ■

To tackle the martingale difference sequence $\{\xi_k\}_k$ in (C.4), we need to control the magnitudes and variances which are established in the following lemma.

Lemma 22 *Let Assumption 1 and Assumption 2 hold. Let $\{\mathbf{w}_t\}_t$ be the sequence produced by (2.11) with $\eta_t = \frac{2}{\sigma_\phi t+2\sigma_f+\sigma_\phi t_1}$, where $t_1 \geq 4C_1\sigma_\phi^{-1}$. Let $\{\xi_k\}_{k=1}^t$ be defined by (C.5). Then for all $k \leq T$ we have*

$$|\xi_k| \leq C_{10}\log(eT) \quad \textit{and} \quad \mathbb{E}_{z_k}\big[\big(\xi_k - \mathbb{E}_{z_k}[\xi_k]\big)^2\big] \leq C_1\phi(\mathbf{w}_k)\|\mathbf{w}_k^*-\mathbf{w}_k\|_2^2,$$

where $C_{10} = 34\kappa^2C_4\sigma_\phi^{-1} + 2\kappa\|\mathbf{w}_1^*\|_2 + (8\kappa\|\mathbf{w}_1^*\|_2+1)^2$.

Proof It follows from the inequality $\|\mathbf{w}_k - \mathbf{w}_k^*\|_2 \leq \|\mathbf{w}_k - \mathbf{w}_1^*\|_2$ and (5.30) that

$$\begin{aligned}\langle \mathbf{w}_k^* - \mathbf{w}_k, \widetilde{F}'(\mathbf{w}_k; z_k) - \nabla f(\mathbf{w}_k)\rangle &\leq \|\mathbf{w}_1^* - \mathbf{w}_k\|_2 \big(\|\widetilde{F}'(\mathbf{w}_k; z_k)\|_2 + \|\nabla f(\mathbf{w}_k)\|_2\big) \\ &\leq 2\big(\|\mathbf{w}_1^*\|_2 + \|\mathbf{w}_k\|_2\big)\big(8\kappa^2\|\mathbf{w}_k\|_2 + \kappa\big) \\ &= 16\kappa^2\|\mathbf{w}_k\|_2^2 + 2\kappa\|\mathbf{w}_1^*\|_2 + 2\|\mathbf{w}_k\|_2(8\kappa^2\|\mathbf{w}_1^*\|_2 + \kappa) \\ &\leq 17\kappa^2\|\mathbf{w}_k\|_2^2 + 2\kappa\|\mathbf{w}_1^*\|_2 + (8\kappa\|\mathbf{w}_1^*\|_2 + 1)^2 \\ &\leq 34\kappa^2 C_4\sigma_\phi^{-1}\log(ek) + 2\kappa\|\mathbf{w}_1^*\|_2 + (8\kappa\|\mathbf{w}_1^*\|_2 + 1)^2 \leq C_{10}\log(eT),\end{aligned}$$

where we have used (C.3).

It is clear from Proposition 1 that $\mathbb{E}_{z_k}[\xi_k] = 0$ and therefore it follows from $\mathbb{E}[(\xi - \mathbb{E}[\xi])^2] \leq \mathbb{E}[\xi^2]$ for any real-valued random variables ξ that

$$\begin{aligned}\mathbb{E}_{z_k}\big[\big(\xi_k - \mathbb{E}_{z_k}[\xi_k]\big)^2\big] = \mathbb{E}_{z_k}[\xi_k^2] &\leq \mathbb{E}_{z_k}\big[\langle \mathbf{w}_k^* - \mathbf{w}_k, \widetilde{F}'(\mathbf{w}_k; z_k)\rangle^2\big] \\ &\leq \|\mathbf{w}_k^* - \mathbf{w}_k\|_2^2 \mathbb{E}_{z_k}[\|\widetilde{F}'(\mathbf{w}_k; z_k)\|_2^2] \leq \|\mathbf{w}_k^* - \mathbf{w}_k\|_2^2 C_1 f(\mathbf{w}_k) \\ &\leq C_1\phi(\mathbf{w}_k)\|\mathbf{w}_k - \mathbf{w}_k^*\|_2^2.\end{aligned}$$

where we have used $\mathbb{E}_{z_k}[\|\widetilde{F}'(\mathbf{w}_k; z_k)\|_2^2] \leq C_1\mathbb{E}_{z_k}[\widetilde{F}(\mathbf{w}_k; z_k)] = C_1 f(\mathbf{w}_k)$ which can be shown analogously to the proof of Lemma 11. The proof is complete. ■

We are now ready to prove Theorem 7. Our key idea is to apply Part (b) of Lemma 19 in the Appendix to show that $\sum_{k=1}^t (k + t_1 + 1)\xi_k$ can be controlled by $\sum_{k=1}^t \big(\phi(\mathbf{w}_k) - \phi(\mathbf{w}_k^*)\big)(k + t_1 + 1)$, which can be offset by the first term of (C.4). Then we can apply the induction strategy to derive the stated bound.

Proof of Theorem 7 Since $t_1 \geq 32C_1\sigma_\phi^{-1}\log\frac{2T}{\delta}$ and $T \geq 2$, we know $t_1 \geq 4C_1\sigma_\phi^{-1}$ and therefore Lemmas 20, 21, 22 hold. According to Lemma 21, there exists a set $\Omega_T^{(1)} = \{(z_1, \ldots, z_T)\}$ with $\Pr(\Omega_T^{(1)}) \geq 1 - \delta/2$ such that for all $(z_1, \ldots, z_T) \in \Omega_T^{(1)}$ we have

$$\begin{aligned}\frac{\sum_{k=1}^t (k + t_1 + 1)(\phi(\mathbf{w}_k) - \phi(\mathbf{w}_k^*))}{(t + t_1 + 1)(t + t_1 + 2)\sigma_\phi} + \|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^*\|_2^2 &\leq \frac{(t_1 + 1)(t_1 + 2)\|\mathbf{w}_1^*\|_2^2}{(t + t_1 + 1)(t + t_1 + 2)} + \\ &+ \frac{4\sum_{k=1}^t (k + t_1 + 1)\xi_k}{(t + t_1 + 1)(t + t_1 + 2)\sigma_\phi} + \frac{2\log^2(eT)(2C_{\delta/(2T)}^2 + C_9)}{(t + t_1 + 2)\sigma_\phi^2}. \quad \text{(C.8)}\end{aligned}$$

According to Lemma 22, we know the following inequalities for $k = 1, \ldots, t$

$$\begin{aligned}|(k + t_1 + 1)\xi_k| &\leq C_{10}(t + t_1 + 1)\log(eT) \\ \mathbb{E}_{z_k}\big[\big((k + t_1 + 1)\xi_k - \mathbb{E}_{z_k}[(k + t_1 + 1)\xi_k]\big)^2\big] &\leq (k + t_1 + 1)^2 C_1\phi(\mathbf{w}_k)\|\mathbf{w}_k^* - \mathbf{w}_k\|_2^2.\end{aligned}$$

Let $\rho \in (0, 1]$ to be fixed later. It then follows from Part (b) of Lemma 19 the following inequality with probability $1 - \delta/(2T)$

$$\sum_{k=1}^t (k+t_1+1)\xi_k \leq \frac{C_1\rho\sum_{k=1}^t \phi(\mathbf{w}_k)(k + t_1 + 1)^2\|\mathbf{w}_k^* - \mathbf{w}_k\|_2^2}{C_{10}(t + t_1 + 1)\log(eT)} + \frac{C_{10}(t + t_1 + 1)\log(eT)\log\frac{2T}{\delta}}{\rho}. \quad \text{(C.9)}$$

By the union bounds of probability, we know the existence of $\Omega_T^{(2)} = \{(z_1, \ldots, z_T)\}$ with probability $\Pr(\Omega_T^{(2)}) \geq 1 - \delta/2$ such that (C.9) holds under the event $\Omega_T^{(2)}$ simultaneously for all $t = 1, \ldots, T$. In the remainder of the proof, we always assume that $\Omega_T^{(1)} \cap \Omega_T^{(2)}$ holds (with probability $1 - \delta$), and show by induction that $\|\mathbf{w}_{\tilde{t}+1} - \mathbf{w}_{\tilde{t}+1}^*\|_2^2 \leq C_{T,\delta}/(\tilde{t} + t_1 + 2)$ for all $\tilde{t} = 0, 1, \ldots, T-1$ conditioned on $\Omega_T^{(1)} \cap \Omega_T^{(2)}$, where we introduce

$$C_{T,\delta} = \max\Big\{2(t_1+1)\|\mathbf{w}_1^*\|_2^2 + \frac{3t_1\phi(\mathbf{w}^*)}{2\sigma_\phi} + \frac{4\log^2(eT)(2C_{\delta/(2T)}^2 + C_9)}{\sigma_\phi^2}, \frac{C_{10}t_1\log(eT)}{4C_1}\Big\}$$

and $\rho = \frac{C_{10}t_1\log(eT)}{4C_1C_{T,\delta}}$. It is clear that $\rho \leq 1$. The case with $\tilde{t} = 0$ is clear from the definition of $C_{T,\delta}$. We now show $\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^*\|_2^2 \leq C_{T,\delta}/(t + t_1 + 2)$ under the induction assumption

$$\|\mathbf{w}_{\tilde{t}+1} - \mathbf{w}_{\tilde{t}+1}^*\|_2^2 \leq C_{T,\delta}/(\tilde{t} + t_1 + 2) \tag{C.10}$$

for $\tilde{t} = 0, 1, \ldots, t-1$.

Plugging the induction assumption (C.10) into (C.9) gives ($\phi(\mathbf{w}_k^*)$ is the same for all k)

$$\begin{aligned}
&\sum_{k=1}^{t}(k+t_1+1)\xi_k \leq \frac{C_1\rho C_{T,\delta}\sum_{k=1}^t\phi(\mathbf{w}_k)(k+t_1+1)}{C_{10}(t+t_1+1)\log(eT)} + \frac{C_{10}(t+t_1+1)\log(eT)\log\frac{2T}{\delta}}{\rho} \\
&\leq \frac{t_1\sum_{k=1}^t\big(\phi(\mathbf{w}_k)-\phi(\mathbf{w}_k^*)\big)(k+t_1+1)}{4(t+t_1+1)} + \frac{t_1\phi(\mathbf{w}_k^*)\sum_{k=1}^t(k+t_1+1)}{4(t+t_1+1)} + \frac{4C_1(t+t_1+1)C_{T,\delta}\log\frac{2T}{\delta}}{t_1} \\
&\leq \frac{t_1\sum_{k=1}^t\big(\phi(\mathbf{w}_k)-\phi(\mathbf{w}_k^*)\big)(k+t_1+1)}{4(t+t_1+1)} + \frac{3t_1\phi(\mathbf{w}_k^*)(t+t_1+1)}{16} + \frac{4C_1(t+t_1+1)C_{T,\delta}\log\frac{2T}{\delta}}{t_1},
\end{aligned}$$

where the second inequality is due to the definition of ρ and the last inequality is due to $\sum_{k=1}^t(k+t_1+1) \leq \frac{3(t+t_1+1)^2}{4}$.

Plugging the above inequality back into (C.8) yields the following inequality

$$\begin{aligned}
&\Big(1 - \frac{t_1}{t+t_1+1}\Big)\frac{\sum_{k=1}^t(k+t_1+1)(\phi(\mathbf{w}_k)-\phi(\mathbf{w}_k^*))}{(t+t_1+1)(t+t_1+2)\sigma_\phi} + \|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^*\|_2^2 \\
&\leq \frac{(t_1+1)\|\mathbf{w}_1^*\|_2^2}{t+t_1+2} + \frac{3t_1\phi(\mathbf{w}^*)}{4\sigma_\phi(t+t_1+2)} + \frac{16C_1C_{T,\delta}\log\frac{2T}{\delta}}{t_1(t+t_1+2)\sigma_\phi} + \frac{2\log^2(eT)(2C_{\delta/2T}^2+C_9)}{(t+t_1+2)\sigma_\phi^2} \\
&\leq \frac{(t_1+1)\|\mathbf{w}_1^*\|_2^2}{t+t_1+2} + \frac{3t_1\phi(\mathbf{w}^*)}{4\sigma_\phi(t+t_1+2)} + \frac{C_{T,\delta}}{2(t+t_1+2)} + \frac{2\log^2(eT)(2C_{\delta/2T}^2+C_9)}{(t+t_1+2)\sigma_\phi^2},
\end{aligned} \tag{C.11}$$

where the last inequality is due to $t_1 \geq 32C_1\sigma_\phi^{-1}\log\frac{2T}{\delta}$. By the definition of $C_{T,\delta}$, it is clear that the right-hand side of (C.11) is less than or equal to $\frac{C_{T,\delta}}{t+t_1+2}$. Therefore, we finish the induction process and show (C.10) for $\tilde{t} = t$.

We now prove the second inequality of (3.5). It follows from the convexity of ϕ and (C.11) that

$$\phi(\bar{\mathbf{w}}_t^{(2)}) - \phi(\mathbf{w}_1^*) \le \Big(\sum_{k=1}^{t}(k+t_1+1)\Big)^{-1}\Big(\sum_{k=1}^{t}(k+t_1+1)\big(\phi(\mathbf{w}_k)-\phi(\mathbf{w}^*)\big)\Big)$$
$$\le \frac{2\sigma_\phi(t+t_1+1)^2}{t(t+1)(t+2t_1+3)}\Big((t_1+1)\|\mathbf{w}_1^*\|_2^2 + \frac{3t_1\phi(\mathbf{w}^*)}{4\sigma_\phi} + \frac{C_{T,\delta}}{2} + \frac{2\log^2(eT)(2C_{\delta/2T}^2+C_9)}{\sigma_\phi^2}\Big).$$

The second inequality of (3.5) then follows. The proof is complete. ■

References

A. Agarwal, M. Wainwright, P. Bartlett, and P. Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pages 1–9, 2009.

S. Agarwal. Surrogate regret bounds for the area under the roc curve via strongly proper losses. In *Conference on Learning Theory*, pages 338–353, 2013.

S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the roc curve. *Journal of Machine Learning Research*, 6(Apr):393–425, 2005.

M. Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10(4):1116–1135, 2000.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Advances in neural information processing systems*, pages 773–781, 2013.

L. Bottou and Y. Cun. Large scale online learning. In *Advances in neural information processing systems*, 2004.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

A. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

R. Caruana, T. Joachims, and L. Backstrom. Kdd-cup 2004: results and analysis. *ACM SIGKDD Explorations Newsletter*, 6(2):95–108, 2004.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of u-statistics. *Annals of Statistics*, pages 844–874, 2008.

C. Cortes and M. Mohri. Auc optimization vs. error rate minimization. In *Advances in neural information processing systems*, pages 313–320, 2004.

G. Denevi, C. Ciliberto, R. Grazzi, and M. Pontil. Learning-to-learn stochastic gradient descent with biased regularization. In *International Conference on Machine Learning*, pages 1566–1575, 2019.

J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(Dec):2899–2934, 2009.

T. Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

W. Gao and Z.-H. Zhou. On the consistency of AUC pairwise optimization. In *International Joint Conferences on Artificial Intelligence*, pages 939–945, 2015.

W. Gao, R. Jin, S. Zhu, and Z.-H. Zhou. One-pass AUC optimization. In *International Conference on Machine Learning*, pages 906–914, 2013.

H. Güvenir and M. Kurtcephe. Ranking instances by maximizing the area under ROC curve. *IEEE Transactions on Knowledge and Data Engineering*, 25(10):2356–2366, 2013.

J. Hanley and B. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

A. Herschtal and B. Raskutti. Optimising area under the ROC curve using gradient descent. In *International Conference on Machine Learning*, page 49. ACM, 2004.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

X. Jiang and Y. Zhou. Dissecting android malware: Characterization and evolution. In *IEEE Symposium on Security and Privacy*, pages 95–109. IEEE, 2012.

T. Joachims. A support vector method for multivariate performance measures. In *International Conference on Machine Learning*, pages 377–384. ACM, 2005.

P. Kar, B. Sriperumbudur, P. Jain, and H. Karnick. On the generalization ability of online learning algorithms for pairwise loss functions. In *International Conference on Machine Learning*, pages 441–449, 2013.

W. Kotlowski, K. Dembczynski, and E. Huellermeier. Bipartite ranking through minimization of univariate loss. In *International Conference on Machine Learning*, pages 1113–1120, 2011.

Y. Lei and K. Tang. Stochastic composite mirror descent: Optimal bounds with high probabilities. In *Advance in Neural Information Processing Systems*, pages 1524–1534, 2018.

M. Liu, X. Zhang, Z. Chen, X. Wang, and T. Yang. Fast stochastic AUC maximization with o(1/n)-convergence rate. In *International Conference on Machine Learning*, pages 3195–3203, 2018.

A. Maurer and M. Pontil. Estimating weighted areas under the ROC curve. *Advances in Neural Information Processing Systems*, 33, 2020.

H. Narasimhan and S. Agarwal. Support vector algorithms for optimizing the partial area under the roc curve. *Neural Computation*, 29(7):1919–1963, 2017.

M. Natole, Y. Ying, and S. Lyu. Stochastic proximal algorithms for AUC maximization. In *International Conference on Machine Learning*, pages 3707–3716, 2018.

I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, pages 1–39, 2018.

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

Y. Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

Y. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming*, 110(2):245–259, 2007.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

F. Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems*, pages 1116–1124, 2014.

F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

B. Palaniappan and F. Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, pages 1416–1424, 2016.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *International Conference on Machine Learning*, pages 449–456, 2012.

L. Rosasco, S. Villa, and B. Vũ. Convergence of stochastic proximal gradient algorithm. *arXiv preprint arXiv:1403.5074*, 2014.

C. Rudin and R. Schapire. Margin-based ranking and an equivalence between adaboost and rankboost. *Journal of Machine Learning Research*, 10:2193–2232, 2009.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

N. Srebro and A. Tewari. Stochastic optimization for machine learning. *ICML Tutorial*, 2010.

D. Wang, D. Irani, and C. Pu. Evolutionary study of web spam: Webb spam corpus 2011 versus webb spam corpus 2006. In *Collaborative Computing: Networking, Applications and Worksharing*, pages 40–49. IEEE, 2012a.

M. Wang, J. Liu, and E. Fang. Accelerating stochastic composition optimization. In *Advances in Neural Information Processing Systems*, pages 1714–1722, 2016.

M. Wang, E. Fang, and H. Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161 (1-2):419–449, 2017.

Y. Wang, R. Khardon, D. Pechyony, and R. Jones. Generalization bounds for online learning algorithms with pairwise loss functions. In *Conference on Learning Theory*, volume 23, pages 13–1, 2012b.

Y. Ying and D.-X. Zhou. Online pairwise learning algorithms. *Neural computation*, 28(4): 743–777, 2016.

Y. Ying, L. Wen, and S. Lyu. Stochastic online AUC maximization. In *Advances in Neural Information Processing Systems*, 2016a.

Y. Ying, L. Wen, and S. Lyu. Stochastic online AUC maximization. In *Advances in Neural Information Processing Systems*, pages 451–459, 2016b.

T. Zhang. Data dependent concentration bounds for sequential prediction algorithms. In *Conference on Learning Theory*, pages 173–187, 2005.

X. Zhang, A. Saha, and S. V. N. Vishwanathan. Smoothing multivariate performance measures. *Journal of Machine Learning Research*, 13:3623–3680, 2012.

P. Zhao, S. Hoi, R. Jin, and T. Yang. Online AUC maximization. In *International Conference on Machine Learning*, pages 233–240, 2011.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, pages 928–936, 2003.