Stochastic Proximal Algorithms for AUC Maximization

Michael Natole Jr. [1] **Yiming Ying** [1] **Siwei Lyu** [2]

Abstract

Stochastic optimization algorithms such as stochastic gradient descent (SGD) update the model sequentially with cheap per-iteration costs, making them amenable for large-scale data analysis. Most of the existing studies focus on the classification accuracy. However, these can not be directly applied to the important problems of maximizing the Area under the ROC curve (AUC) in imbalanced classification and bipartite ranking. In this paper, we develop a novel stochastic proximal algorithm for AUC maximization which is referred to as SPAM. Compared with the previous literature, our algorithm SPAM applies to a non-smooth penalty function, and achieves a convergence rate of $\mathcal{O}(\frac{\log t}{t})$ for strongly convex functions while both space and per-iteration costs are of one datum.

1. Introduction

Stochastic gradient algorithms (Robbins & Monro, 1951; Bottou & Cun, 2004; Srebro & Tewari, 2010; Moulines & Bach, 2011; Duchi et al., 2011) and online learning algorithms (e.g. (Bottou & Cun, 2004; Srebro & Tewari, 2010; Shalev-Shwartz et al., 2011; Hazan & Kale, 2012; Rakhlin et al., 2012a; Orabona, 2014)) can update the model sequentially with computationally cheap per-iteration costs, making them amenable for large-scale streaming data analysis. Most of the existing studies focus on classification error or prediction accuracy where the empirical objective function is a summation of losses over individual samples.

However, accuracy is not suitable for important learning tasks such as imbalanced classification (Elkan, 2001). Area under the ROC curve (AUC) (Hanley & McNeil, 1982; Bradley, 1997; Fawcett, 2006) is a widely used metric for measuring the performance in these tasks. In particular, minimization of the rank loss in bipartite ranking is equivalent to maximizing the AUC criterion (Kotlowski et al., 2011). There are considerable efforts (Herschtal & Raskutti, 2004; Rakotomamonjy, 2004; Joachims, 2005; Zhang et al., 2012) that have been devoted to developing batch AUC maximization algorithms. These appealing algorithms have a convergence rate of $\mathcal{O}\big(\min\big(\frac{1}{\varepsilon}, \frac{1}{\sqrt{\lambda\varepsilon}}\big)\big)$, 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 sequentially.

Recently, there is considerable progress on online learning algorithms (Zhao et al., 2011; Wang et al., 2012; Kar et al., 2013) for AUC maximization. Due to the fact that the empirical risk for AUC maximization is a summation of pairwise losses over pairs of samples, such algorithms, at time t, need to pair the currently-received data (x_t, y_t) with all previous data $\{(x_i, y_i) : i = 1, \ldots, t-1\}$. As such, one needs to access all previous examples which leads to expensive space and per-iteration complexity of $\mathcal{O}(td)$ for d-dimensional data at iteration t. The studies (Zhao et al., 2011; Kar et al., 2013) introduced the technique of buffering to alleviate the above hurdle which reduces the per-iteration space and time complexity to $\mathcal{O}(Bd)$. However, to achieve good generalization performance, the size B needs to be sufficiently large which is typically of $\mathcal{O}(\sqrt{T})$ if the size T of the finite training data is known. The work (Gao et al., 2013) requires to update the covariance matrix of the training data with the space and per-iteration complexity $\mathcal{O}(d^2)$ which is inefficient for high-dimensional data.

The most recent work (Ying et al., 2016) reformulated the problem of AUC maximization with the least square loss as a stochastic saddle point problem (SPP). They proposed an online learning algorithm which conducts stochastic gradient descent/ascent on both the primal and dual variables. The convergence of such first-order primal-dual algorithms for solving stochastic SPPs is at most $\mathcal{O}(\frac{1}{\sqrt{t}})$ as argued in (*e.g.* Chen et al. (2014)). This is, however, inferior to the optimal rate of $\mathcal{O}(\frac{1}{t})$, up to a logarithmic term, of SGDs for the accuracy as a performance measure (Rakhlin et al., 2012a; Shamir & Zhang, 2013). In addition, the work

[1] Department of Mathematics and Statistics, SUNY at Albany, Albany, NY, USA [2] Department of Computer Science, SUNY at Albany, Albany, NY, USA. Correspondence to: Yiming Ying <yying@albany.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

(Ying et al., 2016) only considered smooth penalty terms (*i.e.*, the Frobenius norm).

In this paper, we propose a novel stochastic proximal algorithm for AUC maximization which we refer to as SPAM. The algorithm SPAM applies to general non-smooth regularization terms. In particular, we show under the assumption of strong convexity that SPAM can achieve a convergence rate of $\mathcal{O}(\frac{\log t}{t})$. The time and space complexities of our new algorithm are of one datum. To the best of our knowledge, this is the first stochastic (online) algorithm for AUC maximization with convergence rate of $\mathcal{O}(\frac{\log t}{t})$ while per-iteration and space complexities are of one datum $\mathcal{O}(d)$.

The paper is organized as follows. The next section introduces the problem of AUC maximization and our proposed algorithm. Section 3 establishes the convergence of our algorithm. We validate the performance of our algorithm in Section 4. The paper is concluded in Section 5.

2. Formulation and Algorithm

For a linear scoring function $g(x) = \mathbf{w}^\top x$, its AUC score, denoted by $\text{AUC}(\mathbf{w})$, is the probability of a random positive sample ranking higher than a random negative sample (Hanley & McNeil, 1982; Clémençon et al., 2008). To be specific, suppose $z = (x, y)$ and $z' = (x', y')$ are independently drawn from an unknown distribution ρ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is a bounded domain and $\mathcal{Y} = \{\pm 1\}$. Then, the AUC score is given by

$$\begin{aligned}\text{AUC}(\mathbf{w}) &= \Pr(\mathbf{w}^\top x \geq \mathbf{w}^\top x' | y = 1, y' = -1) \\ &= \mathbb{E}\big[\mathbb{I}_{[\mathbf{w}^\top (x - x') \geq 0]} \big| y = 1, y' = -1\big]. \qquad (1)\end{aligned}$$

In practice, one often replaces the indicator function $\mathbb{I}_{[\cdot]}$ by a *convex surrogate loss* $\phi : \mathbb{R} \to \mathbb{R}^+$ which satisfies $\mathbb{I}_{[\mathbf{w}^\top (x-x')<0]} \leq \phi(\mathbf{w}^\top (x - x'))$. Common choices are the least square loss, $\phi(s) = (1 - s)^2$, or the hinge loss, $\phi(s) = (1 - s)_+$. Throughout the paper, we focus on the least square loss as the hinge loss is not statistically consistent (Gao & Zhou, 2015). To summarize, we consider the following regularization framework for AUC maximization:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \Big\{ p(1-p)\mathbb{E}\big[(1 - \mathbf{w}^\top (x - x'))^2 \big| y = 1, y' = -1\big] + \Omega(\mathbf{w}) \Big\}. \qquad (2)$$

Here, $p = \Pr(y = 1)$ and $\Omega(\cdot)$ is a convex penalty term. The constant $p(1 - p)$ is introduced for simplicity of formulation to cancel the denominator appeared in the conditional expectation $\mathbb{E}\big[(1 - \mathbf{w}^\top (x - x'))^2 \big| y = 1, y' = -1\big] = \frac{1}{p(1-p)} \iint (1 - \mathbf{w}^\top (x - x'))^2 \mathbb{I}_{[y=1]}\mathbb{I}_{[y'=-1]} d\rho(x, y) d\rho(x', y')$. The paper (Ying et al., 2016) considered the case when $\mathbf{w}$ is restricted to a bounded ball with radius R, *i.e.*, $\Omega(\mathbf{w}) = 0$ if $\|\mathbf{w}\| \leq R$ and $\Omega(\mathbf{w}) = \infty$ otherwise. Throughout this paper, we assume that Ω is strongly convex with parameter $\beta > 0$, *i.e.*, for any $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$, $\Omega(\mathbf{w}) \geq \Omega(\mathbf{w}') + \langle \partial\Omega(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \rangle + \frac{\beta}{2}\|\mathbf{w} - \mathbf{w}'\|^2$. Examples of such penalty terms include the Frobenius norm, $\Omega(\mathbf{w}) = \beta\|\mathbf{w}\|^2$, or elastic net (Zou & Hastie, 2005), $\Omega(\mathbf{w}) = \beta\|\mathbf{w}\|^2 + \nu\|\mathbf{w}\|_1$, where β and ν are positive regularization parameters.

2.1. Equivalent Formulation

We can establish a similar theorem for formulation. Here, the proof is generalized from (Ying et al., 2016) for the general regularization framework (2). The present proof is much simpler and more intuitive.

Theorem 1. *The AUC optimization* (2) *in the linear case is equivalent to*

$$\min_{\mathbf{w},a,b} \max_{\alpha\in\mathbb{R}} \big\{ \mathbb{E}[F(\mathbf{w}, a, b, \alpha; z)] + \Omega(\mathbf{w}) \big\}, \qquad (3)$$

where the expectation is with respect to $z = (x, y)$*, and* $F(\mathbf{w}, a, b, \alpha; z) = (1 - p)(\mathbf{w}^\top x - a)^2\mathbb{I}_{[y=1]} + p(\mathbf{w}^\top x - b)^2\mathbb{I}_{[y=-1]} + 2(1 + \alpha)\mathbf{w}^\top x(p\mathbb{I}_{[y=-1]} - (1 - p)\mathbb{I}_{[y=1]}) - p(1 - p)\alpha^2$.

Proof. Specifically, the double integral mainly comes from the multiplication of two single integrals:

$$\begin{aligned}&\mathbb{E}[(1 - \mathbf{w}^\top (x - x'))^2 | y = 1, y' = -1] \\ &= 1 - 2\mathbb{E}[\mathbf{w}^\top x | y = 1] + 2\mathbb{E}[\mathbf{w}^\top x' | y' = -1] \\ &+ (\mathbb{E}[\mathbf{w}^\top x | y = 1] - \mathbb{E}[\mathbf{w}^\top x' | y' = -1])^2 \\ &+ \text{Var}[\mathbf{w}^\top x | y = 1]) + \text{Var}[\mathbf{w}^\top x' | y' = -1]). \qquad (4)\end{aligned}$$

Observe the fact that

$$\begin{aligned}(\mathbb{E}[\mathbf{w}^\top x | y = 1] - \mathbb{E}[\mathbf{w}^\top x' | y' = -1])^2 = \max_\alpha \{-\alpha^2 \\ + 2\alpha(\mathbb{E}[\mathbf{w}^\top x' | y' = -1] - \mathbb{E}[\mathbf{w}^\top x | y = 1])\}. \qquad (5)\end{aligned}$$

In addition,

$$\text{Var}[\mathbf{w}^\top x | y = 1] = \min_a \mathbb{E}[(\mathbf{w}^\top x - a)^2 | y = 1], \qquad (6)$$

and

$$\text{Var}[\mathbf{w}^\top x' | y' = -1] = \min_b \mathbb{E}[(\mathbf{w}^\top x' - b)^2 | y' = -1]. \qquad (7)$$

It is easy to see that the optima for (6), (7), and (5) are respectively achieved at

$$a(\mathbf{w}) = \mathbf{w}^\top \mathbb{E}[x | y = 1], \quad b(\mathbf{w}) = \mathbf{w}^\top \mathbb{E}[x | y = -1], \qquad (8)$$

$$\alpha(\mathbf{w}) = \mathbf{w}^\top (\mathbb{E}[x | y' = -1] - \mathbb{E}[x | y = 1]). \qquad (9)$$

Algorithm 1 Stochastic Proximal AUC Maximization (SPAM)

```
Input: Step sizes $\{\eta_t > 0 : t \in \mathbb{N}\}$
Initialize $\mathbf{w}_1 \in R^d$.
for $t = 1$ to $T$ do
    Receive sample $z_t = (x_t, y_t)$
    Compute $a(\mathbf{w}_t)$, $b(\mathbf{w}_t)$, and $\alpha(\mathbf{w}_t)$ according to (8) and (9).
    $\hat{\mathbf{w}}_{t+1} = \mathbf{w}_t - \eta_t \partial_1 F(\mathbf{w}_t, a(w_t), b(\mathbf{w}_t), \alpha(\mathbf{w}_t); z_t)$
    $\mathbf{w}_{t+1} = \mathrm{prox}_{\eta_t \Omega}(\hat{\mathbf{w}}_{t+1})$
end for
```

Putting the above observations together, one can see now that, for any $\mathbf{w}$, there holds

$$p(1-p)\mathbb{E}[(1-\mathbf{w}^\top(x-x'))^2|y=1, y'=-1] = p(1-p) + \min_{a,b}\max_{\alpha} \mathbb{E}[F(\mathbf{w}, a, b, \alpha; z)].$$

This completes the proof. □

The problem (3) is a standard stochastic saddle point problem (see *e.g.* (Nemirovski et al., 2009)). It is easy to show that its objective function is convex with respect to $\mathbf{w}$, a, and b and concave with respect to α. We later refer to $\mathbf{w}$, a, and b as primal variables and α as a dual variable.

2.2. Proposed Algorithm and Interpretation

The algorithm proposed in (Ying et al., 2016) essentially performs stochastic gradient descent on the primal variables $\mathbf{w}$, a, and b and stochastic gradient ascent on the dual variable α. The critical observation in this paper is that, for fixed $\mathbf{w}$, the optima for a, b, and α in saddle formulation (3) has the exact formulations as given by (8) and (9).

This motivates us to conduct stochastic gradient descent only on $\mathbf{w}$, while a, b, and α are then updated using equations (8) and (9), rather than doing stochastic gradient updates. More specifically, upon receiving data z_t, we update $\mathbf{w}$ by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \partial_1 F(\mathbf{w}_t, a(\mathbf{w}_t), b(\mathbf{w}_t), \alpha(\mathbf{w}_t); z_t), \quad (10)$$

where $\partial_1 F$ denotes the gradient with respect to the first argument and the η_t's are the step sizes. To accommodate the possibly non-smooth penalty term $\Omega(\cdot)$, the next step is to perform a proximal mapping. Specifically, the proximal mapping associated with a convex function $\Omega : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\mathrm{prox}_{\eta_t \Omega}(u) = \arg\min\{\frac{1}{2}\|u - \mathbf{w}\|^2 + \eta_t \Omega(\mathbf{w})\}. \quad (11)$$

The pseudo-code of the proposed algorithm is outlined in Algorithm 1. This new online algorithm has per-iteration and storage cost of one datum. In the algorithm, it is assumed that the probability of class 1, *i.e.*, $p = \Pr(y = 1)$, and $\mathbb{E}(x|y = 1)$ and $\mathbb{E}(x|y = -1)$ are known. In practice, using a portion of the training data, one can efficiently estimate p by the proportion of samples of class 1, and the population means $\mathbb{E}(x|y = 1)$ and $\mathbb{E}(x|y = -1)$ by sample means.

Before we present the rigorous convergence rate of SPAM, let us briefly illustrate the intuition as to why it can be expected to achieve a faster rate of $\mathcal{O}(\frac{1}{t})$ in contrast to $\mathcal{O}(\frac{1}{\sqrt{t}})$ of SOLAM in (Ying et al., 2016). To see this, let us present a simple but critical lemma as follows. For this purpose, let $f(\mathbf{w}) = p(1-p)\mathbb{E}\big[(1-\mathbf{w}^\top(x-x'))^2\big|y=1, y'=-1\big]$ which is identical to $\min_{a,b}\max_\alpha \mathbb{E}[F(\mathbf{w}, a, b, \alpha; z)]$.

Lemma 1. *Let $\mathbf{w}_t$ be given by SPAM described in Algorithm 1. Then, we have that*

$$\partial f(\mathbf{w}_t) = \mathbb{E}_{z_t}[\partial_1 F(\mathbf{w}_t, a(\mathbf{w}_t), b(\mathbf{w}_t), \alpha(\mathbf{w}_t); z_t)],$$

where $\mathbb{E}_{z_t}[\cdot]$ denotes the expectation with respect to $z_t = (x_t, y_t)$.

Proof. Denote by $\partial_i F$ the partial derivative of F with respect to the ith argument. Applying the chain rule gives

$$\begin{aligned}
\partial_{\mathbf{w}} f(\mathbf{w}_t) &= \partial_{\mathbf{w}} \mathbb{E}_{z_t}[F(\mathbf{w}_t, a(\mathbf{w}_t), b(\mathbf{w}_t), \alpha(\mathbf{w}_t); z_t)] \\
&= \mathbb{E}_{z_t}\Big[\partial_{\mathbf{w}} F(\mathbf{w}_t, a(\mathbf{w}_t), b(\mathbf{w}_t), \alpha(\mathbf{w}_t); z_t)\Big] \\
&= \mathbb{E}_{z_t}\Big[\partial_1 F(\mathbf{w}_t, a(\mathbf{w}_t), b(\mathbf{w}_t), \alpha(\mathbf{w}_t); z_t)\Big] \\
&+ \mathbb{E}_{z_t}\Big[\partial_2 F(\mathbf{w}_t, a(\mathbf{w}_t), b(\mathbf{w}_t), \alpha(\mathbf{w}); z_t)\, \partial_{\mathbf{w}} a(\mathbf{w}_t)\Big] \\
&+ \mathbb{E}_{z_t}\Big[\partial_3 F(\mathbf{w}_t, a(\mathbf{w}_t), b(\mathbf{w}_t), \alpha(\mathbf{w}_t); z_t)\, \partial_{\mathbf{w}} b(\mathbf{w}_t)\Big] \\
&+ \mathbb{E}_{z_t}\Big[\partial_4 F(\mathbf{w}_t, a(\mathbf{w}_t), b(\mathbf{w}_t), \alpha(\mathbf{w}_t); z_t) \partial_{\mathbf{w}} \alpha(\mathbf{w}_t)\Big]. \quad (12)
\end{aligned}$$

The second inequality of interchanging differentiation and integration follows from the Leibniz's Integral rule since $F(\mathbf{w}_t, a(\mathbf{w}_t), b(\mathbf{w}_t), \alpha(\mathbf{w}); z_t)$ is quadratic and the input space $\mathcal{X}$ is a bounded domain. In the last equality, the fact that $\mathbf{w}_t$ only depends on $\{z_1, z_2, \ldots, z_{t-1}\}$ implies that $\mathbb{E}_{z_t}\Big[\partial_2 F(\mathbf{w}_t, a(\mathbf{w}_t), b(\mathbf{w}_t), \alpha(\mathbf{w}); z_t)\, \partial_{\mathbf{w}} a(\mathbf{w}_t)\Big] = \partial_2 \mathbb{E}_{z_t}\Big[F(\mathbf{w}_t, a(\mathbf{w}_t), b(\mathbf{w}_t), \alpha(\mathbf{w}); z_t)\Big][\mathbb{E}(x|y=1)]$.

Since $a(\mathbf{w}_t)$ is the minimizer of $\min_{a,b}\max_\alpha \mathbb{E}_{z_t}[F(\mathbf{w}_t, a, b, \alpha; z_t)]$, the first order optimality condition gives, for any b and α, that $\partial_2 \mathbb{E}_{z_t}\Big[F(\mathbf{w}_t, a(\mathbf{w}_t), b, \alpha; z_t)\Big] = 0$. Therefore we have that $\partial_2 \mathbb{E}_{z_t}\Big[F(\mathbf{w}_t, a(\mathbf{w}_t), b(\mathbf{w}_t), \alpha(\mathbf{w}_t); z_t)\Big] = 0$. Hence, $\mathbb{E}_{z_t}\Big[\partial_2 F(\mathbf{w}_t, a(\mathbf{w}_t), b(\mathbf{w}_t), \alpha(\mathbf{w}); z_t)\, \partial_{\mathbf{w}} a(\mathbf{w}_t)\Big] = 0$. Likewise, the third and fourth terms on the righthand side of (12) equal to zero. This completes the proof of the lemma. □

The above lemma implies, conditioned on $\{z_1, \ldots, z_{t-1}\}$, that $\partial_1 F(\mathbf{w}_t, a(\mathbf{w}_t), b(\mathbf{w}_t), \alpha(\mathbf{w}_t); z_t)$ is an unbiased estimator of the true gradient $\partial_{\mathbf{w}} f(\mathbf{w}_t)$. This strongly indicates that SPAM will have a fast convergence rate similar to SGD algorithms (Rakhlin et al., 2012a; Shalev-Shwartz, 2012) for a strongly convex objective function. We will leverage this intuition to prove the fast convergence rate in the next section.

More related work: We should point out that our proposed algorithm has similar spirit to the online forward-backward splitting (Duchi & Singer, 2009) and stochastic proximal gradient methods (Rosasco et al., 2014). However, there are two main differences between our proposed algorithm and their algorithms. Firstly, these algorithms focused on the accuracy performance where the objective function is a single summation/integral over individual samples. Secondly, the convergence proofs in (Duchi & Singer, 2009; Rosasco et al., 2014) critically depend on the boundedness assumptions: the iterates and the stochastic gradients are uniformly bounded or the conditional variance of the stochastic gradient is bounded by the square norm of the true gradient plus a constant, which may not be true and is difficult to verify in practice. Our proof techniques for the convergence of SPAM do not need these boundedness assumptions as shown in the next section. Lastly, the very recent work (Palaniappan & 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 least square loss. However, their saddle point formulation is different from (3) and the algorithm there needs to assume strong convexity on both the primal and dual variables. In addition, the algorithm has per-iteration complexity $O(n+d)$ where n is the total number of training samples and d is the dimension of the data.

3. Convergence Analysis

Before we present the convergence rate of SPAM, let us introduce some notations. Recall that $f(\mathbf{w}) = p(1-p)\mathbb{E}\big[(1 - \mathbf{w}^\top(x - x'))^2 \big| y = 1, y' = -1\big]$. Let $\mathbf{w}^*$ denote the optimal solution of formulation (2), i.e.,

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \{f(\mathbf{w}) + \Omega(\mathbf{w})\}.$$

Define

$$\mathbb{E}[\|G(\mathbf{w}^*; z) - \partial f(\mathbf{w}^*)\|^2] = \sigma_*^2, \tag{13}$$

where, for notional simplicity, we denote $G(\mathbf{w}; z) = \partial_1 F(\mathbf{w}, a(\mathbf{w}), b(\mathbf{w}), \alpha(\mathbf{w}); z)$. The convergence results are established based on the following two assumptions:

- (A1) Assume that $\Omega(\cdot)$ is β-strongly convex.
- (A2) There exists an $M > 0$ such that $\|\mathbf{x}\| \leq M$ for any $x \in \mathcal{X}$.

Furthermore, let $C_{\beta,M} := \frac{\beta}{128M^4}$, $\widetilde{C}_{\beta,M} = \frac{\beta}{(1+\frac{\beta^2}{128M^4})^2}$, and $\bar{C}_{\beta,M} = \widetilde{C}_{\beta,M} C_{\beta,M} = \frac{128M^4\beta^2}{(128M^4+\beta^2)^2}$. We use the conventional notation that for any $T \in \mathbb{N}$, $\mathbb{N}_T = \{1, \ldots, T\}$.

The proofs for Theorems 2 and 3 critically depend on the following lemma which clearly describes how $\|\mathbf{w}_t - \mathbf{w}^*\|$ evolves along time t.

Lemma 2. *Under the assumptions of (A1) and (A2), let $\{\mathbf{w}_t : t \in \mathbb{N}_{T+1}\}$ be generated by SPAM. Then, the following statements hold true.*

(i) For any $t \in \mathbb{N}$ there holds

$$\begin{aligned}&\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2] \\ &\leq \frac{1 + 128M^4\eta_t^2}{(1+\eta_t\beta)^2}\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2] + 2\sigma_*^2\eta_t^2.\end{aligned} \tag{14}$$

(ii) If, furthermore, $0 < \eta_t \leq C_{\beta,M} := \dfrac{\beta}{128M^4}$ for any $t \in \mathbb{N}_T$, then we have , for any $t \in \mathbb{N}_T$,

$$\begin{aligned}&\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2] \\ &\leq \big(1 - \widetilde{C}_{\beta,M}\,\eta_t\big)\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2] + 2\sigma_*^2\eta_t^2.\end{aligned} \tag{15}$$

Proof. Recall that $\mathbf{w}^*$ is the optimal solution of (2). One can directly derive from the first-order optimality condition using subgradients that, for any $\eta_t > 0$,

$$\mathbf{w}^* = \text{prox}_{\eta_t\Omega}(\mathbf{w}^* - \eta_t\partial f(\mathbf{w}^*)).$$

The above observation together with the definition of $\mathbf{w}_{t+1}$ in algorithm SPAM yields that

$$\begin{aligned}&\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 \\ &= \|\text{prox}_{\eta_t\Omega}(\hat{\mathbf{w}}_{t+1}) - \text{prox}_{\eta_t\Omega}(\mathbf{w}^* - \eta_t\partial f(\mathbf{w}^*)\|^2.\end{aligned} \tag{16}$$

Now since $\eta_t\Omega(\mathbf{w})$ is $\eta_t\beta$-strongly convex due to (A1), then by Proposition 23.11 in (Bauschke & Combettes, 2011), $\text{prox}_{\eta_t\Omega}(\cdot)$ is $(1+\eta_t\beta)$-cocoercive, i.e., for any $\mathbf{u}$ and $\mathbf{w}$, there holds $\langle \mathbf{u} - \mathbf{w}, \text{prox}_{\eta_t\Omega}(\mathbf{u}) - \text{prox}_{\eta_t\Omega}(\mathbf{w})\rangle \geq (1+\eta_t\beta)\|\text{prox}_{\eta_t\Omega}(\mathbf{u}) - \text{prox}_{\eta_t\Omega}(\mathbf{w})\|^2$. This, by Cauchy-Schwartz inequality, implies that

$$\|\text{prox}_{\eta_t\Omega}(\mathbf{u}) - \text{prox}_{\eta_t\Omega}(\mathbf{w})\| \leq \frac{1}{1+\eta_t\beta}\|\mathbf{u} - \mathbf{w}\|.$$

Putting this back into (16), we get

$$\begin{aligned}&\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 \\ &= \|\text{prox}_{\eta_t\Omega}(\hat{\mathbf{w}}_{t+1}) - \text{prox}_{\eta_t\Omega}(\mathbf{w}^* - \eta_t\partial f(\mathbf{w}^*)\|^2 \\ &\leq \frac{1}{(1+\eta_t\beta)^2}\|\hat{\mathbf{w}}_{t+1} - (\mathbf{w}^* - \eta_t\partial f(\mathbf{w}^*))\|^2 \\ &= \frac{1}{(1+\eta_t\beta)^2}\|(\mathbf{w}_t - \mathbf{w}^*) - \eta_t(G(\mathbf{w}_t, z_t) - \partial f(\mathbf{w}^*))\|^2,\end{aligned}$$

where in the last equality we recall the notation that $G(\mathbf{w}_t; z_t) = \partial_1 F(\mathbf{w}_t, a(\mathbf{w}_t), b(\mathbf{w}_t), \alpha(\mathbf{w}_t); z_t)$. Now taking the expectation of both sides of the above inequality, and expanding out the right hand side, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2] \leq \frac{1}{(1+\eta_t\beta)^2}\Big(&\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2] \\ &- 2\eta_t \mathbb{E}[\langle \mathbf{w}_t - \mathbf{w}^*, G(\mathbf{w}_t; z_t) - \partial f(\mathbf{w}^*)\rangle] \\ &+ \eta_t^2 \mathbb{E}[\|G(\mathbf{w}_t; z_t) - \partial f(\mathbf{w}^*)\|^2\Big). \end{aligned} \tag{17}$$

We first bound the middle term of the righthand side of (17). By Lemma 1, we know that

$$\begin{aligned}&\mathbb{E}[\langle \mathbf{w}_t - \mathbf{w}^*, G(\mathbf{w}_t; z_t) - \partial f(\mathbf{w}^*)\rangle] \\ &= \mathbb{E}[\langle \mathbf{w}_t - \mathbf{w}^*, \mathbb{E}_{z_t}[G(\mathbf{w}_t; z_t)] - \partial f(\mathbf{w}^*)\rangle] \\ &= \mathbb{E}[\langle \mathbf{w}_t - \mathbf{w}^*, \partial f(\mathbf{w}_t) - \partial f(\mathbf{w}^*)\rangle] \geq 0, \end{aligned} \tag{18}$$

where the last inequality follows from the convexity of f.

For the last term on the righthand side of (17), we proceed as follows: $\mathbb{E}[\|G(\mathbf{w}_t; z_t) - \partial f(\mathbf{w}^*)\|^2] \leq 2\mathbb{E}[\|G(\mathbf{w}_t; z_t) - G(\mathbf{w}^*; z_t)\|^2] + 2\mathbb{E}[\|G(\mathbf{w}^*; z_t) - \partial f(\mathbf{w}^*)\|^2]$. Note that $G(\mathbf{w}_t; z_t)$ is a linear function of $\mathbf{w}_t$. So by the assumption that $\|\mathbf{x}_t\| \leq M$, it is easy to see that

$$\begin{aligned}\|G(\mathbf{w}_t; z_t) - G(\mathbf{w}^*; z_t)\| & \\ \leq 4M^2(1-p)&\|\mathbf{w}_t - \mathbf{w}^*\|\mathbb{I}_{[y_t=1]} \\ + 4M^2 p&\|\mathbf{w}_t - \mathbf{w}^*\|\mathbb{I}_{[y_t=-1]} \\ + 4M^2 |p& - \mathbb{I}_{[y_t=1]}|\|\mathbf{w}_t - \mathbf{w}^*\| \\ \leq 8M^2&\|\mathbf{w}_t - \mathbf{w}^*\|. \end{aligned} \tag{19}$$

Furthermore, from (13), we have $\mathbb{E}[\|G(\mathbf{w}^*; z_t) - \partial f(\mathbf{w}^*)\|^2] = \mathbb{E}_{z_t}[\|G(\mathbf{w}^*; z_t) - \partial f(\mathbf{w}^*)\|^2] = \sigma_*^2$. Hence,

$$\begin{aligned}\mathbb{E}[\|G(\mathbf{w}_t; z_t) - &\partial f(\mathbf{w}^*)\|^2] \\ &\leq 2(8M^2)^2 \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2] + 2\sigma_*^2. \end{aligned} \tag{20}$$

Putting together (17), (18) and (20), we get

$$\begin{aligned}&\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2] \\ &\leq \frac{1}{(1+\eta_t\beta)^2}\Big(\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2] \\ &+ 2(8M^2)^2\eta_t^2\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2] + 2\sigma_*^2\eta_t^2\Big) \\ &\leq \frac{1 + 128M^4\eta_t^2}{(1+\eta_t\beta)^2}\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2] + 2\sigma_*^2\eta_t^2. \end{aligned} \tag{21}$$

This finishes part (i) of the lemma.

For the second part of the lemma, notice that $\eta_t \leq C_{\beta,M} := \dfrac{\beta}{128M^4}$. The coefficient in (21) can be rewritten as follows:

$$\begin{aligned}\frac{1 + 128M^4\eta_t^2}{(1+\eta_t\beta)^2} &= 1 - \Big(1 - \frac{1 + 128M^4\eta_t^2}{(1+\eta_t\beta)^2}\Big) \\ &= 1 - \frac{[2\beta + \beta^2\eta_t - 128M^4\eta_t]\eta_t}{(1+\eta_t\beta)^2}. \end{aligned} \tag{22}$$

Applying the assumption that $\eta_t \leq \dfrac{\beta}{128M^4}$ gives that

$$\frac{[2\beta + \beta^2\eta_t - 128M^4\eta_t]}{(1+\eta_t\beta)^2}\eta_t \geq \frac{\beta}{\big(1 + \frac{\beta^2}{128M^4}\big)^2}\eta_t. \tag{23}$$

In addition, notice that $\frac{\beta}{128M^4} \leq \frac{\left(1+\frac{\beta^2}{128M^4}\right)^2}{\beta}$. This implies the assumption $\eta_t \leq \frac{\beta}{128M^4}$ guarantees that $1 - \frac{\beta}{\left(1+\frac{\beta^2}{128M^4}\right)^2}\eta_t \geq 0$. Combining together (22) with (23) yields the desired result in part (ii). This completes the proof of the lemma. □

The following lemma is from (Smale & Yao, 2006) and will be used to prove Theorems 2 and 3.

Lemma 3. *For any* $0 < \nu \leq 1$, $0 < \alpha < 1$, $t < T$, *and* $0 < \theta \leq 1$, *the following estimations hold true.*

(i) $\sum_{j=t+1}^T j^{-\alpha} \geq \frac{1}{1-\alpha}[(T+1)^{1-\alpha} - (t+1)^{1-\alpha}]$,

(ii) $\sum_{t=1}^{T-1} \frac{1}{t^{2\alpha}} \exp\Big\{ - \nu\sum_{j=t+1}^T j^{-\alpha}\Big\} \leq \frac{18}{\nu T^{\alpha}} + \frac{9T^{1-\alpha}}{(1-\alpha)2^{1-\alpha}} \exp\{-\frac{\nu(1-2^{\alpha-1})}{1-\alpha}(T+1)^{1-\alpha}\}$,

(iii) $e^{-cx} \leq \left(\frac{b}{ce}\right)^b x^{-b}$ *for* $x > 0, c > 0$ *and* $b > 0$.

We now present the convergence analysis.

Theorem 2. *Under the assumptions (A1), (A2), and choosing step sizes with some* $\theta \in (0,1)$ *in the form of* $\{\eta_t = \frac{C_{\beta,M}}{t^\theta} : t \in \mathbb{N}\}$, *the algorithm SPAM achieves the following:*

$$\begin{aligned}&\mathbb{E}[\|\mathbf{w}_{T+1} - \mathbf{w}^*\|^2] \\ &\leq \Big[\exp\Big(\frac{\bar{C}_{\beta,M}}{1-\theta}\Big)\Big(\frac{\theta}{\bar{C}_{\beta,M}e}\Big)^{\frac{\theta}{1-\theta}}\mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}^*\|^2] \\ &+ 2\sigma_*^2 C_{\beta,M}^2\Big(\frac{9}{(1-\theta)2^{1-\theta}}\Big(\frac{1}{\bar{C}_{\beta,M}(1-2^{\theta-1})e}\Big)^{\frac{1}{1-\theta}} \\ &+ \frac{18}{\bar{C}_{\beta,M}} + 1\Big)\Big]T^{-\theta}. \end{aligned}$$

Proof. Denote $r_t = \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2]$. The choice of the step sizes $\eta_t = \frac{C_{\beta,M}}{t^\theta}$ satisfies the condition in Lemma 2, i.e. $\eta_t \leq C_{\beta,M}$. Recall that $C_{\beta,M} = \frac{\beta}{128M^4}$, $\widetilde{C}_{\beta,M} = \frac{\beta}{(1+\frac{\beta^2}{128M^4})^2}$, and $\bar{C}_{\beta,M} = \widetilde{C}_{\beta,M}C_{\beta,M}$ which guarantees that $1 - \widetilde{C}_{\beta,M}\eta_t \geq 1 - \widetilde{C}_{\beta,M}C_{\beta,M} = 1 - \bar{C}_{\beta,M} \geq 0$ for any $t \in \mathbb{N}_T$. Then, it is easy to see from (15) that, after T iterations, there holds

$$\begin{aligned}r_{T+1} &\leq r_1 \prod_{k=1}^T \Big(1 - \widetilde{C}_{\beta,M}\eta_k\Big) \\ &+ 2\sigma_*^2 \sum_{k=1}^{T-1} \prod_{i=k+1}^T \Big(1 - \widetilde{C}_{\beta,M}\,\eta_i\Big)\eta_k^2 + 2\sigma_*^2\eta_T^2. \end{aligned} \tag{24}$$

The first term on the right hand side can be bounded using the fact that $1 - x \le \exp(-x)$ for all $x \in \mathbb{R}$, giving that

$$\begin{aligned} r_1 \prod_{k=1}^{T} \Big(1 - \widetilde{C}_{\beta,M}\eta_k\Big) &= r_1 \prod_{k=1}^{T} \Big(1 - \widetilde{C}_{\beta,M} C_{\beta,M}/k^{\theta}\Big) \\ &\le r_1 \exp\Big(- \bar{C}_{\beta,M} \sum_{k=1}^{T} \frac{1}{k^{\theta}}\Big), \end{aligned} \tag{25}$$

where $\bar{C}_{\beta,M} = \widetilde{C}_{\beta,M} C_{\beta,M} = \frac{128M^4\beta^2}{(128M^4+\beta^2)^2}$. Applying part (i) in Lemma 3 gives that

$$\begin{aligned} & r_1 \exp\Big(- \bar{C}_{\beta,M} \sum_{k=1}^{T} \frac{1}{k^{\theta}}\Big) \\ &\le r_1 \exp\Big(\frac{\bar{C}_{\beta,M}}{1-\theta} \big[1 - (T+1)^{1-\theta}\big]\Big) \\ &= r_1 \exp\Big(\frac{\bar{C}_{\beta,M}}{1-\theta}\Big) \exp\Big(- \frac{\bar{C}_{\beta,M}}{1-\theta}(T+1)^{1-\theta}\Big). \end{aligned}$$

Applying part (iii) in Lemma 3 with $b = \frac{\theta}{1-\theta}$, $x = (T+1)^{1-\theta}$and $c = \frac{\bar{C}_{\beta,M}}{1-\theta}$ yields that

$$\exp\Big(-\frac{\bar{C}_{\beta,M}}{1-\theta}(T+1)^{1-\theta}\Big) \le \Big(\frac{\theta}{\bar{C}_{\beta,M} e}\Big)^{\frac{\theta}{1-\theta}} (T+1)^{-\theta}.$$

Putting the above two inequalities back into (25), we have

$$\begin{aligned} & r_1 \prod_{k=1}^{T} \Big(1 - \widetilde{C}_{\beta,M}\eta_k\Big) \\ &\qquad \le r_1 \exp\Big(\frac{\bar{C}_{\beta,M}}{1-\theta}\Big)\Big(\frac{\theta}{\bar{C}_{\beta,M} e}\Big)^{\frac{\theta}{1-\theta}} T^{-\theta}. \end{aligned} \tag{26}$$

To bound the second term on the righthand side of (24), we proceed again as in the first term:

$$\begin{aligned} & \sum_{k=1}^{T-1} \prod_{i=k+1}^{T} \Big(1 - \widetilde{C}_{\beta,M}\eta_i\Big) \eta_k^2 \\ &= C_{\beta,M}^2 \sum_{k=1}^{T-1} \frac{1}{k^{2\theta}} \prod_{i=k+1}^{T} \Big(1 - \frac{\bar{C}_{\beta,M}}{i^{\theta}}\Big) \\ &\le C_{\beta,M}^2 \sum_{k=1}^{T-1} \frac{1}{k^{2\theta}} \exp\Big(- \bar{C}_{\beta,M} \sum_{i=k+1}^{T} \frac{1}{i^{\theta}}\Big). \end{aligned} \tag{27}$$

Applying Lemma 3 (ii) with $\nu = \bar{C}_{\beta,M}$ and $\alpha = \theta$ gives that the above is bounded by

$$\begin{aligned} & \sum_{k=1}^{T-1} \frac{1}{k^{2\theta}} \exp\Big(- \bar{C}_{\beta,M} \sum_{i=k+1}^{T} \frac{1}{i^{\theta}}\Big) \\ &\le \frac{9T^{1-\theta}}{(1-\theta)(2^{1-\theta})} \exp\Big(- \frac{\bar{C}_{\beta,M}(1-2^{\theta-1})}{1-\theta}(T+1)^{1-\theta}\Big) \\ &+ \frac{18}{\bar{C}_{\beta,M} T^{\theta}}. \end{aligned} \tag{28}$$

And again, applying Lemma 3 (iii) with $b = \frac{1}{1-\theta}, x = (T+1)^{1-\theta}$ and $c = \frac{\bar{C}_{\beta,M}(1-2^{\theta-1})}{1-\theta}$ to (28) gives that

$$\begin{aligned} & \exp\Big(- \frac{\bar{C}_{\beta,M}(1-2^{\theta-1})}{1-\theta}(T+1)^{1-\theta}\Big) \\ &\le \Big(\frac{1}{\bar{C}_{\beta,M}(1-2^{\theta-1})e}\Big)^{\frac{1}{1-\theta}} (T+1)^{-1}. \end{aligned} \tag{29}$$

Putting (28) and (29) back into (27), we have

$$\begin{aligned} & 2\sigma_*^2 \sum_{k=1}^{T-1} \prod_{i=k+1}^{T} \Big(1 - \widetilde{C}_{\beta,M}\,\eta_i\Big) \eta_k^2 \\ &\le 2\sigma_*^2 C_{\beta,M}^2 \Big[\frac{9}{(1-\theta)2^{1-\theta}} \Big(\frac{1}{\bar{C}_{\beta,M}(1-2^{\theta-1})e}\Big)^{\frac{1}{1-\theta}} \\ &+ \frac{18}{\bar{C}_{\beta,M}}\Big] T^{-\theta}. \end{aligned} \tag{30}$$

The last term on the righthand side of (24) is straightforward: $2\sigma_*^2\eta_T^2 \le 2\sigma_*^2 C_{\beta,M}^2 T^{-\theta}$. This, in combination of (26) and (30), yields the desired result. □

This theorem indicates the last output of SPAM achieves the convergence rate of $\mathcal{O}(T^{-\theta})$ with polynomial decaying step sizes in the form of $\eta_t = \mathcal{O}(t^{-\theta})$ for $\theta \in (0,1)$. For $\theta = 1$, we can obtain the following result.

Theorem 3. *Under the assumptions of (A1), (A2), and choosing step sizes $\{\eta_t = [\widetilde{C}_{\beta,M}(t+1)]^{-1} : t \in \mathbb{N}\}$, the algorithm SPAM achieves the following:*

$$\begin{aligned} & \mathbb{E}[\|\mathbf{w}_{T+1} - \mathbf{w}^*\|^2] \\ &\qquad \le \big(t_0 \mathbb{E}[\|\mathbf{w}_{t_0} - \mathbf{w}^*\|^2]\big)\frac{1}{T} + \frac{4\sigma_*^2}{\widetilde{C}_{\beta,M}^2}\frac{\log T}{T}. \end{aligned}$$

where $t_0 = \max\Big(2, \Big\lceil 1 + \frac{(128M^4+\beta^2)^2}{128M^4\beta^2}\Big\rceil\Big)$.

Proof. The condition that $t \ge t_0$ guarantees the assumption in part (ii) of Lemma 2 that $\eta_t = \Big[\widetilde{C}_{\beta,M}(t+1)\Big]^{-1} \le C_{\beta,M}$ is satisfied. Now by letting $r_t = \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2]$ we have

$$r_{t+1} \le \Big(1 - \widetilde{C}_{\beta,M}\,\eta_t\Big) r_t + 2\sigma_*^2\eta_t^2. \tag{31}$$

Then, we have

$$\begin{aligned} r_{T+1} &\le r_{t_0} \prod_{k=t_0}^{T} \Big(1 - \widetilde{C}_{\beta,M}\eta_k\Big) + 2\sigma_*^2\eta_T^2 \\ &+ 2\sigma_*^2 \sum_{k=t_0}^{T-1} \prod_{i=k+1}^{T} \Big(1 - \widetilde{C}_{\beta,M}\eta_i\Big) \eta_k^2. \end{aligned} \tag{32}$$

The first term on the right hand side of the above inequality can be estimated as follows: $r_{t_0} \prod_{k=t_0}^{T}(1 - \widetilde{C}_{\beta,M}\eta_k) =$

$r_{t_0} \prod_{k=t_0}^{T} \frac{k}{k+1} = \frac{t_0 r_{t_0}}{T+1} \leq \frac{t_0 r_{t_0}}{T}$. For the second term on the righthand side of (32), there holds $2\sigma_*^2 \eta_T = \frac{2\sigma_*^2}{\widetilde{C}_{\beta,M}^2 (T+1)^2} \leq \frac{2\sigma_*^2}{\widetilde{C}_{\beta,M}^2 T}$. To bound the third term on the righthand side of (32), we can do the following $\sum_{k=t_0}^{T-1} \prod_{i=k+1}^{T} (1 - \widetilde{C}_{\beta,M} \eta_i) \eta_k^2 = \widetilde{C}_{\beta,M}^{-2} \sum_{k=t_0}^{T-1} \prod_{i=k+1}^{T-1} \left(1 - \frac{1}{i+1}\right) \frac{1}{(k+1)^2} = \widetilde{C}_{\beta,M}^{-2} \frac{1}{T} \sum_{k=t_0}^{T-1} \frac{1}{k+1} \leq \widetilde{C}_{\beta,M}^{-2} \frac{\log(T-1) - \log t_0}{T} \leq \widetilde{C}_{\beta,M}^{-2} \frac{\log T}{T}$. Putting all the above estimations together yields the desired result. □

The convergence of SPAM proved in the above theorem shows that it can achieve $\mathcal{O}(\frac{\log T}{T})$. The convergence rate of $\mathcal{O}(\frac{1}{T})$ could be obtained using averaging schemes developed by (Lacoste-Julien et al., 2012; Rakhlin et al., 2012b; Shamir & Zhang, 2013).

The term $\mathbb{E}[\|\mathbf{w}_{t_0} - \mathbf{w}^*\|^2]$ can indeed be estimated as follows if $\eta_t = [\widetilde{C}_{\beta,M}(t+1)]^{-1}$ for $t \in \mathbb{N}$. From part (i) of Lemma 2, we have, for any $t \in \mathbb{N}$, $\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2] \leq \frac{1+128M^4\eta_t^2}{(1+\eta_t\beta)^2} \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2] + 2\sigma_*^2\eta_t^2 \leq (1 + 128M^4\eta_t^2)\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2] + 2\sigma_*^2\eta_t^2$. Therefore, $\mathbb{E}[\|\mathbf{w}_{t_0} - \mathbf{w}^*\|^2] \leq \prod_{k=1}^{t_0-1}(1 + 128M^4\eta_k^2) + 2\sigma_*^2 \sum_{k=1}^{t_0-1} \prod_{j=k}^{t_0-1} (1 + 128M^4\eta_j^2)\eta_k^2 \leq \left(\prod_{k=1}^{t_0-1}(1+128M^4\eta_k^2)\right)\left(1+2\sigma_*^2 \sum_{k=1}^{t_0-1} \eta_k^2\right)$. Notice that $\prod_{k=1}^{t_0-1}(1+128M^4\eta_k^2) \leq \exp\left(\frac{128M^4}{\widetilde{C}_{\beta,M}^2} \sum_{k=1}^{t_0-1}(k+1)^{-2}\right) \leq \exp\left(\frac{128M^4}{\widetilde{C}_{\beta,M}^2}\right)$. and $2\sigma_*^2 \sum_{k=1}^{t_0-1} \eta_k^2 = \frac{2\sigma_*^2}{\widetilde{C}_{\beta,M}^2} \sum_{k=1}^{t_0-1}(k+1)^{-2} \leq \frac{2\sigma_*^2}{\widetilde{C}_{\beta,M}^2}$. Hence, one can have the following bound depending on β and M:

$$\mathbb{E}[\|\mathbf{w}_{t_0} - \mathbf{w}^*\|^2] \leq \frac{2\sigma_*^2}{\widetilde{C}_{\beta,M}^2} + \exp\Big(\frac{128M^4}{\widetilde{C}_{\beta,M}^2}\Big).$$

4. Experiments

In this section, we report the experimental evaluation of SPAM by comparing it against existing algorithms for AUC optimization.

In particular, we use SPAM-L^2 to denote SPAM with the Frobenius norm, i.e., $\Omega(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{w}\|^2$. The solution to the proximal step using the Frobenius norm is very straight forward. The other version, SPAM-NET, uses the elastic net norm (Zou & Hastie, 2005), i.e., $\Omega(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{w}\|^2 + \beta_1\|\mathbf{w}\|_1$. The proximal step can be written as

$$\arg\min_{\mathbf{w}} \Big\{ \frac{1}{2} \left\| \mathbf{w} - \frac{\hat{\mathbf{w}}_{t+1}}{\eta_t\beta + 1} \right\|^2 + \frac{\eta_t\beta_1}{\eta_t\beta + 1} \|\mathbf{w}\|_1 \Big\},$$

for which the optimal solution is the soft-thresholding operator (e.g. Parikh et al. (2014)).

Table 1. Basic information about the datasets.

DATA	NAME	# INSTANCES	# FEATURES
1	DIABETES	768	8
2	FOURCLASS	862	8
3	GERMAN	1000	24
4	SPLICE	3175	60
5	USPS	9,298	256
6	A9A	32,561	123
7	MNIST	60,000	780
8	ACOUSTIC	78,823	50
9	IJCNN1	141,691	22
10	COVTYPE	581,012	54
11	SECTOR	9,619	55,197
12	NEWS20	15,935	62,061

We compare our algorithms with both batch and online AUC optimization algorithms. To ensure a fair comparison with (Ying et al., 2016), the algorithm SOLAM was modified to include the Frobenius-norm regularization term instead of the original bounded restriction on the norm of $\|\mathbf{w}\|$. We also compare our algorithm against the one-pass AUC optimization algorithm (Gao et al., 2013) with the least square loss and the OAMseq and OAMgra algorithms (Zhao et al., 2011) with hinge loss. Lastly, we include the B-LS-SVM algorithm (Joachims, 2006), a batch learning algorithm for AUC maximization with least square loss.

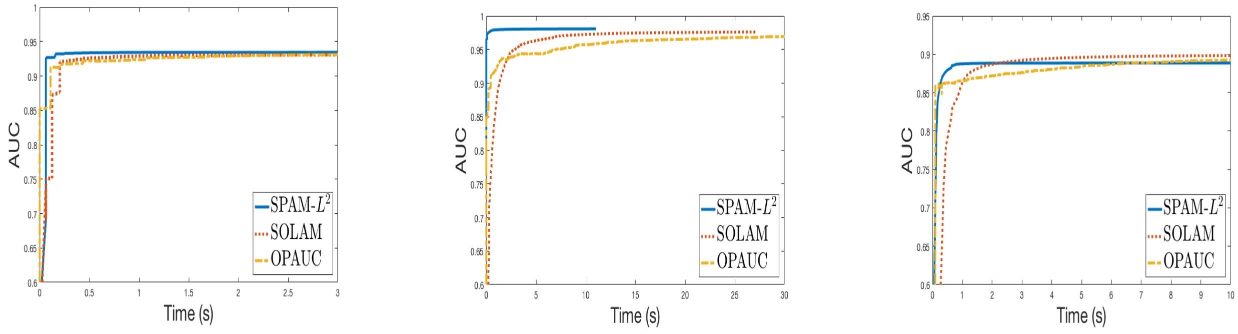
Table 1 summarizes the details of each of the data sets we used for comparison. All of these datasets are available to download from the LIBSVM (Chang & Lin, 2011) and UCI machine learning repository (Frank & Asuncion, 2010). It is worthy of noting that some of the datasets are multi-class. We converted them to binary data by randomly partitioning the data into two groups, where each group includes the same number of classes.

We used 80% of the data for training and the remaining 20% for testing. The results are based on 20 runs for each dataset for which we used to calculate the average AUC score and standard deviation. To determine the proper parameters for each dataset, we conduct 5-fold cross validation on the training sets to determine the parameter $\beta \in 10^{[-5:5]}$ for SPAM-L^2 and $\beta_1 \in 10^{[-5:5]}$ for SPAM-NET. All experiments were conducted with MATLAB and the MATLAB codes for the compared methods were obtained from the authors.

Classification performance on all of the data sets is summarized in Table 2. SPAM-L^2 and SPAM-NET both achieve a similar performance as the other state of the art AUC maximization algorithms in both the online and batch settings. This validates the algorithms we proposed in this paper. The data set `sector` shows the advantage of using elastic net. Next, we compared the CPU running time of SPAM-

Table 2. Comparison of the testing AUC values (mean±std.). To accelerate the experiments, the values for OPAUC, OAMseq, OAMgra, and B-LS-SVM were taken from (Gao et al., 2013).

DATA	SPAM-L^2	SPAM-NET	SOLAM	OPAUC	OAM$_{\text{SEQ}}$	OAM$_{\text{GRA}}$	B-LS-SVM
1	.8272±.0277	.8085±.0431	.8128±.0304	.8309±.0350	.8264±.0367	.8262±.0338	.8325±.0329
2	.8210±.0203	.8211±.0205	.8213±.0209	.8310±.0251	.8306±.0247	.8295±.0251	.8309±.0309
3	.7942±.0388	.7937±.0386	.7778±.0373	.7978±.0347	.7747±.0411	.7723±.0358	.7994±.0343
4	.9263±.0091	.9267±.0090	.9246±.0087	.9232±.0099	.8594±.0194	.8864±.0166	.9245±.0092
5	.9868±.0032	.9855±.0029	.9822±.0036	.9620±.0040	.9310±.0159	.9348±.0122	.9634±.0045
6	.8998±.0046	.8980±.0047	.8966±.0043	.9002±.0047	.8420±.0174	.8571±.0173	.8982±.0028
7	.9254±.0025	.9132±.0026	.9118±.0029	.9242±.0021	.8615±.0087	.8643±.0112	.9336±.0025
8	.8120±.0030	.8109±.0028	.8099±.0036	.8192±.0032	.7113±.0590	.7711±.0217	.8210±.0033
9	.9174±.0024	.9155±.0024	.9129±.0030	.9269±.0021	.9209±.0079	.9100±.0092	.9320±.0037
10	.9504±.0011	.9508±.0011	.9503±.0012	.8244±.0014	.7361±.0317	.7403±.0289	.8222±.0014
11	.8768±.0126	.9077±.0104	.8767±.0129	.9292±.0081	.9163±.0087	.9043±.0100	-
12	.8708±.0069	.8704± .0070	.8712±.0073	.8871±.0083	.8543±.0099	.8346±.0094	-



(a) splice

(b) usps

(c) a9a

Figure 1. AUC vs. CPU running time curves of SPAM-L^2 against SOLAM (Ying et al., 2016) and OPAUC (Gao et al., 2013).

L^2 versus SOLAM and the OPAUC algorithm. We did not compare the running time of SPAM against OAM (Zhao et al., 2011) since it used hinge loss. It was observed that the running time is inferior to OPAUC as shown in (Gao et al., 2013) and to SOLAM (Ying et al., 2016).

The main advantage of SPAM is the running efficiency. As we pointed out in the introduction, it has a faster convergence rate of $\mathcal{O}(\frac{1}{t})$ than SOLAM's convergence rate of $\mathcal{O}(\frac{1}{\sqrt{t}})$, and its per-iteration running time and space complexity are linear in data dimension. The running time performance against OPAUC and SOLAM is depicted in Figure 1 on `splice`, `usps` and `a9a` datasets. Theses results show that SPAM demonstrates a competitive performance while achieving a faster rate of performance.

5. Conclusion

In this paper, we proposed a novel stochastic proximal algorithm (SPAM) for AUC maximization with general penalty terms. We showed that the algorithm can achieve a convergence rate of $\mathcal{O}(1/T)$ up to a logarithmic term for strongly convex objective functions while the space and per-iteration complexity are of one datum.

There are several directions for future work. Firstly, it would be very interesting to extend the ideas of this paper to design stochastic variance reduction algorithms (Johnson & Zhang, 2013) and stochastic primal-dual algorithms (Zhang & Xiao, 2017) for AUC maximization, which can achieve the linear convergence rate. Secondly, it remains unclear to us whether SPAM can achieve convergence rate $\mathcal{O}(1/T)$ without strong convexity (e.g. SPAM-L^2 with $\beta = 0$). One possible approach is to adapt the proof techniques in (Bach & Moulines, 2013; Yang & Lin, 2015) to the setting of AUC maximization.

Acknowledgement

Yiming Ying is supported by the Simons Foundation (#422504) and the 2016-2017 Presidential Innovation Fund for Research and Scholarship (PIFRS) program from SUNY Albany. Siwei Lyu is supported by the National Science Foundation (NSF, Grant IIS1537257).

References

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

Bauschke, H H and Combettes, P L. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

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

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

Chang, C.-C. and Lin, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

Chen, Y, Lan, G, and Ouyang, Y. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.

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

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

Duchi, J., Hazan., E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Elkan, C. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, 2001.

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

Frank, Andrew and Asuncion, Arthur. Uci machine learning repository [http://archive. ics. uci. edu/ml]. irvine, ca: University of california. *School of information and computer science*, 213, 2010.

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

Gao, Wei and Zhou, Zhi-Hua. On the consistency of auc pairwise optimization. In *IJCAI*, pp. 939–945, 2015.

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

Hazan, E. and Kale, S. Projection-free online learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012.

Herschtal, A and Raskutti, B. Optimising area under the roc curve using gradient descent. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 49. ACM, 2004.

Joachims, Thorsten. A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on Machine learning*, pp. 377–384. ACM, 2005.

Joachims, Thorsten. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pp. 217–226, New York, NY, USA, 2006. ACM. ISBN 1-59593-339-5. doi: 10.1145/1150402.1150429. URL http://doi.acm.org/10.1145/1150402.1150429.

Johnson, Rie and Zhang, Tong. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pp. 315–323, 2013.

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

Kotlowski, W, Dembczynski, K J, and Huellermeier, E. Bipartite ranking through minimization of univariate loss. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 1113–1120, 2011.

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

Moulines, Eric and Bach, Francis R. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pp. 451–459, 2011.

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

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

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

Parikh, Neal, Boyd, Stephen, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 449–456, 2012a.

Rakhlin, A, Shamir, O, and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 449–456, 2012b.

Rakotomamonjy, Alain. Optimizing auc with support vector machine. In *European Conference on Artificial Intelligence Workshop on ROC Curve and AI*, 2004.

Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Rosasco, Lorenzo, Villa, Silvia, and Vũ, Bang Công. Convergence of stochastic proximal gradient algorithm. *arXiv preprint arXiv:1403.5074*, 2014.

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

Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

Shamir, O and Zhang, T. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pp. 71–79, 2013.

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

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

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

Yang, Tianbao and Lin, Qihang. Rsg: Beating subgradient method without smoothness and strong convexity. *arXiv preprint arXiv:1512.03107*, 2015.

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

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

Zhang, Yuchen and Xiao, Lin. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *The Journal of Machine Learning Research*, 18(1): 2939–2980, 2017.

Zhao, P., Jin, R., Yang, T., and Hoi, S. C. Online auc maximization. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011.

Zou, H and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.