COMMUNICATIONS ON
PURE AND APPLIED ANALYSIS
Volume **19**, Number **8**, August **2020**

doi:10.3934/cpaa.2020188

pp. **4191–4212**

STOCHASTIC AUC OPTIMIZATION WITH GENERAL LOSS

Zhenhuan Yang

Department of Mathematics and Statistics, State University of New York at Albany
Albany, NY 12206, USA

Wei Shen

Department of Mathematics, Hong Kong Baptist University
Kowloon Tong, Kowloon, Hong Kong, China

Yiming Ying*

Department of Mathematics and Statistics, State University of New York at Albany
Albany, NY 12206, USA

Xiaoming Yuan

Department of Mathematics, The University of Hong Kong
Hong Kong, China

Abstract. Recently, there is considerable work on developing efficient stochastic optimization algorithms for AUC maximization. However, most of them focus on the least square loss which may be not the best option in practice. The main difficulty for dealing with the general convex loss is the pairwise nonlinearity w.r.t. the sampling distribution generating the data. In this paper, we use Bernstein polynomials to uniformly approximate the general losses which are able to decouple the pairwise nonlinearity. In particular, we show that this reduction for AUC maximization with a general loss is equivalent to a weakly convex (nonconvex) min-max formulation. Then, we develop a novel SGD algorithm for AUC maximization with per-iteration cost linearly w.r.t. the data dimension, making it amenable for streaming data analysis. Despite its non-convexity, we prove its global convergence by exploring the appealing convexity-preserving property of Bernstein polynomials and the intrinsic structure of the min-max formulation. Experiments are performed to validate the effectiveness of the proposed approach.

1. **Introduction.** Area under the ROC curve (AUC) [2, 6, 11, 14] is a widely used metric for measuring classification performance in imbalanced classification and bipartite ranking. In imbalanced classification, the instances in one class are much more than the other class. Imbalanced data sets exist in many real-world domains such as fraud detection, information retrieval and medical diagnosis. It is of fundamental importance to develop efficient optimization algorithms for analyzing streaming data which is prevalent at the big data era.

There are considerable efforts on developing batch (offline) algorithms for AUC maximization, which use the entire data once, including the cutting plane algorithm

2010 *Mathematics Subject Classification.* Primary: 58F15, 58F17; Secondary: 53C35.

Key words and phrases. Stochastic optimization, AUC maximization, Bernstein polynomial.

This work was completed when Wei Shen was a visiting student at SUNY Albany. Yiming Ying is supported by the National Science Foundation (NSF, Grant IIS1816227).

* Corresponding author.

[16] and gradient descent methods [3, 15, 35]. These algorithms have convergence rates of $\mathcal{O}\big(\min\big(1/\varepsilon, 1/\sqrt{\lambda\varepsilon}\big)\big)$ to achieve precision ε which, however, needs high per-iteration cost $\mathcal{O}(nd)$. Here, λ, n, d are the regularization parameter, the number of examples, and the dimension of the data, respectively. Such algorithms are not suitable for analyzing massive streaming data due to the expensive per-iteration cost.

Stochastic optimization algorithms such as SGD [1, 18, 19, 28, 29, 30, 34, 33] are iterative and incremental in nature and process each new sample (input) with a computationally cheap update, making them amenable for large-scale streaming data analysis. However, most of existing studies focus on classification error where the objective function is linear w.r.t. the sampling distribution. This means that the expectation in the expected risk is taken w.r.t. a single data point. In contrast, the problem of AUC maximization involves the expectation of a *pairwise* loss function which depends on pairs of data points which makes the direct employment of standard SGD infeasible.

The studies [17, 31, 34, 36] developed SGD or online gradient descent algorithms for AUC maximization. Such appealing algorithms do gradient descent based on gradient of the local error which compares the current example with all previous ones. As a result, one needs to access previous examples which leads to expensive space and per-iteration complexity of $\mathcal{O}(td)$ for d-dimensional data at iteration t. Although this problem is partially mitigated by the use of buffers with a fixed size B, this reduction is not necessarily an ideal approach. The work [12] followed the same approach and noticed that such algorithms for the least square loss only need to update the covariance matrix of the training data with per-iteration complexity $\mathcal{O}(d^2)$, which could be not scalable well to high-dimensional data.

The recent work [32, 22, 21] used the min-max reformulation of AUC maximization with the least square loss. The main idea is to reduce the double integral w.r.t. pairs of examples in the original objective function to a single integral w.r.t. an individual example by introducing auxiliary variables. However, one shortcoming of the above studies is that such methods depend critically on the structure of the least square loss and can not apply to the general losses such as logistic loss and hinge one. This largely limits the practical applications of AUC optimization algorithms since the least square loss is arguably not the best suitable loss in practice. The very recent work [20] considered AUC maximization with deep neural networks associated with the least square loss, which resutls in a nonconvex-concave minmax problem.

In this paper, we make efforts to develop novel SGD-type algorithms for AUC maximization with a general loss. In particular, we first propose to use Bernstein polynomials [25, 26] to uniformly approximate the general loss. Then, we derive its equivalent (non-convex) min-max formulation which removes the pairwise structure in the original AUC objective function. We show that this non-convex min-max formulation is weakly convex [10, 24] in the primal variables and develop novel SGD-type algorithms inspired by the recent work [27]. In contrast to the local convergence proved in [27], we are able to show that our novel algorithms enjoy the global convergence. The novel idea is the introduction of proximal terms only on partial primal variables instead of all of them in our algorithmic design, and an appealing relation between the original AUC objective function, which is convex due to the convexity-preserving property of Bernstein polynomials, and the duality gap

arising from the special structure of the min-max formulation (see more discussions in Section 3).

2. **AUC maximization with general loss and min-max reformulation.** The AUC score [14] has an elegant probabilistic formulation. Specifically, suppose $z = (x, y)$ and $z' = (x', y')$ are independently drawn from an unknown (sampling) distribution $\mathcal{P}$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y} = \{\pm 1\}$. Then, the AUC score is the probability of a positive sample ranking higher than a negative sample (e.g., [5, 14]) which is given by

$$\mathrm{AUC}(\mathbf{w}) = \Pr(\langle \mathbf{w}, x\rangle \geq \langle \mathbf{w}, x'\rangle | y=1, y'=-1) = \mathbb{E}\big[\mathbb{I}_{[\langle \mathbf{w}, x-x'\rangle \geq 0]} \big| y=1, y'=-1\big], \quad (2.1)$$

where the expectation is w.r.t. (z, z'). Hence, maximizing $\mathrm{AUC}(\mathbf{w})$ is equivalent to minimizing $1 - \mathrm{AUC}(\mathbf{w})$ which is given by $\mathbb{E}\big[\mathbb{I}_{[\langle \mathbf{w}, x-x'\rangle < 0]} \big| y = 1, y' = -1\big]$. It involves a *pairwise loss* $\mathbb{I}_{[\langle \mathbf{w}, x-x'\rangle < 0]} \mathbb{I}_{[y=1]} \mathbb{I}_{[y'=-1]}$, i.e. the loss depends on a pair of examples (x, y) and (x', y'). In practice, one often replaces the indicator function $\mathbb{I}_{[\cdot]}$ by a *convex surrogate loss* $\ell : \mathbb{R} \to \mathbb{R}^+$ which satisfies $\mathbb{I}_{[\langle \mathbf{w}, x-x'\rangle < 0]} \leq \ell(\langle \mathbf{w}, x - x'\rangle)$. It can be any convex loss such as the hinge loss $\ell(s) = (1-s)_+$ or logistic regression loss $\ell(s) = \log(1 + e^{-s})$. One can find appealing results in [13] on how statistical consistency is related to the choice of different loss functions.

Now AUC maximization can be equivalently formulated as

$$\min_{\|\mathbf{w}\| \leq R} \big\{ g(\mathbf{w}) := \mathbb{E}\big[\ell(\mathbf{w}^\top x - \mathbf{w}^\top x') \mathbb{I}_{[y=1]} \mathbb{I}_{[y'=-1]}\big] \big\}, \quad (2.2)$$

where the constant term $\frac{1}{\Pr(y=1)P(y'=-1)}$ in the original formulation is ignored.

Initial motivation for using Bernstein polynomials. We consider the (stochactic) online setting where individual data points $z = (x, y)$ are i.i.d from the distribution $\mathcal{P}$. The main difficulty for developing AUC optimization algorithms for streaming data is that the population (expected) risk in (2.2) depends on pairs of examples (z, z') which are statistically dependent as pairs of examples may share one common individual example. The work [32] showed, for the least square loss, that the original problem (2.2) is equivalent to a convex-concave (saddle) point problem [23] where the new objective function depends on only one individual example.

Following the same spirit, since the least square loss is a polynomial function of degree 2, one would naturally think of approximating the general loss ℓ by high-order polynomial functions and then expect an equivalently saddle point reformulation. One plain idea is to use m-th order Taylor polynomials (Taylor series) to approximate ℓ which, however, is not convex even if ℓ is convex. Instead, we propose to use the Bernstein polynomials (e.g., [25, 26]), useful tools from approximation theory, to uniformly approximate a convex loss function. Specifically, the Bernstein polynomial of degree m for a function $\varphi : [0,1] \to \mathbb{R}$ are defined, for any $u \in [0,1]$, by

$$B_m(\varphi; u) = \sum_{k=0}^{m} \varphi\Big(\frac{k}{m}\Big) \binom{m}{k} u^k (1-u)^{m-k} = \sum_{k=0}^{m} \binom{m}{k} \Delta^k \varphi(0)\, u^k, \quad (2.3)$$

where $\binom{m}{k}$ denotes the binomial coefficients and $\Delta^k \varphi(0) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} \varphi(\frac{j}{m})$ is the forward difference operator on φ at 0. If φ is Lipschitz continuous, then $B_m(\varphi; \cdot)$ converges uniformly to φ with a rate of $\mathcal{O}(\frac{1}{\sqrt{m}})$, and the rate is $\mathcal{O}(\frac{1}{m})$ if φ has Lipschitz continuous gradient (See a self-contained proof in Part D of the Appendix). More importantly, Bernstein polynomials are convexity-preserving, i.e. if φ is convex then $B_m(\varphi; \cdot)$ is convex which is critical for deriving the global convergence of our proposed algorithm later.

To approximate the loss ℓ defined on the general interval, we assume $D = \sup_{x\in\mathcal{X}} \|x\| < \infty$ and thus $s := \mathbf{w}^\top x - \mathbf{w}^\top x'$ satisfies $|s| \le 2RD := L$ for any $\mathbf{w}, x$ and x' as $\|\mathbf{w}\| \le R$. Now by changing variables $u = \frac{L+s}{2L}$ (i.e. $s = L(2u-1)$), the loss ℓ induces a function on the unit interval $[0,1]$ by letting $\varphi(u) = \ell(s)$ for any $s \in [-L, L]$. Consequently, there holds

$$\begin{aligned}
&\ell\big(\mathbf{w}^\top x - \mathbf{w}^\top x'\big) \approx B_m\big(\varphi; \frac{L + \mathbf{w}^\top x - \mathbf{w}^\top x'}{2L}\big)\\
=&\sum_{k=0}^{m} \binom{m}{k} \Delta^k\varphi(0) \big(\frac{L + \mathbf{w}^\top x - \mathbf{w}^\top x'}{2L}\big)^k\\
=&\frac{1}{m+1}\sum_{k=0}^{m}\sum_{i=0}^{k}(m+1)\binom{m}{k}\binom{k}{i}\frac{\Delta^k\varphi(0)}{(2L)^k}[(L/2+\mathbf{w}^\top x)^i][(L/2-\mathbf{w}^\top x')^{k-i}]\\
=&\frac{1}{m+1}\sum_{i=0}^{m}\Big\{\big[(L/2+\mathbf{w}^\top x)^i\big]\times\big[\sum_{k=i}^{m}\binom{m}{k}\binom{k}{i}\frac{(m+1)\Delta^k\varphi(0)}{(2L)^k}(L/2-\mathbf{w}^\top x')^{k-i}\big]\Big\}\\
=&\frac{1}{m+1}\sum_{i=0}^{m}\Big[f_i(\mathbf{w};x)\tilde{f}_i(\mathbf{w};x')\Big], \qquad (2.4)
\end{aligned}$$

where $f_i(\mathbf{w};x) = (\frac{L}{2} + \mathbf{w}^\top x)^i$, and $\tilde{f}_i(\mathbf{w};x) = \sum_{k=i}^{m}\binom{m}{k}\binom{k}{i}\frac{(m+1)\Delta^k\varphi(0)}{(2L)^k}(\frac{L}{2} - \mathbf{w}^\top x)^{k-i}$.

As argued in (2.4), AUC maximization with a general loss now becomes

$$\min_{\|\mathbf{w}\|\le R}\Big\{f(\mathbf{w}) := \frac{1}{(m+1)}\sum_{i=0}^{m}\mathbb{E}\big[f_i(\mathbf{w};x)\mathbb{I}_{[y=1]}\tilde{f}_i(\mathbf{w};x')\mathbb{I}_{[y'=-1]}\big]\Big\}, \qquad (2.5)$$

which is convex due to the convexity-preserving property of Bernstein polynomials [25, 26].

2.1. **Min-max formulation.** Here, we show that the AUC maximization (2.5) is equivalent to a (non-convex) min-max formulation, and discuss some properties of its objective function. For the simplicity of notation, denote $\mathbf{e}^+ = \{e_i^+(\mathbf{w},z)\}_{i=0}^m$ where $e_i^+(\mathbf{w},z) = f_i(\mathbf{w};x)\mathbb{I}_{[y=1]}$ and denote $\mathbf{e}^- = \{e_i^-(\mathbf{w},z)\}_{i=0}^m$ where $e_i^-(\mathbf{w},z) = \tilde{f}_i(\mathbf{w};x)\mathbb{I}_{[y=-1]}$. Likewise, we define their expectations by $\mathbf{E}^+(\mathbf{w}) = \{E_i^+(\mathbf{w}) := \mathbb{E}_z[e_i^+(\mathbf{w},z)]\}_{i=0}^m$ and $\mathbf{E}^-(\mathbf{w}) = \{E_i^-(\mathbf{w}) := \mathbb{E}_z[e_i^-(\mathbf{w},z)]\}_{i=0}^m$. In addition, we define, for any $\mathbf{v} = (\mathbf{w},\mathbf{a},\mathbf{b}) \in \mathbb{R}^d\times\mathbb{R}^{m+1}\times\mathbb{R}^{m+1}$, $\boldsymbol{\alpha}\in\mathbb{R}^{m+1}$ and $z = (x,y)\in\mathcal{Z}$, that

$$F(\mathbf{v},\boldsymbol{\alpha};z) = \frac{1}{2(m+1)}\big\{-\|\boldsymbol{\alpha}\|^2 + 2\boldsymbol{\alpha}^\top(\mathbf{e}^+ + \mathbf{e}^-) + \|\mathbf{a}\|^2 - 2\mathbf{a}^\top\mathbf{e}^+ + \|\mathbf{b}\|^2 - 2\mathbf{b}^\top\mathbf{e}^-\big\}. \qquad (2.6)$$

Theorem 2.1. *AUC optimization (2.5) is equivalent to*

$$\min_{\substack{\mathbf{v}=(\mathbf{w},\mathbf{a},\mathbf{b})\\ \|\mathbf{w}\|\le R}}\ \max_{\boldsymbol{\alpha}}\Big\{\phi(\mathbf{v},\boldsymbol{\alpha}) := \mathbb{E}_z[F(\mathbf{v},\boldsymbol{\alpha};z)]\Big\} \qquad (2.7)$$

where the expectation $\mathbb{E}_z[\cdot]$ is taken w.r.t. $z = (x,y)$.

Proof. Notice, since (x,y) and (x',y') are independent, that the expectation terms in (2.5) can be written by

$$\begin{aligned}
\mathbb{E}\big[f_i(\mathbf{w};x)\tilde{f}_i(\mathbf{w};x')\mathbb{I}_{[y=1]}\mathbb{I}_{[y'=-1]}\big] &= \mathbb{E}_z\big[f_i(\mathbf{w};x)\mathbb{I}_{[y=1]}\big]\cdot\mathbb{E}_z\big[\tilde{f}_i(\mathbf{w};x)\mathbb{I}_{[y=-1]}\big]\\
&= E_i^+(\mathbf{w})E_i^-(\mathbf{w}).
\end{aligned}$$

Thus, the objective function in (2.5) can be rewritten as

$$2(m+1)f(\mathbf{w}) = 2(\mathbf{E}^+)^\top \mathbf{E}^- = \|\mathbf{E}^+ + \mathbf{E}^-\|^2 - \|\mathbf{E}^+\|^2 - \|\mathbf{E}^-\|^2.$$

Notice that

$$\begin{aligned}
&\|\mathbf{E}^+ + \mathbf{E}^-\|^2 - \|\mathbf{E}^+\|^2 - \|\mathbf{E}^-\|^2 \\
=&\max_{\boldsymbol{\alpha}} \{-\|\boldsymbol{\alpha}\|^2 + 2\boldsymbol{\alpha}^\top (\mathbf{E}^+ + \mathbf{E}^-)\} + \min_{\mathbf{a}} \{\|\mathbf{a}\|^2 - 2\mathbf{a}^\top \mathbf{E}^+\} + \min_{\mathbf{b}} \{\|\mathbf{b}\|^2 - 2\mathbf{b}^\top \mathbf{E}^-\} \\
=&\max_{\boldsymbol{\alpha}} \mathbb{E}_z \{-\|\boldsymbol{\alpha}\|^2 + 2\boldsymbol{\alpha}^\top (\mathbf{e}^+ + \mathbf{e}^-)\} + \min_{\mathbf{a}} \mathbb{E}_z \{\|\mathbf{a}\|^2 - 2\mathbf{a}^\top \mathbf{e}^+\} \\
&\qquad + \min_{\mathbf{b}} \mathbb{E}_z \{\|\mathbf{b}\|^2 - 2\mathbf{b}^\top \mathbf{e}^-\} \\
=&2(m+1) \min_{\mathbf{a},\mathbf{b}} \max_{\boldsymbol{\alpha}} \mathbb{E}_z \{F(\mathbf{w}, \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha}; z)\}.
\end{aligned}$$

This means that, for every $\mathbf{w}$,

$$f(\mathbf{w}) = \min_{\mathbf{a},\mathbf{b}} \max_{\boldsymbol{\alpha}} \phi(\mathbf{w}, \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha}), \tag{2.8}$$

and the optima are achieved at

$$\mathbf{a}(\mathbf{w}) = \mathbf{E}^+(\mathbf{w}), \ \mathbf{b}(\mathbf{w}) = \mathbf{E}^-(\mathbf{w}), \ \boldsymbol{\alpha}(\mathbf{w}) = \mathbf{E}^+(\mathbf{w}) + \mathbf{E}^-(\mathbf{w}) \tag{2.9}$$

This completes the proof of the theorem. □

Properties of the min-max formulation. We discuss useful properties of the min-max formulation and the function F. Firstly, we can show that $\mathbf{u} = (\mathbf{v}, \boldsymbol{\alpha}) = (\mathbf{w}, \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha})$ in formulation (2.7) can all be restricted to a bounded domain. To see this, notice from (2.8), (2.9) that any optimal point $(\mathbf{v}^*, \boldsymbol{\alpha}^*)$ satisfies $\mathbf{a}^* = \mathbf{a}(\mathbf{w}^*) = \mathbf{E}^+(\mathbf{w}^*), \mathbf{b}^* = \mathbf{b}(\mathbf{w}^*) = \mathbf{E}^-(\mathbf{w}^*), \boldsymbol{\alpha}^* = \boldsymbol{\alpha}(\mathbf{w}^*) = \mathbf{E}^+(\mathbf{w}^*) + \mathbf{E}^-(\mathbf{w}^*)$. Therefore, by the definitions of $\mathbf{E}^+$ and $\mathbf{E}^-$ and noting that $|\mathbf{w}^\top x| \le \|\mathbf{w}\| \|x\| \le RD = \frac{L}{2}$, we have

$$\begin{aligned}
&\|\mathbf{a}^*\| \le \sum_{i=0}^{m} L^i := R_1, \quad \|\boldsymbol{\alpha}^*\| = \|\mathbf{E}^+(\mathbf{w}^*) + \mathbf{E}^-(\mathbf{w}^*)\| \le R_1 + R_2, \\
&\|\mathbf{b}^*\| \le \sum_{i=0}^{m} \sum_{k=i}^{m} \binom{m}{k} \binom{k}{i} \frac{(m+1)|\Delta^k \varphi(0)|}{2^k L^i} := R_2.
\end{aligned}$$

Therefore, without loss of generality, the variables $(\mathbf{w}, \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha})$ in formulation (2.7) can be restricted to the bounded set $\mathbf{v} \in \Omega_1 = \{(\mathbf{w}, \mathbf{a}, \mathbf{b}) : \|\mathbf{w}\| \le R, \|\mathbf{a}\| \le R_1, \|\mathbf{b}\| \le R_2\}$ and $\boldsymbol{\alpha} \in \Omega_2 = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\| \le R_1 + R_2\}$.

Secondly, it is easy to see that the involved function $F(\mathbf{v}, \boldsymbol{\alpha}; z)$ is not convex with respect to $\mathbf{v} = (\mathbf{w}, \mathbf{a}, \mathbf{b})$ and is strongly concave with respect to $\boldsymbol{\alpha}$. Hence the min-max formulation (2.7) is not a standard convex-concave saddle point problem [23]. However, one can show that F is ρ-weakly convex on $\mathbf{v}$ for some $\rho > 0$, i.e. $F(\mathbf{v}, \boldsymbol{\alpha}; z) + \frac{\rho}{2}\|\mathbf{v}\|^2$ is convex on $\mathbf{v}$ for any $\boldsymbol{\alpha}$ and z [24, 8] (See its proof in Part A of the Appendix). Perhaps most importantly, one can further show that adding a partial regularization term $\|\mathbf{w}\|^2$ to $F(\mathbf{v}, \boldsymbol{\alpha}; z)$, instead of the square norm of all primal variables $\|\mathbf{v}\|^2$, will play the same convexity-inducing effect, i.e. $F(\mathbf{v}, \boldsymbol{\alpha}; z) + \frac{\gamma}{2}\|\mathbf{w}\|^2$ is convex w.r.t. $\mathbf{v}$ for a sufficient large $\gamma > 0$. To show this, let us introduce some notations

$$S_1^+ = \sum_{i=0}^{m} i L^{i-1}, \quad S_2^+ = \sum_{i=0}^{m} i(i-1) L^{i-2},$$

$$S_1^- = \sum_{i=0}^{m}\sum_{k=i}^{m}\binom{m}{k}\binom{k}{i}\frac{(m+1)(k-i)}{2^k L^{i+1}}|\Delta^k\varphi(0)|,$$
$$S_2^- = \sum_{i=0}^{m}\sum_{k=i}^{m}\binom{m}{k}\binom{k}{i}\frac{(m+1)(k-i)(k-i-1)}{2^k L^{i+2}}|\Delta^k\varphi(0)|.$$

Let

$$\gamma_0 := \frac{1}{m+1}\max\{(2R_1+R_2)D^2S_2^+ + D^2(S_1^+)^2, (R_1+2R_2)D^2S_2^- + D^2(S_1^-)^2\} \quad (2.10)$$

and consider

$$\Phi_\gamma^t(\mathbf{v},\boldsymbol{\alpha};z) := F(\mathbf{v},\boldsymbol{\alpha};z) + \frac{\gamma}{2}\|\mathbf{w}-\bar{\mathbf{w}}_{t-1}\|^2. \quad (2.11)$$

Proposition 1. *Assume $\gamma \geq \gamma_0$ where γ_0 is given by (2.10). Then, for any fixed $\bar{\mathbf{w}}_{t-1}$ and z, we have that $\Phi_\gamma^t(\mathbf{v},\boldsymbol{\alpha};z)$ is convex w.r.t. $\mathbf{v}$ and concave w.r.t. $\boldsymbol{\alpha}$.*

As we will see in the next section, this proposition plays a key role in designing an SGD-type algorithm which enjoys the global convergence (see the proof in Part B of the Appendix).

3. Algorithm and convergence analysis.

In this section, we propose a stochastic optimization algorithm for our novel min-max formulation (2.7). Our proposed algorithm called SAUC is described in Algorithm 1 which is inspired by the recent work [27]. In particular, the appealing work [27] studied a family of non-convex min-max problems where the minimization component is weakly convex and the maximization component is concave, which is motivated by the studies on weakly convex minimization problems [8, 7].

Algorithm 1 Stochastic AUC Optimization (SAUC)

1: **Input:** $R > 0$, $\gamma \geq \gamma_0$ and $\beta > 0$.
2: Initialize $\bar{\mathbf{v}}_0 = 0$ and $\bar{\boldsymbol{\alpha}}_0 = 0$.
3: **for** $t = 1$ **to** $T-1$ **do**
4: Set $\mathbf{v}_0^t = \bar{\mathbf{v}}_{t-1}, \boldsymbol{\alpha}_0^t = \bar{\boldsymbol{\alpha}}_{t-1}$ and $\eta_t = \frac{\beta}{\sqrt{t}}$.
5: **for** $j = 1$ **to** t **do**
6: Randomly sample $z_j^t = (x_j^t, y_j^t)$ and compute
$$\mathbf{v}_j^t = \mathbf{Proj}_{\Omega_1}\big(\mathbf{v}_{j-1}^t - \eta_t\nabla_{\mathbf{v}}\Phi_\gamma^t(\mathbf{v}_{j-1}^t, \boldsymbol{\alpha}_{j-1}^t; z_j^t)\big),$$
$$\boldsymbol{\alpha}_j^t = \mathbf{Proj}_{\Omega_2}\big(\boldsymbol{\alpha}_{j-1}^t + \eta_t\nabla_{\boldsymbol{\alpha}}\Phi_\gamma^t(\mathbf{v}_{j-1}^t, \boldsymbol{\alpha}_{j-1}^t; z_j^t)\big)$$
7: **end for**
8: Compute $\bar{\mathbf{v}}_t = \frac{1}{t}\sum_{j=0}^{t-1}\mathbf{v}_j^t$ and $\bar{\boldsymbol{\alpha}}_t = \frac{1}{t}\sum_{j=0}^{t-1}\boldsymbol{\alpha}_j^t$.
9: **end for**
10: **Output:** $\widetilde{\mathbf{v}}_T := \frac{1}{T}\sum_{t=0}^{T-1}\bar{\mathbf{v}}_t$ and $\widetilde{\boldsymbol{\alpha}}_T := \frac{1}{T}\sum_{t=0}^{T-1}\bar{\boldsymbol{\alpha}}_t$.

Our algorithm SAUC has a cheap storage and per-iteration cost. Indeed, at line 6, the samples $\{z_j^t = (x_j^t, y_j^t)\}$ are i.i.d from the distribution $\mathcal{P}$ on $\mathcal{X}\times\mathcal{Y}$. As such, one only need to store the current sample with space $\mathcal{O}(d)$, which is linear w.r.t. the data dimension. It makes SAUC the first truly online algorithm for AUC maximization with general loss. The main per-iteration cost comes from lines 6

and 7, which is the standard SGD-type algorithm by [23] for solving the standard (convex-concave) stochastic saddle point problem:

$$\min_{\mathbf{v}\in\Omega_1}\max_{\boldsymbol{\alpha}\in\Omega_2}\{\varphi_\gamma^t(\mathbf{v},\boldsymbol{\alpha}) := \mathbb{E}_z(\Phi_\gamma^t(\mathbf{v},\boldsymbol{\alpha};z))\}. \tag{3.1}$$

Note that the projections at line 6 can be easily computed since Ω_1 and Ω_2 are bounded ℓ_2-balls. The computation of $\nabla_{\mathbf{v}}\Phi_\gamma^t$ (involving the computation of $\nabla f_i(\mathbf{w};x)$ and $\nabla\tilde{f}_i(\mathbf{w};x)$ etc.) only needs to compute and save the inner product w^Tx, which costs $\mathcal{O}(d)$ then do arithmetic operations $\mathcal{O}(m)$ times. Thus, the per-iteration cost is $\mathcal{O}(m+d)$, where the m can be ignored for large d.

Compared with the work [27], the main difference of our algorithm from that in [27] is that we propose to use the proximal term $\frac{\gamma}{2}\|\mathbf{w}-\bar{\mathbf{w}}_{t-1}\|^2$ in (3.1) instead of $\frac{\rho}{2}\|\mathbf{v}-\bar{\mathbf{v}}_{t-1}\|^2$. This simple design is important to prove the convergence of the global convergence of Algorithm 1. To be more specific, we can interpret the result in [27] under our setting of AUC maximization as follows. Let $\psi(\mathbf{v}) := \max_{\boldsymbol{\alpha}}\phi(\mathbf{v},\boldsymbol{\alpha})$, where $\phi(\mathbf{v},\boldsymbol{\alpha})$ is defined in (2.5), and $\psi_r(\mathbf{v}) := \psi(\mathbf{v})+r(\mathbf{v})$, where $r(\mathbf{v})$ is the indicator function of Ω_1 (i.e. if $\mathbf{v}\in\Omega_1$, then $r(\mathbf{v})=0$; otherwise, $r(\mathbf{v})=\infty$). The work [27] transformed a constrained optimization problem, namely $\min_{\mathbf{v}\in\Omega_1}\psi(\mathbf{v})$ to an unconstrained, however non-smooth optimization problem, i.e. $\min_{\mathbf{v}}\psi_r(\mathbf{v})$. They used proximal term $\frac{\rho}{2}\|\mathbf{v}-\bar{\mathbf{v}}_{t-1}\|^2$ in (3.1) and accordingly defined $\hat{\mathbf{v}}_t := \operatorname{argmin}_{\mathbf{v}\in\Omega_1}\psi(\mathbf{v})+\frac{\rho}{2}\|\mathbf{v}-\bar{\mathbf{v}}_t\|^2$. Then for τ chosen uniformly from $\{0,\cdots,T-1\}$, the following result, with appropriately choosing stepsizes, was established $\mathbb{E}\left[\text{dist}^2(0,\partial\psi_r(\hat{\mathbf{v}}_\tau))\right] \le \rho^2\mathbb{E}\|\hat{\mathbf{v}}_\tau-\bar{\mathbf{v}}_\tau\|^2 \le O(\frac{\log(T)}{T})$, namely, $\bar{\mathbf{v}}_\tau$ is close to an ε-stationary point (i.e. $\hat{\mathbf{v}}_\tau$) of $\psi_r(\mathbf{v})$ with $\varepsilon = O(\sqrt{\log(T)/T})$. Note that this only means the local convergence of $\bar{\mathbf{v}}_t$ and it is difficult to show the global convergence since $\psi(\mathbf{v}) = (\|\mathbf{E}^+ + \mathbf{E}^-\|^2 + \|\mathbf{a}\|_2^2 - 2\mathbf{a}^\top\mathbf{E}^+ + \|\mathbf{b}\|_2^2 - 2\mathbf{b}^\top\mathbf{E}^-)/(2(m+1))$, as a function of $\mathbf{v}=(\mathbf{w},\mathbf{a},\mathbf{b})$, is not convex.

The global convergence result for Algorithm 1 is stated as follows. Let $\mathbf{w}^* := \operatorname{argmin}_{\|\mathbf{w}\|\le R} f(\mathbf{w})$ to be an optimal solution of the original AUC maximization problem (2.5).

Theorem 3.1. *Assume the data $\{z_j^t=(x_j^t,y_j^t)\in\mathcal{X}\times\mathcal{Y} : t\in[1,T-1],1\le j\le t\}$ are i.i.d and consider the sequence $\{\bar{\mathbf{w}}_t\}_{t=1}^{T-1}$ generated by Algorithm 1 with stepsizes $\{\eta_t=\beta/\sqrt{t}\}$. For the output defined by $\widetilde{\mathbf{w}}_T := \frac{1}{T}\sum_{t=0}^{T-1}\bar{\mathbf{w}}_t$, we get $\mathbb{E}[f(\widetilde{\mathbf{w}}_T)-f(\mathbf{w}^*)]^2 \le \frac{C_m}{\sqrt{T}}$, where C_m is an absolute constant depending on m but independent of T.*

The proof of Theorem 3.1 requires the following lemma. It shows the convergence of the SGD-type algorithm for solving the saddle point problem (3.1) (the inner loop from line 5 to line 8 in Algorithm 1 with fixed t in the outer loop). Recall that φ_γ^t is defined by (3.1) and define the duality gap at $\bar{\mathbf{u}}_t := (\bar{\mathbf{v}}_t,\bar{\boldsymbol{\alpha}}_t)$ by

$$\epsilon_t(\bar{\mathbf{u}}_t) := \max_{\boldsymbol{\alpha}\in\Omega_2}\varphi_\gamma^t(\bar{\mathbf{v}}_t,\boldsymbol{\alpha}) - \min_{\mathbf{v}\in\Omega_1}\varphi_\gamma^t(\mathbf{v},\bar{\boldsymbol{\alpha}}_t). \tag{3.2}$$

Lemma 3.2. *Assume that the data $\{z_j^k=(x_j^k,y_j^k) : k\in[1,t],j\in[1,k]\}$ is i.i.d. and consider the sequence $\{\mathbf{u}_j^t\}_{j=1}^t$ generated by the inner loop from line 5 to line 8 in Algorithm 1 with the stepsize $\eta_t=\beta/\sqrt{t}$. Then we have $\mathbb{E}[\epsilon_t(\bar{\mathbf{u}}_t)] \le C_1/\sqrt{t}$, where C_1 is an absolute constant independent of t (see its explicit expression in the proof).*

The proof of Lemma 3.2 is standard [23]. A self-contained proof can be found in Part C of the Appendix. To prove Theorem 3.1, we further define, for any $\bar{\mathbf{w}}_t$, that $\hat{\mathbf{w}}_t = \operatorname{argmin}_{\|\mathbf{w}\|\le R}\{f(\mathbf{w})+\frac{\gamma}{2}\|\mathbf{w}-\bar{\mathbf{w}}_t\|^2\}$. Now we are ready to prove Theorem 3.1.

Proof of Theorem 3.1. We first estimate $f(\bar{\mathbf{w}}_t) - f(\mathbf{w}^*)$. For this purpose, notice the convexity of $f(\mathbf{w})$ indicates

$$f(\hat{\mathbf{w}}_t) - f(\mathbf{w}^*) \leq (\hat{\mathbf{w}}_t - \mathbf{w}^*)^\top \nabla f(\hat{\mathbf{w}}_t),$$

and the optimality condition of $\hat{\mathbf{w}}_t$ implies

$$(\mathbf{w}^* - \hat{\mathbf{w}}_t)^\top (\nabla f(\hat{\mathbf{w}}_t) + \gamma(\hat{\mathbf{w}}_t - \bar{\mathbf{w}}_t)) \geq 0.$$

Combining these two inequalities implies that

$$f(\hat{\mathbf{w}}_t) - f(\mathbf{w}^*) \leq \gamma(\hat{\mathbf{w}}_t - \bar{\mathbf{w}}_t)^\top (\mathbf{w}^* - \hat{\mathbf{w}}_t).$$

Consequently,

$$f(\hat{\mathbf{w}}_t) - f(\mathbf{w}^*) \leq 2R\gamma \|\hat{\mathbf{w}}_t - \bar{\mathbf{w}}_t\|. \tag{3.3}$$

Moreover, the Lipschitz continuity of $f(\mathbf{w})$ for $\|\mathbf{w}\| \leq R$, we have $|f(\bar{\mathbf{w}}_t) - f(\hat{\mathbf{w}}_t)| \leq L_1 \|\bar{\mathbf{w}}_t - \hat{\mathbf{w}}_t\|$ where $L_1 := D(S_1^+ R_2 + S_1^- R_1)/(m+1)$ estimated as the upper bound of $\|\nabla f(\mathbf{w})\|$. Combining this with (3.3) implies, for any t, that

$$f(\bar{\mathbf{w}}_t) - f(\mathbf{w}^*) \leq (L_1 + 2R\gamma)\|\bar{\mathbf{w}}_t - \hat{\mathbf{w}}_t\|.$$

By convexity and Cauchy-Schwarz inequality, the above estimation indicates

$$[f(\widetilde{\mathbf{w}}_T) - f(\mathbf{w}^*)]^2 \leq \left\{ \frac{1}{T} \sum_{t=0}^{T-1} [f(\bar{\mathbf{w}}_t) - f(\mathbf{w}^*)] \right\}^2 \leq \frac{(L_1 + 2R\gamma)^2}{T} \sum_{t=0}^{T-1} \|\bar{\mathbf{w}}_t - \hat{\mathbf{w}}_t\|^2. \tag{3.4}$$

Now observe that

$$\frac{\gamma}{2}\|\bar{\mathbf{w}}_t - \hat{\mathbf{w}}_{t-1}\|^2 \leq f(\bar{\mathbf{w}}_t) + \frac{\gamma}{2}\|\bar{\mathbf{w}}_t - \bar{\mathbf{w}}_{t-1}\|^2 - f(\hat{\mathbf{w}}_{t-1}) - \frac{\gamma}{2}\|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_{t-1}\|^2 \leq \epsilon_t(\bar{\mathbf{u}}_t), \tag{3.5}$$

wherein the first inequality follows the optimality condition of $\hat{\mathbf{w}}_{t-1}$ and the γ-strong convexity of $f(\mathbf{w}) + \frac{\gamma}{2}\|\mathbf{w} - \bar{\mathbf{w}}_{t-1}\|^2$ for $\|\mathbf{w}\| \leq R$, while the second inequality can be derived as follows. Indeed, from (2.8) in the proof of Theorem 2.1 we know that $f(\bar{\mathbf{w}}_t) = \min_{\mathbf{a},\mathbf{b}} \max_{\boldsymbol{\alpha}} \phi(\bar{\mathbf{w}}_t, \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha})$ which implies that

$$\begin{aligned} f(\bar{\mathbf{w}}_t) + \frac{\gamma}{2}\|\bar{\mathbf{w}}_t - \bar{\mathbf{w}}_{t-1}\|^2 &= \min_{\mathbf{a},\mathbf{b}} \max_{\boldsymbol{\alpha}} \phi(\bar{\mathbf{w}}_t, \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha}) + \frac{\gamma}{2}\|\bar{\mathbf{w}}_t - \bar{\mathbf{w}}_{t-1}\|^2 \\ &\leq \max_{\boldsymbol{\alpha}} \phi(\bar{\mathbf{v}}_t, \boldsymbol{\alpha}) + \frac{\gamma}{2}\|\bar{\mathbf{w}}_t - \bar{\mathbf{w}}_{t-1}\|^2 = \max_{\boldsymbol{\alpha} \in \Omega_2} \varphi_\gamma^t(\bar{\mathbf{v}}_t, \boldsymbol{\alpha}). \end{aligned}$$

In addition, by (2.8) again

$$\begin{aligned} & f(\hat{\mathbf{w}}_{t-1}) + \frac{\gamma}{2}\|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_{t-1}\|^2 \\ &= \min_{\mathbf{w} \in \Omega} \Big\{ f(\mathbf{w}) + \frac{\gamma}{2}\|\mathbf{w} - \bar{\mathbf{w}}_{t-1}\|^2 \Big\} = \min_{\mathbf{w} \in \Omega} \Big\{ \min_{\mathbf{a},\mathbf{b}} \max_{\boldsymbol{\alpha}} \phi(\mathbf{v}, \boldsymbol{\alpha}) + \frac{\gamma}{2}\|\mathbf{w} - \bar{\mathbf{w}}_{t-1}\|^2 \Big\} \\ &\geq \min_{\mathbf{w} \in \Omega} \Big\{ \min_{\mathbf{a},\mathbf{b}} \phi(\mathbf{v}, \bar{\boldsymbol{\alpha}}_t) + \frac{\gamma}{2}\|\mathbf{w} - \bar{\mathbf{w}}_{t-1}\|^2 \Big\} = \min_{\mathbf{v} \in \Omega_1} \Big\{ \phi(\mathbf{v}, \bar{\boldsymbol{\alpha}}_t) + \frac{\gamma}{2}\|\mathbf{w} - \bar{\mathbf{w}}_{t-1}\|^2 \Big\} \\ &= \min_{\mathbf{v} \in \Omega_1} \varphi_\gamma^t(\mathbf{v}, \bar{\boldsymbol{\alpha}}_t). \end{aligned}$$

To further bound (3.5), we need the following elementary inequalities:

$$\begin{aligned} & \|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_{t-1}\|^2 - \|\bar{\mathbf{w}}_t - \bar{\mathbf{w}}_{t-1}\|^2 \\ =& (\|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_{t-1}\| - \|\bar{\mathbf{w}}_t - \bar{\mathbf{w}}_{t-1}\|)(\|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_{t-1}\| + \|\bar{\mathbf{w}}_t - \bar{\mathbf{w}}_{t-1}\|) \\ \leq& \|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_t\| \cdot (2\|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_{t-1}\| + \|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_t\|) \\ =& 2\|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_{t-1}\| \cdot \|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_t\| + \|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_t\|^2 \\ \leq& \frac{1}{3}\|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_{t-1}\|^2 + 4\|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_t\|^2. \end{aligned}$$

Combining this with (3.5) implies that

$$\begin{aligned} f(\bar{\mathbf{w}}_t) &\le f(\hat{\mathbf{w}}_{t-1}) + \varepsilon_t(\bar{\mathbf{u}}_t) + \frac{\gamma}{2}(\|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_{t-1}\|^2 - \|\bar{\mathbf{w}}_t - \bar{\mathbf{w}}_{t-1}\|^2) \\ &\le f(\hat{\mathbf{w}}_{t-1}) + \varepsilon_t(\bar{\mathbf{u}}_t) + \frac{\gamma}{6}\|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_{t-1}\|^2 + 2\gamma\|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_t\|^2 \\ &\le f(\hat{\mathbf{w}}_{t-1}) + \frac{\gamma}{6}\|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_{t-1}\|^2 + 5\epsilon_t(\bar{\mathbf{u}}_t), \end{aligned}$$

where the last inequality used the fact $\frac{\gamma}{2}\|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_t\|^2 \le \varepsilon(\bar{\mathbf{u}}_t)$ from (3.5). We also notice, by the definition of $\hat{\mathbf{w}}_{t-1}$, that

$$f(\hat{\mathbf{w}}_{t-1}) + \frac{\gamma}{2}\|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_{t-1}\|^2 \le f(\bar{\mathbf{w}}_{t-1}).$$

Combining the above two estimations, we get

$$f(\bar{\mathbf{w}}_t) \le f(\bar{\mathbf{w}}_{t-1}) - \frac{\gamma}{3}\|\hat{\mathbf{w}}_{t-1} - \bar{\mathbf{w}}_{t-1}\|^2 + 5\epsilon_t(\bar{\mathbf{u}}_t).$$

Adding the inequalities from $t = 1$ to $t = T$, we have

$$\begin{aligned} \sum_{t=0}^{T-1} \|\bar{\mathbf{w}}_t - \hat{\mathbf{w}}_t\|^2 &\le 3\gamma^{-1}(f(\bar{\mathbf{w}}_0) - f(\bar{\mathbf{w}}_T)) + 15\gamma^{-1}\sum_{t=1}^{T}\epsilon_t(\bar{\mathbf{u}}_t) \\ &\le 3\gamma^{-1}(2L_1R) + 15\gamma^{-1}\sum_{t=1}^{T}\epsilon_t(\bar{\mathbf{u}}_t). \end{aligned}$$

Combining this with (3.4) and taking expectation on both sides imply that

$$\mathbb{E}[f(\widetilde{\mathbf{w}}_T) - f(\mathbf{w}^*)]^2 \le \frac{3(L_1 + 2R\gamma)^2}{\gamma T}\big\{2L_1R + 5\sum_{t=1}^{T}\mathbb{E}[\epsilon_t(\bar{\mathbf{u}}_t)]\big\}.$$

Combining it with Lemma 3.2 completes the proof with $C_m = 30(L_1+2R\gamma)^2(L_1R+C_1)\gamma^{-1}$. □

Trading off the approximation error of Bernstein polynomials and Theorem 3.1 yields the following final convergence rate for the original objective function g of AUC maximization defined by (2.2).

Theorem 3.3. *Consider a surrogate loss function ℓ on $[-L, L]$, with Lipschitz constant G, Bernstein approximation with degree m, and the SAUC output $\widetilde{\mathbf{w}}_T$ after T iterations. Then we have*

$$\mathbb{E}[g(\widetilde{\mathbf{w}}_T) - \inf_{\|\mathbf{w}\|\le R} g(\mathbf{w})] \le \frac{2GRD}{\sqrt{m}} + \frac{\widetilde{B}\widetilde{C}^m}{T^{1/4}},$$

where $\widetilde{C}$ and $\widetilde{B}$ are constants depending on G, R, D and β but independent of m and T.

The proof of Theorem 3.3 is given in Part D of the Supplementray Material. From the above theorem, we can choose $m = \log_{\widetilde{C}}(T^{1/4}/\log T) = \mathcal{O}(\log T)$ which yields the final convergence rate of $\mathcal{O}(\sqrt{1/\log T})$. However, this analysis and convergence rate is not satisfactory. It remains a challenging problem to us on how to derive a desirable final convergence rate.

4. **Experiments.** In this section, we compare our proposed algorithm SAUC against existing AUC optimization algorithms. In particular, SAUC-H and SAUC-L denotes SAUC with the hinge loss and the logistic loss, respectively. All experiments were implemented in Python 3 with 16 $\times$ 3.0GHz CPUs and 128GB memory.

We conducted our experiments on 9 benchmark datasets which are downloaded from the LIBSVM [4] and UCI machine learning repository [9]. Multi-class datasets have been converted into binary-class by randomly partitioning classes into positive and negative groups. All data have been normalized with unit ℓ_2-norm. Information about these datasets is summarized in Table 1.

TABLE 1. Statistics of datasets

Name	# Instances	# Features	Name	# Instances	# Features	Name	# Instances	# Features
australian	690	14	leu	38	7,129	sector	6,412	55,197
cod-rna	59,535	8	madelon	2,000	500	skin nonskin	245,057	3
dna	2,000	180	news20	15,935	62,061	svmguide1	3,089	4

Generalization performance: We compare SAUC with following state-of-the-art online learning algorithms for AUC optimization: 1) Online AUC Maximization(OAM) [36] with focus on OAM_{gra} assocaited with the hinge loss. The buffer sizes N_+, N_- for positive and negative classes are set as 100 as in the original paper; 2) One-Pass AUC Maximization(OPAUC) [12] optimizes square loss with ℓ_2 regularizing term. 3) Stochastic Online AUC Maximization(SOLAM) [32] optimizes square loss on bounded domain; 4) Stochastic Proximal Algorithm for AUC Maximization(SPAM) [22] optimizes square loss with ℓ_2 regularizer; 5) Fast Stochastic AUC Maximization(FSAUC) [21] optimizes square loss on bounded ℓ_1 domain. The probability parameter δ is set as 0.1 as in the codes author provided. All the algorithms with ℓ_2 regularizer is converted to ℓ_2-norm constraint for the sake of fair comparison.

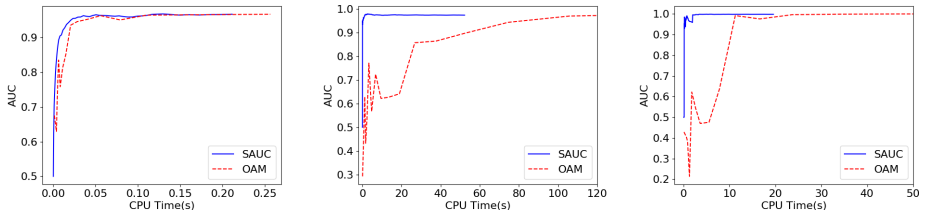
In the training phase of each algorithms, we use 5-fold cross validation to determine the bound radius $R \in 10^{[-2:2]}$ and the learning rate parameter $\beta \in 10^{[-2:2]}$. The proximal parameter γ is chosen as γ_0 as given in (2.10). Throughout our experiments, the degree of the Bernstein polynomials m is chosen to be 10. The performance of each algorithms is evaluated by averaging results from 10 epochs of 5-fold cross validations.

Testing performances of all methods are summarized in Table 2. These results show that SAUC achieves similar or competitive performances as other state-of-the-art online or stochastic methods based on AUC maximization. In particular, SAUC-H performs similarly as OAM_{gra} on all datasets. In addition, on datasets `australian`, `leu`, `sector`, `skin_nonskin` and `svmguide1`, the hinge loss based algorithm SAUC-H outperforms square loss based algorithms. This suggests that the different losses (hinge loss and logistic loss) may be more suitable for these datasets.

Convergence speed: We compare convergence versus CPU time (in seconds) of SAUC-H and OAM_{gra} on datasets `dna`, `news20` and `sector`. The reason for choosing these two algorithms is because they both maximize AUC under the hinge loss. The results are summarized in Figure 1. These results show that SAUC-H converges much faster than OAM_{gra} when the data feature dimension d is high. One important reason may be that SAUC only needs use $\mathcal{O}(d)$ memory and per-iteration cost while OAM needs $\mathcal{O}(Bd)$, where B is the buffer size.

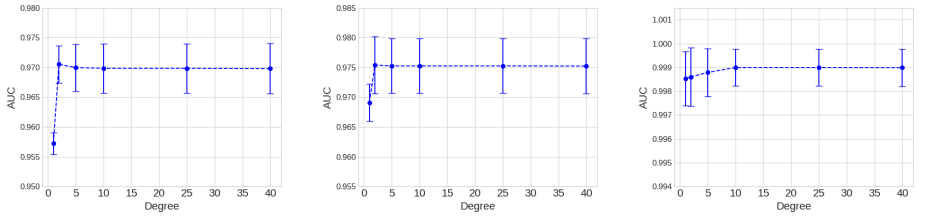
TABLE 2. Comparison of AUC score (mean±std) on test data; OPAUC on `news20` and `sector` does not converge in a reasonable time limit. Best AUC value on each dataset is in **bold** and second is underlined.

Dataset	SAUC-H	SAUC-L	OAM_{gra}	OPAUC	SOLAM	FSAUC	SPAM
australian	**.9250±.0057**	.9249±.0045	.9238±.0031	.9127±.0016	.9202±.0065	.9217±.0064	.9233±.0024
cod-rna	.9190±.0030	.9189±.0031	**.9194±.0030**	.9193±.0030	.9190±.0031	.9190±.0031	.9190±.0031
dna leu	.9702±.0038	.9710±.0040	.9716±.0025	.9675±.0023	.9735±.0015	**.9796±.0031**	.9753±.0046
madelon	**1.000±.0000**	**1.000±.0000**	**1.000±.0000**	.9810±.0269	.9907±.0131	.9778±.0314	.9905±.0135
news20	.6319±.0203	.6315±.0219	.6295±.0249	.6315±.0242	.6317±.0184	.6300±.0228	**.6321±.0151**
sector	.9759±.0013	.9766±.0019	.9750±.0031	-	**.9794±.0019**	.9758±.0030	.9774±.0029
skin nonskin	**.9994±.0004**	**.9994±.0003**	.9991±.0003	-	.9987±.0008	.9975±.0016	.9990±.0006
svmguide1	.9948±.0009	**.9966±.0003**	.9960±.0005	.9405±.0011	.9456±.0012	9407±.0011	.9388±.0013
	.8848±.0062	.8842±.0013	.8845±.0099	.8818±.0033	.8804±.0072	.8804±.0084	.8796±.0101



(a) dna (b) news20 (c) sector

FIGURE 1. Comparison of convergence speed between SAUC-H and OAM_{gra}.



(a) dna (b) news20 (c) sector

FIGURE 2. Evaluation of AUC scores vesus the degree of the Bernstein polynomial.

Sensitivity of the degree of the Bernstein polynomial: Here we investigate the sensitivity of the degree m of Bernstein polynomials to the empirical performance. Figure 2 evaluates the AUC scores of SAUC-H with varied Bernstein polynomial degrees m on datasets `dna`, `news20` and `sector`.

First of all, we find that on `sector`, when the Bernstein polynomial degree is too small (e.g. $m = 2$ or $m = 5$), SAUC achieves lowest AUC scores. This matches the intuition that the Bernstein polynomial approximates the original loss badly. Interestingly, we also find that on `dna` and `news20` when $m = 2$, the approximated loss becomes a square loss and the performance is improved, which coincides with the results in Table 2. Finally, we find that when m is large enough, the AUC scores

tend to become saturated. This is consistent with theoretically optimal choice of m is $\mathcal{O}(\log T)$ from Theorem 3.3 and a larger degree does not necessarily lead to better performance.

5. **Conclusion.** In this paper we proposed to use the Bernstein polynomials to uniformly approximate a general convex loss, and then showed that AUC maximization is equivalent to a (non-convex) weakly convex saddle point (min-max) problem. From this equivalent formulation, we proposed a novel SGD-type AUC optimization algorithm for streaming data analysis. Although the min-max formulation is non-convex, we showed that the proposed algorithm still enjoys the global convergence through sufficiently exploring the intrinsic structure of the min-max formulation and the convex-preserving property of Bernstein polynomials. Finally, we performed experiments to validate effectiveness of the proposed algorithm.

There are several directions for future research. Firstly, the decomposition of $f(\mathbf{w})$ into $\phi(\mathbf{v}, \boldsymbol{\alpha})$ is not unique. It would be interesting to find other possible ones and investigate their theoretical differences and the resulting algorithmic performances in practice. Secondly, the choice of parameter γ_0 in (2.10) for SAUC is potentially large if m or R is large. It would be interesting if we can choose the γ_0 adapatively using some strategies of line search. Another closely related question is the final convergence rate for the original objective function (2.2) of AUC maximization is not desirable as the estimation of the constant C_m in terms of γ_0 and R is complicated. It remians unclear to us how to derive a fast final convergence rate.

Acknowledgement. The authors would like to thank the reviewers for their constructive comments and suggestions.

Appendix.

A. **Weak-convexity of F w.r.t. $\mathbf{v}$.** Remind that a C^1-smooth function with ρ_0-Lipschitz gradient is ρ_0-weakly convex. Thus we prove F is weakly convex by calculating its Lipschitz gradient constant. Denote $\rho_0^+ = \max\{DS_1^+ + (2R_1 + R_2)D^2S_2^+, 1 + DS_1^+\}$, and $\rho_0^- = \max\{DS_1^- + (2R_2 + R_1)D^2S_2^-, 1 + DS_1^-\}$. We study two cases $y = 1$ and $y = -1$ separately. If $y = 1$, we have

$$\nabla_{\mathbf{v}} F(\mathbf{v}, \boldsymbol{\alpha}; z) = \frac{1}{m+1} \begin{pmatrix} (\nabla_{\mathbf{w}} \mathbf{e}^+)(\boldsymbol{\alpha} - \mathbf{a}) \\ \mathbf{a} - \mathbf{e}^+ \\ \mathbf{b} \end{pmatrix}.$$

So we have

$$\begin{aligned} &(m+1)[\nabla_{\mathbf{v}} F(\mathbf{v}, \boldsymbol{\alpha}; z) - \nabla_{\mathbf{v}} F(\mathbf{v}', \boldsymbol{\alpha}; z)] \\ &= \begin{pmatrix} (\nabla_{\mathbf{w}} \mathbf{e}^+(\mathbf{w}))(\boldsymbol{\alpha} - \mathbf{a}) - (\nabla_{\mathbf{w}} \mathbf{e}^+(\mathbf{w}'))(\boldsymbol{\alpha} - \mathbf{a}') \\ \mathbf{a} - \mathbf{e}^+(\mathbf{w}) - \mathbf{a}' + \mathbf{e}^+(\mathbf{w}') \\ \mathbf{b} - \mathbf{b}' \end{pmatrix}. \end{aligned}$$

Combining this with the fact $(\nabla_{\mathbf{w}} \mathbf{e}^+(\mathbf{w}))(\boldsymbol{\alpha} - \mathbf{a}) - (\nabla_{\mathbf{w}} \mathbf{e}^+(\mathbf{w}'))(\boldsymbol{\alpha} - \mathbf{a}') = (\nabla_{\mathbf{w}} \mathbf{e}^+(\mathbf{w}) - \nabla_{\mathbf{w}} \mathbf{e}^+(\mathbf{w}'))(\boldsymbol{\alpha} - \mathbf{a}') + (\nabla_{\mathbf{w}} \mathbf{e}^+(\mathbf{w}))(\mathbf{a}' - \mathbf{a})$, we have

$$\begin{aligned} &\|\nabla_{\mathbf{v}} F(\mathbf{v}, \boldsymbol{\alpha}; z) - \nabla_{\mathbf{v}} F(\mathbf{v}', \boldsymbol{\alpha}; z)\| \\ \leq &\frac{\sqrt{3}}{m+1} \Big[\|\boldsymbol{\alpha} - \mathbf{a}'\| \cdot \|\nabla \mathbf{e}^+(\mathbf{w}) - \nabla \mathbf{e}^+(\mathbf{w}')\| + \|\mathbf{e}^+(\mathbf{w}) - \mathbf{e}^+(\mathbf{w}')\| \\ &+ \|\mathbf{b} - \mathbf{b}'\| + (1 + \|\nabla \mathbf{e}^+(\mathbf{w})\|) \cdot \|\mathbf{a} - \mathbf{a}'\| \Big] \end{aligned}$$

$$\begin{aligned}
&\leq \frac{\sqrt{3}}{m+1}\Big[(2R_1+R_2)D^2S_2^+\cdot\|\mathbf{w}-\mathbf{w}'\|+DS_1^+\cdot\|\mathbf{w}-\mathbf{w}'\| \\
&\quad +\|\mathbf{b}-\mathbf{b}'\|+(1+DS_1^+)\cdot\|\mathbf{a}-\mathbf{a}'\|\Big] \\
&\leq \frac{\sqrt{3}\rho_0^+}{m+1}\|\mathbf{v}-\mathbf{v}'\|.
\end{aligned}$$

If $y=-1$, we can obtain

$$\nabla_{\mathbf{v}}F(\mathbf{v},\boldsymbol{\alpha};z)=\frac{1}{m+1}\begin{pmatrix}(\nabla_{\mathbf{w}}\mathbf{e}^-)(\boldsymbol{\alpha}-\mathbf{b})\\ \mathbf{a}\\ \mathbf{b}-\mathbf{e}^-\end{pmatrix}.$$

So we have

$$\begin{aligned}
&(m+1)[\nabla_{\mathbf{v}}F(\mathbf{v},\boldsymbol{\alpha};z)-\nabla_{\mathbf{v}}F(\mathbf{v}',\boldsymbol{\alpha};z)] \\
&=\begin{pmatrix}(\nabla_{\mathbf{w}}\mathbf{e}^-(\mathbf{w}))(\boldsymbol{\alpha}-\mathbf{b})-(\nabla_{\mathbf{w}}\mathbf{e}^-(\mathbf{w}'))(\boldsymbol{\alpha}-\mathbf{b}')\\ \mathbf{a}-\mathbf{a}'\\ \mathbf{b}-\mathbf{e}^-(\mathbf{w})-\mathbf{b}'+\mathbf{e}^-(\mathbf{w}')\end{pmatrix}.
\end{aligned}$$

Combining this with the fact $(\nabla_{\mathbf{w}}\mathbf{e}^-(\mathbf{w}))(\boldsymbol{\alpha}-\mathbf{b})-(\nabla_{\mathbf{w}}\mathbf{e}^-(\mathbf{w}'))(\boldsymbol{\alpha}-\mathbf{b}')=(\nabla_{\mathbf{w}}\mathbf{e}^-(\mathbf{w})-\nabla_{\mathbf{w}}\mathbf{e}^-(\mathbf{w}'))(\boldsymbol{\alpha}-\mathbf{b}')+(\nabla_{\mathbf{w}}\mathbf{e}^-(\mathbf{w}))(\mathbf{b}'-\mathbf{b})$, we have

$$\begin{aligned}
&\|\nabla_{\mathbf{v}}F(\mathbf{v},\boldsymbol{\alpha};z)-\nabla_{\mathbf{v}}F(\mathbf{v}',\boldsymbol{\alpha};z)\| \\
&\leq \frac{\sqrt{3}}{m+1}\Big[\|\boldsymbol{\alpha}-\mathbf{b}'\|\cdot\|\nabla\mathbf{e}^-(\mathbf{w})-\nabla\mathbf{e}^-(\mathbf{w}')\|+\|\mathbf{e}^-(\mathbf{w})-\mathbf{e}^-(\mathbf{w}')\| \\
&\quad +\|\mathbf{a}-\mathbf{a}'\|+(1+\|\nabla\mathbf{e}^-(\mathbf{w})\|)\cdot\|\mathbf{b}-\mathbf{b}'\|\Big] \\
&\leq \frac{\sqrt{3}}{m+1}\Big[(2R_2+R_1)D^2S_2^-\cdot\|\mathbf{w}-\mathbf{w}'\|+DS_1^-\cdot\|\mathbf{w}-\mathbf{w}'\| \\
&\quad +\|\mathbf{a}-\mathbf{a}'\|+(1+DS_1^-)\cdot\|\mathbf{b}-\mathbf{b}'\|\Big] \\
&\leq \frac{\sqrt{3}\rho_0^-}{m+1}\|\mathbf{v}-\mathbf{v}'\|.
\end{aligned}$$

In conclusion, we obtain the Lipschitz gradient parameter $\rho_0=\frac{\sqrt{3}}{m+1}\max\{\rho_0^+,\rho_0^-\}$.

B. **Proof of Proposition 1.** Note that it is straight forward to see that $\Phi_\gamma(\mathbf{v},\boldsymbol{\alpha};z) := F(\mathbf{v},\boldsymbol{\alpha};z)+\frac{\gamma}{2}\|\mathbf{w}-\bar{\mathbf{w}}\|^2$ is concave w.r.t $\boldsymbol{\alpha}$ no matter what γ is chosen. We next show how to choose a baseline $\gamma_0>0$ such that when $\gamma\geq\gamma_0$, $\Phi_\gamma(\mathbf{v},\boldsymbol{\alpha};z)$ is convex w.r.t $\mathbf{v}$. Firstly, we can write $\nabla^2_{\mathbf{w}}F(\mathbf{v},\boldsymbol{\alpha};z)=\frac{C_m(\mathbf{u},z)}{m+1}\cdot xx^\top$, where

$$\begin{aligned}
C_m(\mathbf{u},z)=&\sum_{i=0}^{m}\{h_i^+(\mathbf{u},x)\mathbb{I}_{[y=1]}+h_i^-(\mathbf{u},x)\mathbb{I}_{[y=-1]}\}, \\
h_i^+(\mathbf{u},x)=&(\alpha_i-a_i)i(i-1)(L/2+\mathbf{w}^\top x)^{i-2}:=h_i^+, \\
h_i^-(\mathbf{u},x)=&(\alpha_i-b_i)\sum_{k=i}^{m}(k-i)(k-i-1)\binom{m}{k}\binom{k}{i} \\
&\cdot\frac{(m+1)\Delta^k\Phi(0)}{(2L)^k}(\frac{L}{2}-\mathbf{w}^\top x)^{k-i-2}:=h_i^-.
\end{aligned}$$

Then we write the Hessian matrix of $\Phi_\gamma(\mathbf{v}, \boldsymbol{\alpha}; z)$ w.r.t. $\mathbf{v}$ as

$$\nabla^2_{\mathbf{v}}\Phi_\gamma = \begin{pmatrix} \nabla^2_{\mathbf{w}}F + \gamma I_d & -\frac{\nabla_{\mathbf{w}}\mathbf{e}^+}{m+1} & -\frac{\nabla_{\mathbf{w}}\mathbf{e}^-}{m+1} \\ -\frac{(\nabla_{\mathbf{w}}\mathbf{e}^+)^\top}{m+1} & \frac{I_{m+1}}{m+1} & 0 \\ -\frac{(\nabla_{\mathbf{w}}\mathbf{e}^-)^\top}{m+1} & 0 & \frac{I_{m+1}}{m+1} \end{pmatrix},$$

where I_d is a $d \times d$ identity matrix and I_{m+1} is an $(m+1)\times(m+1)$ identity matrix. For simplicity, we study two cases $y = 1$ and $y = -1$ separately. On one hand, if $y = 1$, we have

$$\nabla^2_{\mathbf{v}}\Phi_\gamma = \begin{pmatrix} \nabla^2_{\mathbf{w}}F + \gamma I_d & -\frac{\nabla_{\mathbf{w}}\mathbf{e}^+}{m+1} & 0 \\ -\frac{(\nabla_{\mathbf{w}}\mathbf{e}^+)^\top}{m+1} & \frac{I_{m+1}}{m+1} & 0 \\ 0 & 0 & \frac{I_{m+1}}{m+1} \end{pmatrix},$$

where $\nabla^2_{\mathbf{w}}F = \frac{\sum_{i=0}^m h_i^+}{m+1} \cdot xx^\top$. By Schur complement, $\nabla^2_{\mathbf{v}}\Phi_\gamma \succeq 0$ is equivalent to

$$\begin{pmatrix} H^+ & 0 & 0 \\ 0 & \frac{I_{m+1}}{m+1} & 0 \\ 0 & 0 & \frac{I_{m+1}}{m+1} \end{pmatrix} \succeq 0,$$

where $H^+ = \nabla^2_{\mathbf{w}}F + \gamma I_d - \frac{(\nabla_{\mathbf{w}}\mathbf{e}^+)(\nabla_{\mathbf{w}}\mathbf{e}^+)^\top}{m+1}$. Thus we only need $H^+ \succeq 0$. Next we will estimate the lower bound of the minimal eigenvalue of H^+. Firstly, we know $\nabla^2_{\mathbf{w}}F$ has only one none-zero eigenvalue, i.e. $\frac{\sum_{i=0}^m h_i^+}{m+1} \cdot \|x\|^2$ which is lower bounded by $-\frac{(2R_1+R_2)D^2S_2^+}{m+1}$. Thus the minimal eigenvalue of $\nabla^2_{\mathbf{w}}F + \gamma I_d$ is lower bounded by $\gamma - \frac{(2R_1+R_2)D^2S_2^+}{m+1}$. Secondly, we write $(\nabla_{\mathbf{w}}\mathbf{e}^+)(\nabla_{\mathbf{w}}\mathbf{e}^+)^\top = \sum_{i=0}^m (\nabla_{\mathbf{w}}\mathbf{e}_i^+)(\nabla_{\mathbf{w}}\mathbf{e}_i^+)^\top$ and we know $(\nabla_{\mathbf{w}}\mathbf{e}_i^+)(\nabla_{\mathbf{w}}\mathbf{e}_i^+)^\top$ has only one none-zero eigenvalue, i.e. $\|\nabla_{\mathbf{w}}\mathbf{e}_i^+\|^2$. Thus the maximal eigenvalue of $(\nabla_{\mathbf{w}}\mathbf{e}^+)(\nabla_{\mathbf{w}}\mathbf{e}^+)^\top$ is upper bounded by $\sum_{i=0}^m \|\nabla_{\mathbf{w}}\mathbf{e}_i^+\|^2 = \|\nabla_{\mathbf{w}}\mathbf{e}^+\|^2 \le D^2(S_1^+)^2$. Finally, we have the minimal eigenvalue of H^+ is lower bounded by $\gamma - \frac{(2R_1+R_2)D^2S_2^+ + D^2(S_1^+)^2}{m+1}$. This means if $y = 1$, we can choose

$$\gamma \ge \frac{(2R_1 + R_2)D^2S_2^+ + D^2(S_1^+)^2}{m+1}$$

to ensure that the function $\Phi_\gamma(\mathbf{v}, \boldsymbol{\alpha})$ is convex w.r.t $\mathbf{v}$. Similarly, on the other hand, if $y = -1$, we can choose

$$\gamma \ge \frac{(R_1 + 2R_2)D^2S_2^- + D^2(S_1^-)^2}{m+1}$$

to ensure that the function $\Phi_\gamma(\mathbf{v}, \boldsymbol{\alpha})$ is convex w.r.t $\mathbf{v}$. As the deduction is almost the same as the previous case, we omit the details here. In conclusion, we have if the parameter $\gamma \ge \gamma_0$, where

$$\gamma_0 := \frac{1}{m+1}\max\{(2R_1 + R_2)D^2S_2^+ + D^2(S_1^+)^2, (R_1 + 2R_2)D^2S_2^- + D^2(S_1^-)^2\},$$

then $\Phi_\gamma(\mathbf{v}, \boldsymbol{\alpha})$ is convex w.r.t $\mathbf{v}$.

C. **Proof of Lemma 3.2.** The proof of Lemma 3.2 needs the following elementary lemma.

Lemma 5.1. *Let* $\omega_{j+1} := \operatorname{argmin}_{\omega\in\Omega}\{\omega^\top g + \frac{1}{2\eta}\|\omega-\omega_j\|^2\}$ *where* $\eta > 0$. *We have* $(\omega_j-\omega)^\top g \le \frac{\eta}{2}\|g\|^2 - \frac{1}{2\eta}\|\omega-\omega_{j+1}\|^2 + \frac{1}{2\eta}\|\omega-\omega_j\|^2$, *for any* $\omega\in\Omega$.

Proof. By the optimality condition of ω_{j+1}, for any $\omega\in\Omega$, we have

$$(\omega-\omega_{j+1})^\top g \ge \frac{1}{\eta}(\omega-\omega_{j+1})^\top(\omega_j-\omega_{j+1}). \tag{C.1}$$

We decompose left hand side of (C.1) like that

$$(\omega-\omega_j)^\top g + (\omega_j-\omega_{j+1})^\top g \ge \frac{1}{\eta}(\omega-\omega_{j+1})^\top(\omega_j-\omega_{j+1}). \tag{C.2}$$

To further decompose the right hand side of (C.1), we need the following perfect square formulation $\|\omega-\omega_j\|^2 = \|\omega-\omega_{j+1}+\omega_{j+1}-\omega_j\|^2 = \|\omega-\omega_{j+1}\|^2 - 2(\omega-\omega_{j+1})^\top(\omega_j-\omega_{j+1}) + \|\omega_j-\omega_{j+1}\|^2$. Combining this with (C.2) implies that

$$(\omega-\omega_j)^\top g + (\omega_j-\omega_{j+1})^\top g \ge \frac{1}{2\eta}(\|\omega-\omega_{j+1}\|^2 + \|\omega_j-\omega_{j+1}\|^2 - \|\omega-\omega_j\|^2).$$

Thus we have

$$\begin{aligned}(\omega_j-\omega)^\top g &\le (\omega_j-\omega_{j+1})^\top g - \frac{1}{2\eta}\|\omega_j-\omega_{j+1}\|^2 - \frac{1}{2\eta}\|\omega-\omega_{j+1}\|^2 + \frac{1}{2\eta}\|\omega-\omega_j\|^2\\ &\le \frac{\eta}{2}\|g\|^2 - \frac{1}{2\eta}\|\omega-\omega_{j+1}\|^2 + \frac{1}{2\eta}\|\omega-\omega_j\|^2,\end{aligned}$$

where the last inequality used the fact $(\omega_j-\omega_{j+1})^\top g \le \frac{\eta}{2}\|g\|^2 + \frac{1}{2\eta}\|\omega_j-\omega_{j+1}\|^2$. Thus we have proved this lemma. □

We are ready to prove Lemma 3.2.

Proof of Lemma 3.2. By the convexity of $\varphi_\gamma^t(\cdot,\boldsymbol{\alpha})$, we have, for all $\mathbf{v}\in\Omega_1$, that

$$\varphi_\gamma^t(\mathbf{v}_j^t,\boldsymbol{\alpha}_j^t) - \varphi_\gamma^t(\mathbf{v},\boldsymbol{\alpha}_j^t) \le (\mathbf{v}_j^t-\mathbf{v})^\top\nabla_{\mathbf{v}}\varphi_\gamma^t(\mathbf{v}_j^t,\boldsymbol{\alpha}_j^t).$$

Similarly by the concavity of $\varphi_\gamma^t(\mathbf{v},\cdot)$, we have, for all $\boldsymbol{\alpha}\in\Omega_2$, that

$$\varphi_\gamma^t(\mathbf{v}_j^t,\boldsymbol{\alpha}) - \varphi_\gamma^t(\mathbf{v}_j^t,\boldsymbol{\alpha}_j^t) \le (\boldsymbol{\alpha}-\boldsymbol{\alpha}_j^t)^\top\nabla_{\boldsymbol{\alpha}}\varphi_\gamma^t(\mathbf{v}_j^t,\boldsymbol{\alpha}_j^t).$$

Denote

$$g_{j+1}^t := (g_{\mathbf{v}}^{(t,j+1)}, -g_{\boldsymbol{\alpha}}^{(t,j+1)}) := (\nabla_{\mathbf{v}}\varphi_\gamma^t(\mathbf{v}_j^t,\boldsymbol{\alpha}_j^t), -\nabla_{\boldsymbol{\alpha}}\varphi_\gamma^t(\mathbf{v}_j^t,\boldsymbol{\alpha}_j^t)).$$

Thus we obtain

$$\varphi_\gamma^t(\mathbf{v}_j^t,\boldsymbol{\alpha}) - \varphi_\gamma^t(\mathbf{v},\boldsymbol{\alpha}_j^t) \le (\mathbf{v}_j^t-\mathbf{v})^\top g_{\mathbf{v}}^{(t,j+1)} - (\boldsymbol{\alpha}_j^t-\boldsymbol{\alpha})^\top g_{\boldsymbol{\alpha}}^{(t,j+1)} = (\mathbf{u}_j^t-\mathbf{u})^\top g_{j+1}^t, \tag{C.3}$$

for all $\mathbf{u} = (\mathbf{v},\boldsymbol{\alpha})\in\Omega_1\times\Omega_2$, Denote

$$G_{j+1}^t := (G_{\mathbf{v}}^{(t,j+1)}, -G_{\boldsymbol{\alpha}}^{(t,j+1)}) := (\nabla_{\mathbf{v}}\Phi_\gamma^t(\mathbf{v}_j^t,\boldsymbol{\alpha}_j^t;z_{j+1}^t), -\nabla_{\boldsymbol{\alpha}}\Phi_\gamma^t(\mathbf{v}_j^t,\boldsymbol{\alpha}_j^t;z_{j+1}^t)).$$

With this notation, the $(j+1)$-th update step in the t-th inner loop of Algorithm 1 can be written as

$$\mathbf{u}_{j+1}^t := \operatorname*{argmin}_{\mathbf{u}\in\Omega_1\times\Omega_2}\{\mathbf{u}^\top G_{j+1}^t + \frac{1}{2\eta_t}\|\mathbf{u}-\mathbf{u}_j^t\|^2\}.$$

Similarly, we define an auxiliary sequence as

$$\tilde{\mathbf{u}}_{j+1}^t := \operatorname*{argmin}_{\mathbf{u}\in\Omega_1\times\Omega_2}\{-\mathbf{u}^\top\Delta_{j+1}^t + \frac{1}{2\eta_t}\|\mathbf{u}-\tilde{\mathbf{u}}_j^t\|^2\},$$

where $\Delta_{j+1}^t = G_{j+1}^t - g_{j+1}^t$ and $\tilde{\mathbf{u}}_0^t = \mathbf{u}_0^t$. Applying Lemma 5.1 to $\mathbf{u}_{j+1}^t$ and $\tilde{\mathbf{u}}_{j+1}^t$ implies that

$$(\mathbf{u}_j^t - \mathbf{u})^\top G_{j+1}^t \leq \frac{\eta_t}{2}\|G_{j+1}^t\|^2 - \frac{1}{2\eta_t}\|\mathbf{u} - \mathbf{u}_{j+1}^t\|^2 + \frac{1}{2\eta_t}\|\mathbf{u} - \mathbf{u}_j^t\|^2, \tag{C.4}$$

$$-(\tilde{\mathbf{u}}_j^t - \mathbf{u})^\top \Delta_{j+1}^t \leq \frac{\eta_t}{2}\|\Delta_{j+1}^t\|^2 - \frac{1}{2\eta_t}\|\mathbf{u} - \tilde{\mathbf{u}}_{j+1}^t\|^2 + \frac{1}{2\eta_t}\|\mathbf{u} - \tilde{\mathbf{u}}_j^t\|^2. \tag{C.5}$$

Adding up (C.3), (C.4) and (C.5), we have

$$\begin{aligned}
&\varphi_\gamma^t(\mathbf{v}_j^t, \boldsymbol{\alpha}) - \varphi_\gamma^t(\mathbf{v}, \boldsymbol{\alpha}_j^t) \\
\leq&(\tilde{\mathbf{u}}_j^t - \mathbf{u}_j^t)^\top \Delta_{j+1}^t + \frac{\eta_t}{2}(\|G_{j+1}^t\|^2 + \|\Delta_{j+1}^t\|^2) + \frac{1}{2\eta_t}(\|\mathbf{u} - \mathbf{u}_j^t\|^2 - \|\mathbf{u} - \mathbf{u}_{j+1}^t\|^2) \\
&+ \frac{1}{2\eta_t}(\|\mathbf{u} - \tilde{\mathbf{u}}_j^t\|^2 - \|\mathbf{u} - \tilde{\mathbf{u}}_{j+1}^t\|^2).
\end{aligned} \tag{C.6}$$

Again, by the convexity and concavity of φ_γ^t, we obtain

$$\varphi_\gamma^t(\bar{\mathbf{v}}_t, \boldsymbol{\alpha}) - \varphi_\gamma^t(\mathbf{v}, \bar{\boldsymbol{\alpha}}_t) \leq \frac{1}{t}\sum_{j=0}^{t-1}\Big[\varphi_\gamma^t(\mathbf{v}_j^t, \boldsymbol{\alpha}) - \varphi_\gamma^t(\mathbf{v}, \boldsymbol{\alpha}_j^t)\Big].$$

Combining this with (C.6) implies that $\forall \mathbf{u} \in \Omega_1 \times \Omega_2$,

$$\begin{aligned}
\varphi_\gamma^t(\bar{\mathbf{v}}_t, \boldsymbol{\alpha}) - \varphi_\gamma^t(\mathbf{v}, \bar{\boldsymbol{\alpha}}_t) \leq&\frac{1}{t}\sum_{j=0}^{t-1}\Big[(\tilde{\mathbf{u}}_j^t - \mathbf{u}_j^t)^\top \Delta_{j+1}^t + \frac{\eta_t}{2}(\|G_{j+1}^t\|^2 + \|\Delta_{j+1}^t\|^2)\Big] \\
&+ \frac{1}{2t\eta_t}(\|\mathbf{u} - \mathbf{u}_0^t\|^2 + \|\mathbf{u} - \tilde{\mathbf{u}}_0^t\|^2).
\end{aligned} \tag{C.7}$$

Now taking supreme in $\mathbf{u} \in \Omega_1 \times \Omega_2$ on both sides of (C.7) implies

$$\begin{aligned}
\epsilon_t(\bar{\mathbf{u}}_t) :=& \max_{\boldsymbol{\alpha} \in \Omega_2} \varphi_\gamma^t(\bar{\mathbf{v}}_t, \boldsymbol{\alpha}) - \min_{\mathbf{v} \in \Omega_1} \varphi_\gamma^t(\mathbf{v}, \bar{\boldsymbol{\alpha}}_t) \\
\leq&\frac{1}{t}\sum_{j=0}^{t-1}\Big[(\tilde{\mathbf{u}}_j^t - \mathbf{u}_j^t)^\top \Delta_{j+1}^t + \frac{\eta_t}{2}(\|G_{j+1}^t\|^2 + \|\Delta_{j+1}^t\|^2)\Big] \\
&+ \frac{1}{2t\eta_t} \cdot \sup_{\mathbf{u} \in \Omega_1 \times \Omega_2}(\|\mathbf{u} - \mathbf{u}_0^t\|^2 + \|\mathbf{u} - \tilde{\mathbf{u}}_0^t\|^2) \\
\leq&\frac{1}{t}\sum_{j=0}^{t-1}\Big[(\tilde{\mathbf{u}}_j^t - \mathbf{u}_j^t)^\top \Delta_{j+1}^t + \frac{\eta_t}{2}(\|G_{j+1}^t\|^2 + \|\Delta_{j+1}^t\|^2)\Big] \\
&+ \frac{4}{t\eta_t}(R^2 + R_1^2 + R_2^2 + (R_1 + R_2)^2),
\end{aligned} \tag{C.8}$$

where the last inequality used the bounds of $\mathbf{v}$ and $\boldsymbol{\alpha}$. We next show how to uniformly bound $\|G_{j+1}^t\|^2$ and $\|\Delta_{j+1}^t\|^2$. For the simplicity of notation, we write $G_{\mathbf{v}}^{(t,j+1)}$, $G_{\boldsymbol{\alpha}}^{(t,j+1)}$, $g_{\mathbf{v}}^{(t,j+1)}$ and $g_{\boldsymbol{\alpha}}^{(t,j+1)}$ as $G_{\mathbf{v}}$, $G_{\boldsymbol{\alpha}}$, $g_{\mathbf{v}}$ and $g_{\boldsymbol{\alpha}}$ respectively. We also simply denote $\mathbf{e}^+ = \{e_i^+(\mathbf{w}_j^t, z_{j+1}^t)\}_{i=0}^m$ and $\mathbf{e}^- = \{e_i^-(\mathbf{w}_j^t, z_{j+1}^t)\}_{i=0}^m$. Likewise, we simply denote $\mathbf{E}^+ = \{\mathbb{E}_z[e_i^+(\mathbf{w}_j^t, z)]\}_{i=0}^m$ and $\mathbf{E}^- = \{\mathbb{E}_z[e_i^-(\mathbf{w}_j^t, z)]\}_{i=0}^m$. We write

$$\begin{aligned}
G_{\mathbf{w}} :=& \nabla_{\mathbf{w}} \Phi_\gamma^t(\mathbf{v}_j^t, \boldsymbol{\alpha}_j^t; z_{j+1}^t) \\
=& \frac{1}{m+1}\Big\{(\nabla_{\mathbf{w}}\mathbf{e}^+)(\boldsymbol{\alpha}_j^t - \mathbf{a}_j^t) + (\nabla_{\mathbf{w}}\mathbf{e}^-)(\boldsymbol{\alpha}_j^t - \mathbf{b}_j^t)\Big\} + \gamma(\mathbf{w}_j^t - \bar{\mathbf{w}}_{t-1}), \\
G_{\mathbf{a}} :=& \nabla_{\mathbf{a}} \Phi_\gamma^t(\mathbf{v}_j^t, \boldsymbol{\alpha}_j^t; z_{j+1}^t) = \frac{1}{m+1}\{\mathbf{a}_j^t - \mathbf{e}^+\},
\end{aligned}$$

$$G_{\mathbf{b}} := \nabla_{\mathbf{b}} \Phi_\gamma^t(\mathbf{v}_j^t, \boldsymbol{\alpha}_j^t; z_{j+1}^t) = \frac{1}{m+1}\{\mathbf{b}_j^t - \mathbf{e}^-\},$$
$$G_{\boldsymbol{\alpha}} := \nabla_{\boldsymbol{\alpha}} \Phi_\gamma^t(\mathbf{v}_j^t, \boldsymbol{\alpha}_j^t; z_{j+1}^t) = \frac{1}{m+1}\{\mathbf{e}^+ + \mathbf{e}^- - \boldsymbol{\alpha}_j^t\}.$$

Combining these with the facts

$$\|\mathbf{e}^+\| \le \sum_{i=0}^m L^i = R_1, \quad \|\nabla_{\mathbf{w}} \mathbf{e}^+\| \le D S_1^+, \quad \|\nabla_{\mathbf{w}} \mathbf{e}^-\| \le D S_1^-,$$
$$\|\mathbf{e}^-\| \le \sum_{i=0}^m \sum_{k=i}^m \binom{m}{k}\binom{k}{i} \frac{(m+1)|\Delta^k \varphi(0)|}{2^k L^i} = R_2,$$

implies

$$\begin{aligned}\|G_{\mathbf{w}}\|^2 &\le 3(m+1)^{-2}\Big[\|\nabla_{\mathbf{w}} \mathbf{e}^+\|^2 \|\boldsymbol{\alpha}_j^t - \mathbf{a}_j^t\|^2 + \|\nabla_{\mathbf{w}} \mathbf{e}^-\|^2 \|\boldsymbol{\alpha}_j^t - \mathbf{b}_j^t\|^2\Big] + 12\gamma^2 R^2 \\ &\le 3(m+1)^{-2}\Big[(DS_1^+)^2(6R_1^2 + 4R_2^2) + (DS_1^-)^2(6R_2^2 + 4R_1^2)\Big] + 12\gamma^2 R^2,\end{aligned}$$

and

$$\begin{aligned}\|G_{\mathbf{a}}\|^2 &\le 2(m+1)^{-2}\Big(\|\mathbf{a}_j^t\|^2 + \|\mathbf{e}^+\|^2\Big) \le 4(m+1)^{-2} R_1^2, \\ \|G_{\mathbf{b}}\|^2 &\le 2(m+1)^{-2}\Big(\|\mathbf{b}_j^t\|^2 + \|\mathbf{e}^-\|^2\Big) \le 4(m+1)^{-2} R_2^2, \\ \|G_{\boldsymbol{\alpha}}\|^2 &\le \Big(\frac{3}{m+1}\Big)^2 (R_1^2 + R_2^2) := M_2^2.\end{aligned}$$

Moreover we have

$$\begin{aligned}\|G_{\mathbf{v}}\|^2 &= \|G_{\mathbf{w}}\|^2 + \|G_{\mathbf{a}}\|^2 + \|G_{\mathbf{b}}\|^2 \\ &\le 4(m+1)^{-2}(R_1^2 + R_2^2) + 12\gamma^2 R^2 + 6(m+1)^{-2}\Big[3(DS_1^+)^2 + 2(DS_1^-)^2\Big] R_1^2 \\ &\quad + 6(m+1)^{-2}\Big[2(DS_1^+)^2 + 3(DS_1^-)^2\Big] R_2^2 := \ M_1^2.\end{aligned}$$

Then replacing $\mathbf{e}^+$ and $\mathbf{e}^-$ with $\mathbf{E}^+$ and $\mathbf{E}^-$ respectively and using almost the same deductions as the above, we have

$$\|g_{\mathbf{v}}\|^2 \le M_1^2, \ \|g_{\boldsymbol{\alpha}}\|^2 \le M_2^2.$$

Therefore we have $\|G_{j+1}^t\|^2 \le M_1^2 + M_2^2$, and $\|\Delta_{j+1}^t\|^2] \le 2\|G_{j+1}^t\|^2 + 2\|g_{j+1}^t\|^2 \le 4(M_1^2 + M_2^2)$. Combining this with (C.8) implies

$$\begin{aligned}\epsilon_t(\bar{\mathbf{u}}_t) &:= \max_{\boldsymbol{\alpha} \in \Omega_2} \varphi_\gamma^t(\bar{\mathbf{v}}_t, \boldsymbol{\alpha}) - \min_{\mathbf{v} \in \Omega_1} \varphi_\gamma^t(\mathbf{v}, \bar{\boldsymbol{\alpha}}_t) \\ &\le \frac{1}{t}\sum_{j=0}^{t-1}\Big[(\tilde{\mathbf{u}}_j^t - \mathbf{u}_j^t)^\top \Delta_{j+1}^t\Big] + \frac{5\eta_t}{2}(M_1^2 + M_2^2) \\ &\quad + \frac{4}{t\eta_t}(R^2 + R_1^2 + R_2^2 + (R_1 + R_2)^2), \qquad \text{(C.9)}\end{aligned}$$

It suffices to bound $\mathbb{E}_{j+1}^t[(\tilde{\mathbf{u}}_j^t - \mathbf{u}_j^t)^\top \Delta_{j+1}^t]$ where the expectation $\mathbb{E}_{j+1}^t$ is taken w.r.t. all the randomness from $(z_1^1, z_1^2, z_2^2, \ldots, z_1^t, z_2^t, \ldots, z_{j+1}^t) := [z_{j+1}^t]$. Take into account that $\tilde{\mathbf{u}}_j^t$ and $\mathbf{u}_j^t$ are deterministic functions of $[z_j^t]$. And note that the conditional expectation of Δ_{j+1}^t, given $[z_j^t]$, vanishes, i.e. $\mathbb{E}_{(j+1|j)}^t[\Delta_{j+1}^t \Big| [z_j^t]] = 0$. Thus we have

$$\mathbb{E}_{j+1}^t\Big\{(\tilde{\mathbf{u}}_j^t - \mathbf{u}_j^t)^\top \Delta_{j+1}^t\Big\} = \mathbb{E}_{j+1}^t\Big\{\mathbb{E}_{(j+1|j)}^t\Big[(\tilde{\mathbf{u}}_j^t - \mathbf{u}_j^t)^\top \Delta_{j+1}^t \Big| \ [z_j^t]\Big]\Big\}$$

$$=\mathbb{E}_{j+1}^t\Big\{(\tilde{\mathbf{u}}_j^t-\mathbf{u}_j^t)^\top\mathbb{E}_{(j+1|j)}^t\Big[\Delta_{j+1}^t\Big|\ [z_j^t]\Big]\Big\}=0.$$

Taking expectation on both sides of (C.9), we conclude that

$$\begin{aligned}\mathbb{E}[\epsilon_t(\bar{\mathbf{u}}_t))]\leq&\frac{1}{t}\sum_{j=0}^{t-1}\mathbb{E}_{j+1}^t[(\tilde{\mathbf{u}}_j^t-\mathbf{u}_j^t)^\top\Delta_{j+1}^t]+\frac{4}{t\eta_t}(R^2+R_1^2+R_2^2+(R_1+R_2)^2)\\&+\frac{5\eta_t}{2}(M_1^2+M_2^2)\\=&\frac{4}{t\eta_t}(R^2+R_1^2+R_2^2+(R_1+R_2)^2)+\frac{5\eta_t}{2}(M_1^2+M_2^2).\end{aligned}$$

Then by choosing $\eta_t=\beta/\sqrt{t}$ and denoting $C_1:=4\beta^{-1}[R^2+R_1^2+R_2^2+(R_1+R_2)^2]+5\beta(M_1^2+M_2^2)/2$, we obtain the desired result. □

D. **Final convergence rate.** In this subsection, we investigate the final convergence of the output $\widetilde{\mathbf{w}}_T$ of SAUC on the original objective function g of AUC maximization given by (2.2) through theoretically optimal choice of the degree m of Bernstein polynomials.

The proof of Theorem 3.3 requires the following proposition on the approximation error of Bernstein polynomials [25, 26]. Recall that the surrogate loss ℓ on general interval $[-L,L]$ induces a function φ on the unit interval $[0,1]$ by interchanging $u=\frac{L+s}{2L}$ and letting $\varphi(u)=\ell(s)$ for any $s\in[-L,L]$.

Proposition 2. *Given any surrogate loss function ℓ on $[-L,L]$, with Lipschitz constant G. Its Bernstein approximation is uniformly bounded by*

$$\sup_{s\in[-L,L]}|\ell(s)-B_m(\varphi;\frac{L+s}{2L})|\leq\frac{GL}{\sqrt{m}}.\tag{D.1}$$

Proof. It is straightforward to see that if ℓ is Lipschitz continuous with constant G then φ is also Lipschitz continuous with constant $2GL$. Furthermore, we have $|\ell(s)-B_m(\varphi;\frac{L+s}{2L})|=|\varphi(u)-B_m(\varphi;u)|$ for any $s\in[-L,L]$.

Notice that $B_m(\varphi;u)=\mathbb{E}[\varphi(\frac{X}{m})]$, where X is a random variable obeying binomial distribution out of m Bernoulli trials with success probability u. Hence, for any $u\in[0,1]$, we have

$$\begin{aligned}|\varphi(u)-B_m(\varphi;u)|=&|\varphi(u)-\mathbb{E}[\varphi(\frac{X}{m})]|=|\mathbb{E}[\varphi(u)-\varphi(\frac{X}{m})]|\\\leq&\mathbb{E}[|\varphi(u)-\varphi(\frac{X}{m})|]\leq\mathbb{E}[2GL|u-\frac{X}{m}|]=2GL\mathbb{E}[|u-\frac{X}{m}|].\end{aligned}$$

The first inequality is Jensen's inequality. The second inequality holds because φ is Lipschitz continuous. On the other hand, The variance of X satisfies, $\mathbb{E}[|X-mu|^2]=mu(1-u)\leq m/4$. It follows that

$$\begin{aligned}|\varphi(u)-B_m(\varphi;u)|\leq&2GL\mathbb{E}[|u-\frac{X}{m}|]\leq2GL\sqrt{\mathbb{E}[|u-\frac{X}{m}|^2]}=\frac{2GL}{m}\sqrt{\mathbb{E}[mu-X|^2]}\\\leq&\frac{GL}{\sqrt{m}}.\end{aligned}$$

The first inequality is Cauchy-Schwartz inequality. And the proposition follows by taking the supremum. □

Now we are ready to prove Theorem 3.3.

Proof of Theorem 3.3. Notice that

$$\mathbb{E}[g(\widetilde{\mathbf{w}}_T) - \inf_{\|\mathbf{w}\|\le R} g(\mathbf{w})] \le \overbrace{2\sup_{\|\mathbf{w}\|\le R}|f(\mathbf{w}) - g(\mathbf{w})|}^{\text{Part I}} + \overbrace{|\mathbb{E}[f(\widetilde{\mathbf{w}}_T) - f(\mathbf{w}^*)]|}^{\text{Part II}}. \quad \text{(D.2)}$$

Below we will bound both parts of (D.2). For Part I, recall that $g(\mathbf{w}) = \mathbb{E}\big[\ell(\mathbf{w}^\top x - \mathbf{w}^\top x')\mathbb{I}_{[y=1]}\mathbb{I}_{[y'=-1]}\big]$ and

$$\begin{aligned} f(\mathbf{w}) =& \frac{1}{(m+1)}\sum_{i=0}^{m}\mathbb{E}\big[f_i(\mathbf{w};x)\mathbb{I}_{[y=1]}\tilde{f}_i(\mathbf{w};x')\mathbb{I}_{[y'=-1]}\big] \\ =& \mathbb{E}\big[B_m\big(\varphi;\frac{L+\mathbf{w}^\top x - \mathbf{w}^\top x'}{2L}\big)\mathbb{I}_{[y=1]}\mathbb{I}_{[y'=-1]}\big]. \end{aligned}$$

Apply Proposition 2 we have

$$\begin{aligned} &\sup_{\|\mathbf{w}\|\le R}|f(\mathbf{w}) - g(\mathbf{w})| \\ \le& \sup_{\|\mathbf{w}\|\le R,\|x\|\le D}\Big|\mathbb{E}\big[\big(\ell(\mathbf{w}^\top x - \mathbf{w}^\top x') - B_m(\varphi;\frac{L+\mathbf{w}^\top x - \mathbf{w}^\top x'}{2L})\big)\mathbb{I}_{[y=1]}\mathbb{I}_{[y'=-1]}\big]\Big| \\ \le& \frac{GL}{\sqrt{m}} = \frac{2GRD}{\sqrt{m}}. \end{aligned} \quad \text{(D.3)}$$

For Part II, by Theorem 3.1, one has $\mathbb{E}[f(\widetilde{\mathbf{w}}_T) - f(\mathbf{w}^*)] \le \sqrt{C_m/\sqrt{T}}$ where

$$\begin{aligned} C_m &= 30(L_1 + 2R\gamma)^2(L_1R + C_1)\gamma^{-1}, \quad L_1 = D(S_1^+R_2 + S_1^-R_1)/(m+1) \\ C_1 &= 4\beta^{-1}[R^2 + R_1^2 + R_2^2 + (R_1+R_2)^2] + 5\beta(M_1^2 + M_2^2)/2, \\ M_1^2 &= 4(m+1)^{-2}(R_1^2+R_2^2) + 12\gamma^2R^2 + 6(m+1)^{-2}\Big[3(DS_1^+)^2 + 2(DS_1^-)^2\Big]R_1^2 \\ &\quad + 6(m+1)^{-2}\Big[2(DS_1^+)^2 + 3(DS_1^-)^2\Big]R_2^2, \\ M_2^2 &= \Big(\frac{3}{m+1}\Big)^2(R_1^2 + R_2^2). \end{aligned}$$

To evaluate C_m, we will need the following quantities. Firstly, the k-th order forward difference is upper bounded by

$$|\Delta^k\varphi(0)| = |\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}\varphi(\frac{j}{m})| \le 2GL\sum_{j=0}^{k}\binom{k}{j} = 4GRD\cdot 2^k,$$

where we use the fact that $(a+b)^k = \sum_{j=0}^{k}\binom{k}{j}a^jb^{k-j}$. Secondly, the bound on $\|\mathbf{a}\|$ and $\|\mathbf{b}\|$ can be upper bounded by

$$R_1 = \sum_{i=0}^{m}L^i \le (m+1)\max\{1,(2RD)^m\} \le (2+4RD)^m := B_1^m.$$

Likewise, there exists B_2 independent of m such that

$$\begin{aligned} R_2 =& \sum_{i=0}^{m}\sum_{k=i}^{m}\binom{m}{k}\binom{k}{i}\frac{(m+1)|\Delta^k\varphi(0)|}{2^kL^i} \\ \le& 2GL(m+1)\max\{1,\frac{1}{L^m}\}\sum_{i=0}^{m}\sum_{k=i}^{m}\binom{m}{k}\binom{k}{i} \end{aligned}$$

$$=4GRD(m+1)\max\{1,\frac{1}{(2RD)^m}\}3^m \le B_2^m,$$

the second equality we use the fact that $(a+b+c)^m = \sum_{i=0}^m \sum_{k=i}^m \binom{m}{k}\binom{k}{i} a^i b^{k-i} c^{m-k}$. Similarly, upper bounds on the gradients and the hessian of f_i and $\tilde{f}_i$ can also be derived as

$$S_1^+ \le m(m+1)\max\{1,(2RD)^m\} \le B_3^m,\ S_1^- \le 2Gm(m+1)\max\{1,\frac{1}{(2RD)^m}\}3^m \le B_4^m,$$

$$S_2^+ \le m^3\max\{1,(2RD)^m\} \le B_5^m, \quad S_2^- \le \frac{G}{2RD}m^3\max\{1,\frac{1}{(2RD)^m}\}3^m := B_6^m.$$

Recall that, $\gamma \ge \gamma_0$, hence

$$\begin{aligned}\gamma \ge& \frac{1}{m+1}\max\{(2R_1+R_2)D^2S_2^+ + D^2(S_1^+)^2, (R_1+2R_2)D^2S_2^- + D^2(S_1^-)^2\}\\ \ge& \frac{1}{m+1}D^2(S_1^+)^2 \ge \frac{m}{2}\min\{1,(2RD)^m\} :\ge B_7^m.\end{aligned}$$

If one choose $\gamma = \gamma_0$, it follows

$$\begin{aligned}\gamma \le& \frac{1}{m+1}\left((2R_1+R_2)D^2S_2^+ + D^2(S_1^+)^2\right) + \left((R_1+2R_2)D^2S_2^- + D^2(S_1^-)^2\right)\\ \le& \frac{1}{m+1}\left((2B_1+B_2)D^2B_5 + D^2B_3^2\right) + \left((B_1+2B_2)D^2B_6 + D^2B_4^2\right) \le B_8^m.\end{aligned}$$

Now, one has

$$\begin{aligned}L_1 \le& D((B_3B_2)^m + (B_4B_1)^m) \le B_9^m,\\ M_1^2 \le& 4(m+1)^{-2}(B_1^2+B_2^2) + 12B_8^2R^2 + 6(m+1)^{-2}\left[3(DB_3)^2 + 2(DB_4)^2\right]B_1^2\\ &+ 6(m+1)^{-2}\left[2(DB_3)^2 + 3(DB_4)^2\right]B_2^2 \le B_{10}^m,\end{aligned}$$

In addition, there exists constants B_{11} and B_{12} such that

$$\begin{aligned}M_2^2 \le& \left(\frac{3}{m+1}\right)^2(B_1^2+B_2^2) \le B_{11}^m,\\ C_1 \le& 4\beta^{-1}[R^2 + B_1^2 + B_2^2 + (B_1+B_2)^2] + 5\beta(B_{10}^2 + B_{11}^2)/2 \le B_{12}^m.\end{aligned}$$

Combining all the above estimations, we have

$$\begin{aligned}C_m \le& 60(L_1^2(L_1R+C_1))\gamma^{-1} + 240R^2(L_1R+C_1)\gamma\\ \le& 60((B_9^2)^m(B_9^mR + B_{12}^m))B_7^{-m} + 240R^2(B_9^mR + B_{12}^m)B_8^m\\ \le& (60R + B_1 + 240R^3)\max(B_9^3B_7^{-1}, 2B_7^{-1}, B_9B_8, B_8B_{12})^m.\end{aligned}$$

Letting $\widetilde{B} = (60R + B_1 + 240R^3)$ and $\widetilde{C} = \max(B_9^3B_7^{-1}, 2B_7^{-1}, B_9B_8, B_8B_{12})$ and combine this result with (D.3), the theorem follows. □

REFERENCES

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

[2] A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognit.*, **30** (1997), 1145–1159.

[3] T. Calders and S. Jaroszewicz, Efficient AUC optimization for classification in *PKDD*, Vol. 4702, Springer, (2007), 42–53.

[4] C. C. Chang and C. J. Lin, LIBSVM: a library for support vector machines, *ACM Trans. Intell. Syst. Technol.*, **2** (2011), 21 pp.

[5] S. Clémençon, G. Lugosi and N. Vayatis, Ranking and empirical minimization of U-statistics, *Ann. Statist.*, **36** (2008), 844–874.
[6] C. Cortes and M. Mohri, AUC optimization vs. error rate minimization, in *Advances in Neural Information Processing Systems*, (2004), 313–320.
[7] D. Davis and D. Drusvyatskiy, Stochastic model-based minimization of weakly convex functions, *SIAM J. Optim.*, **29** (2019), 207–239.
[8] D. Davis and B. Grimmer, Proximally Guided Stochastic Subgradient Method for Nonsmooth, Nonconvex Problems, *SIAM J. Optim.*, **29** (2019), 1908–1930.
[9] Dheeru, Dua and Karra Taniskidou, Efi, *UCI Machine Learning Repository*, University of California, Irvine, School of Information and Computer Sciences, 2017. Available from: `http://archive.ics.uci.edu/ml`.
[10] D. Drusvyatskiy, The proximal point method revisited, preprint, `arXiv:1712.06038`.
[11] T. Fawcett, An introduction to ROC analysis, *Pattern Recognit. Lett.*, **27** (2006), 861–874.
[12] W. Gao, R. Jin, S. Zhu and Z. H. Zhou, One-pass AUC optimization in *International Conference on Machine Learning*, (2013), 906–914.
[13] W. Gao and Z. H. Zhou, On the Consistency of AUC Pairwise Optimization, in *IJCAI*, (2015), 939–945.
[14] J. A. Hanley and B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, **143** (1982), 29–36.
[15] A. Herschtal and B. Raskutti, Optimising area under the ROC curve using gradient descent, in *Proceedings of the 21st International Conference on Machine Learning*, ACM, (2004), 49.
[16] T. Joachims, A support vector method for multivariate performance measures, in *Proceedings of the 22nd International Conference on Machine Learning*, ACM, (2005), 377–384.
[17] 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*, (2013), 441–449.
[18] S. Lacoste-Julien, M. Schmidt and F. Bach, A simpler approach to obtaining an O (1/t) convergence rate for the projected stochastic subgradient method, preprint, `arXiv:1212.2002`.
[19] J. Lin and L. Rosasco, Optimal learning for multi-pass stochastic gradient methods, in *Advances in Neural Information Processing Systems*, (2016), 4556–4564.
[20] M. Liu, Z. Yuan, Y. Ying and T. Yang, Stochastic AUC Maximization with Deep Neural Networks, in *International Conference on Learning Representations (ICLR)*, 2020.
[21] 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*, (2018), 3195–3203.
[22] M. Natole, Y. Ying and S. Lyu, Stochastic proximal algorithms for AUC maximization in *International Conference on Machine Learning*, (2018), 3707–3716.
[23] A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro, Robust stochastic approximation approach to stochastic programming, *SIAM J. Optim.*, **19** (2009), 1574–1609.
[24] E. A. Nurminskii, The quasigradient method for the solving of the nonlinear programming problems, *Cybernetics*, **9** (1973), 145–150.
[25] G. M. Phillips, *Interpolation and Approximation by Polynomials*, Vol. 14, Springer Science & Business Media, 2003.
[26] M. J. D. Powell, *Approximation Theory and Methods*, Cambridge University Press, 1981.
[27] H. Rafique, M. Liu, Q. Lin and T. Yang, Non-Convex Min-Max Optimization: Provable Algorithms and Applications in Machine Learning, preprint, `arXiv:1810.02060`.
[28] A. Rakhlin, O. Shamir and K. Sridharan, Making gradient descent optimal for strongly convex stochastic optimization, in *Proceedings of the 29th International Conference on Machine Learning*, (2012), 449–456.
[29] O. Shamir and T. Zhang, Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes, in *International Conference on Machine Learning*, (2013), 71–79.
[30] N. Srebro, A. Tewari, Stochastic optimization for machine learning, *CML Tutorial*, (2010).
[31] Y. Wang, R. Khardon, D. Pechyony and R. Jones, Generalization bounds for online learning algorithms with pairwise loss functions, in *Conference on Learning Theory*, (2012), 13.
[32] Y. Ying, L. Wen and S. Lyu, Stochastic online AUC maximization, in *Advances in Neural Information Processing Systems*, 2016.
[33] Y. Ying and D. X. Zhou, Online regularized classification algorithms, *IEEE Trans. Inform. Theory*, **52** (2006), 4775–4788.

[34] Y. Ying and D. X. Zhou, Online pairwise learning algorithms, *Neural Comput.*, **28** (2016), 743–777.
[35] X. Zhang, A. Saha and S. V. N. Vishwanathan, Smoothing multivariate performance measures, *J. Mach. Learn. Res.*, **13** (2012), 3623–3680.
[36] P. Zhao, R. Jin, T. Yang and S. C. Hoi, Online AUC maximization in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011.

Received October 2019; revised January 2020.

E-mail address: zyang6@albany.edu
E-mail address: 16482530@life.hkbu.edu.hk
E-mail address: yying@albany.edu
E-mail address: xmyuan@hku.hk