Variance Reduced Stochastic Proximal Algorithm for AUC Maximization

Soham Dan* and Dushyant Sahoo*

University of Pennsylvania

Abstract

- Stochastic Gradient Descent (SGD) has been widely studied with classification accuracy as a performance measure.
- These algorithms are not applicable when non-decomposable pairwise performance measures are used, such as Area under the ROC curve (AUC).
- We propose a Variance Reduced Stochastic Proximal algorithm for AUC Maximization (VRSPAM) which converges faster than existing methods.

Introduction

- Class imbalance poses a challenge in several domains for instance, medical diagnosis of rare diseases. [1]
- AUC is commonly used to evaluate the performance of a binary classifier in this setting. AUC measures the ability of a family of classifiers to correctly rank an example from the positive class with respect to a randomly selected example from the negative class.
- In the online setting, AUC metric does not decomposes over individual instances, unlike classification accuracy.
- [2] reformulated the pairwise squared loss surrogate of AUC and gave an algorithm with a convergence rate of $\mathcal{O}\left(\frac{\log t}{t}\right)$, under strong convexity.
- This rate is sub-optimal to the linear rate SGD achieves with classification accuracy as a performance measure. The slow convergence is caused by the high variance of the gradient in each iteration.
- We present VRSPAM which extends previous work [2, 4] for surrogate-AUC maximization by using the Proximal SVRG [3] algorithm and achieves linear convergence rate.

AUC Formulation

- $\text{AUC}(\mathbf{w}) = \mathbb{E}[\mathbb{I}_{\mathbf{w}^T(x-x')\geq 0}|y=1, y'=-1]$
- We consider the below objectve function

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

where $f(\mathbf{w}) = p(1-p)\,\mathbb{E}[(1-\mathbf{w}^T(x-x'))^2|y=1, y'=-1]$ and Ω a convex regularizer (where $p = Pr(y=+1)$)

- The above minimization problem can be reformulated such that stochastic gradient descent can be performed to find the optimum value. Below is an equivalent formulation from Theorem 1 in [2]-

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

where the expectation is with respect to $z = (x, y)$ and

$$F(\mathbf{w}, a, b, \zeta; z) = (1-p)(\mathbf{w}^T x - a)^2\mathbb{I}_{[y=1]} + p(\mathbf{w}^T x - b)^2\mathbb{I}_{[y=-1]} + 2(1+\zeta)\mathbf{w}^T x(p\mathbb{I}_{[y=-1]} - (1-p)\mathbb{I}_{[y=1]}) - p(1-p)\zeta^2$$

[2] shows that the optimal choices for a, b, ζ satisfy

$$a(\mathbf{w}) = \mathbf{w}^T\,\mathbb{E}[x|y=1]$$
$$b(\mathbf{w}) = \mathbf{w}^T\,\mathbb{E}[x|y=-1]$$
$$\zeta(\mathbf{w}) = \mathbf{w}^T(\mathbb{E}[x'|y'=-1] - \mathbb{E}[x|y=1])$$

Algorithm

Let-

- $G(\mathbf{w}; z) = \partial_{\mathbf{w}} F(\mathbf{w}, a(\mathbf{w}), b(\mathbf{w}), \zeta(\mathbf{w}); z)$
- $\tilde{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^n G(\tilde{\mathbf{w}}, z_i)$
- $\mathbf{v}_t = G(\mathbf{w}_t, z_{i_{t-1}}) - G(\tilde{\mathbf{w}}, z_{i_{t-1}}) + \tilde{\boldsymbol{\mu}}$

Algorithm 1 Proximal SVRG for AUC maximization

INPUT Constant step size η and update frequency m

INITIALIZE $\tilde{\mathrm{w}}_0$

for $s = 1, 2, \ldots$ **do**
  $\tilde{\mathbf{w}} = \tilde{\mathbf{w}}_{s-1}$
  $\tilde{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^n G(\tilde{\mathbf{w}}, z_i)$
  $\mathbf{w}_0 = \tilde{\mathbf{w}}$
  for $t = 1, 2, \ldots, m$ **do**
    Randomly pick $i_t \in \{1, .., n\}$ and update weight
    $\hat{\mathbf{w}}_t = \mathbf{w}_{t-1} - \eta(G(\mathbf{w}_{t-1}, z_{i_t}) - G(\tilde{\mathbf{w}}, z_{i_t}) + \tilde{\boldsymbol{\mu}})$
    $\mathbf{w}_t = \text{prox}_{\eta\Omega}(\hat{\mathbf{w}}_t)$
  end for
  $\tilde{\mathbf{w}}_s = \mathbf{w}_m$
end for

Bounded Variance

Lemma 1. *Consider* VRSPAM *(Algorithm 1), then the variance of the* $\mathbf{v}_t$ *is upper bounded as:*

$$\mathbb{E}[\|\mathbf{v}_t - \partial f(\mathbf{w}_t)\|^2]) \leq 4(8M^2)^2\|\mathbf{w}_t - \mathbf{w}^*\|^2 + 2(8M^2)^2\|\tilde{\mathbf{w}} - \mathbf{w}^*\|^2$$

- At the convergence, $\tilde{\mathbf{w}} = \mathbf{w}^*$ and $\mathbf{w}_t = \mathbf{w}^*$
- Variance of the updates are bounded and go to zero as the algorithm converges
- Variance of the gradient in SPAM [2] does not go to zero as it is a stochastic gradient descent based algorithm

Convergence Analysis

Theorem 1. *Consider* VRSPAM *(Algorithm 1) and let* $\mathbf{w}^* = \arg\min_{\mathbf{w}} f(\mathbf{w}) + \Omega(\mathbf{w})$*; if* $\eta < \frac{\beta}{128M^4}$*, then there exists* $\alpha < 1$ *and we have the geometric convergence in expectation:*

$$\mathbb{E}[\|\mathbf{w_s} - \mathbf{w}^*\|^2] \leq \alpha^s\,\mathbb{E}[\|\mathbf{w_0} - \mathbf{w}^*\|^2]$$

- We get a geometric convergence rate of α^s which is much stronger than the $\mathcal{O}(\frac{1}{t})$ convergence rate obtained in [2].

Complexity Analysis:

- For any $0 < \theta < 1$ and $E = \frac{1}{(1+\frac{\theta\beta^2}{128M^4})}$, if we choose $m \approx 2\frac{\log\theta}{\log E}$ then $\alpha \approx 2\theta E^2$
- Thus the time complexity of the algorithm is $\mathcal{O}(n + 2\frac{\log\theta}{\log E})(\log(\frac{1}{\epsilon}))$ when $m = \Theta(\frac{\log\theta}{\log E})$
- As the order has inverse dependency on $\log E = \log\frac{128M^4}{128M^4+\theta\beta^2}$, increase in M will result in increase in number of iterations i.e. as the maximum norm of training samples is increased, larger m is required to reach ϵ accuracy.
- SPAM algorithm takes $\mathcal{O}(\frac{t_0 F}{\epsilon})$ iterations to achieve ϵ accuracy. Thus, SPAM has lower per iteration complexity but slower convergence rate as compared to VRSPAM. Therefore, VRSPAM will take less time to get a good approximation of the solution.

Results

- German: $n = 1000$, $p = 24$; USPS: $n = 9298$, $p = 256$; a9a: $n = 32,561$, $p = 123$

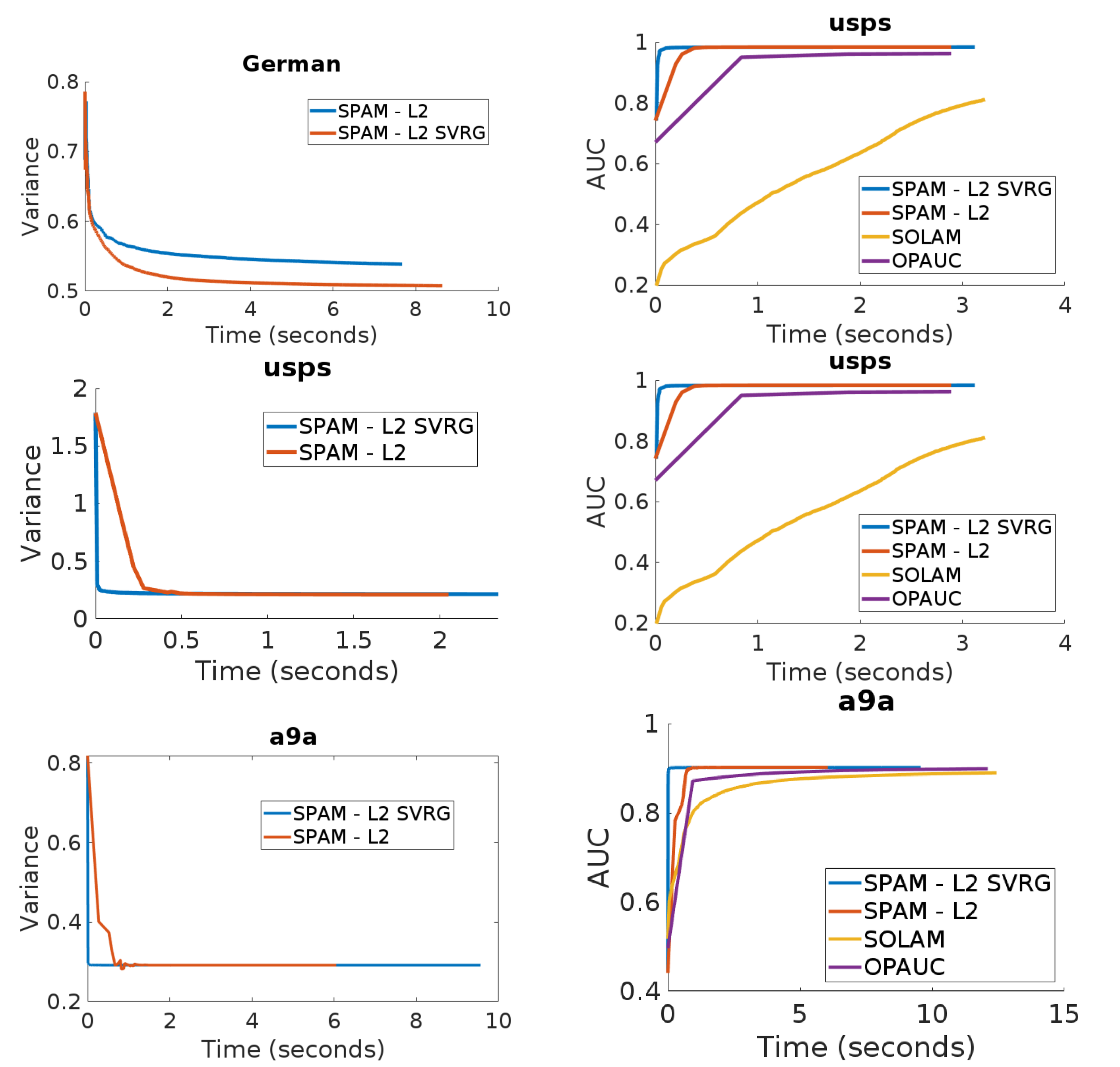

Fig. 1: The top row shows that VRSPAM (SPAM-L2-SVRG) has lower variance than SPAM-L2 across different datasets. The bottom row shows VRSPAM (SPAM-L2-SVRG) converges faster and performs better than existing algorithms on AUC maximization.

Conclusion

- Proposed variance reduced stochastic proximal algorithm for AUC maximization (VRSPAM).
- Obtained convergence rate of $\mathcal{O}(\alpha^t)$ where $\alpha < 1$, improving upon state-of-the-art methods [2] which have a convergence rate of $\mathcal{O}(\frac{1}{t})$.
- Showed theoretically and empirically VRSPAM converges faster than existing methods for AUC maximization.

References

[1] Charles Elkan. "The foundations of cost-sensitive learning". In: *International joint conference on artificial intelligence.* Vol. 17. 1. Lawrence Erlbaum Associates Ltd. 2001, pp. 973–978.

[2] Michael Natole, Yiming Ying, and Siwei Lyu. "Stochastic proximal algorithms for AUC maximization". In: *International Conference on Machine Learning.* 2018, pp. 3707–3716.

[3] Lin Xiao and Tong Zhang. "A proximal stochastic gradient method with progressive variance reduction". In: *SIAM Journal on Optimization* 24.4 (2014), pp. 2057–2075.

[4] Yiming Ying, Longyin Wen, and Siwei Lyu. "Stochastic online AUC maximization". In: *Advances in neural information processing systems.* 2016, pp. 451–459.