View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by CiteSeerX

JMLR: Workshop and Conference Proceedings 45:205–220, 2015 ACML 2015

Proximal Average Approximated Incremental Gradient Method for Composite Penalty Regularized Empirical Risk Minimization*

Yiu-ming Cheung YMC@COMP.HKBU.EDU.HK and **Jian Lou** JIANLOU@COMP.HKBU.EDU.HK
Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China

Editor: Geoffrey Holmes and Tie-Yan Liu

Abstract

Proximal average (PA) is an approximation technique proposed recently to handle nonsmooth composite regularizer in empirical risk minimization problem. For nonsmooth composite regularizer, it is often difficult to directly derive the corresponding proximal update when solving with popular proximal update. While traditional approaches resort to complex splitting methods like ADMM, proximal average provides an alternative, featuring the tractability of implementation and theoretical analysis. Nevertheless, compared to SDCA-ADMM and SAG-ADMM which are examples of ADMM-based methods achieving faster convergence rate and low per-iteration complexity, existing PA-based approaches either converge slowly (e.g. PA-ASGD) or suffer from high per-iteration cost (e.g. PA-APG). In this paper, we therefore propose a new PA-based algorithm called PA-SAGA, which is optimal in both convergence rate and per-iteration cost, by incorporating into incremental gradient-based framework.

1. Introduction

In this paper, we focus on empirical risk minimization which optimizes the model by minimizing the average loss from the training data. The averaged loss is often regularized by a regularizer, e.g. smooth penalties such as l_2 norm or nonsmooth penalties like lasso. More complex regularizers such as group lasso (Yuan et al. (2011)), Mairal et al. (2010)) and graph lasso (Jacob et al. (2009)) have been widely used in bioinformatics, text and other structured data mining tasks to introduce structured sparsity to the model. Optimizing such structure regularized empirical risk minimization problem can be challenging, especially confronted with very large dataset.

Generally, nonsmooth composite penalties such as group lasso and graph lasso are difficult to solve because we can neither directly take derivative as dealing with smooth regularizer like ℓ_2 norm nor have closed form proximal update for simple nonsmooth like ℓ_1 norm. In other words, none of these simple methods mentioned above can be directly applied when these complex regularizers are used. Apparently, other techniques need to be introduced. A popular line of research is a split method called alternating direction method of multipliers (ADMM) (Boyd et al. (2011)). There are already stochastic gradient and incremental gradient ADMM methods. Stochastic Gradient ADMM (Ouyang et al. (2013)) methods include RDA-ADMM (Suzuki (2013)) which incorporates RDA method with ADMM. SADMM and

[*] Yiu-ming Cheung is the corresponding author.

© 2015 Y.-m. Cheung & J. Lou.

optimal-SADMM in (Azadi and Sra (2014)) utilize nonuniform averaging of iterative variable (Lacoste-Julien et al. (2012), Shamir and Zhang (2013)) and accelerated stochastic gradient method (Ghadimi and Lan (2012)). In particular, SA-ADMM (Zhong and Kwok (2014b)) and SDCA-ADMM (Suzuki (2014)) are two recent ADMM methods incorporating incremental gradient method SAG (Roux et al. (2012a)) and SDCA (Shalev-Shwartz and Zhang (2013)) correspondingly. When incorporating into incremental gradient framework, both methods enjoy not only low per-iteration complexity, but also optimal convergence rate. A major drawback of ADMM-based methods is the complex implementation and convergence analysis, which are brought about by the additional variables introduced and the alternating updating scheme.

Recently, proximal average (Bauschke et al. (2008), Yu (2013), Zhong and Kwok (2014a)) has been introduced to deal with component penalties efficiently when each of its component allows cheap computation of proximal operator. It averages the proximal map from each component proximal operator, which is surprisingly simple in terms of implementation. In addition, (Zhong and Kwok (2014a)) shows that this proximal average update scheme is equivalent to solving a proximal mapping operation on a surrogate regularizer. More importantly, the approximation of this surrogate regularizer can be arbitrarily close to the original composite regularizer. Compared with ADMM, (Zhong and Kwok (2014a)) points out that ADMM is also a proximal method by duplicating variables. Apparently, proximal average is much simpler to implement and also much easier to make analysis, which will be introduced later. Pioneers includes (Yu (2013)), which introduces proximal average with accelerated full gradient method FISTA (Beck and Teboulle (2009)). (Zhong and Kwok (2014a)) incorporates proximal average technique with the stochastic variant of optimal gradient method. It has provable superiority over smoothing technique which is also shared by (Yu (2013)). Despite the simplicity advantage in terms of implementation and analysis, when compared to incremental ADMM methods (e.g. SA-ADMM and SDCA-ADMM), existing PA-based approaches either converge slowly (e.g. PA-ASGD) or suffer from high per-iteration cost (e.g. PA-APG). To make the PA-based method more appealing, it is worth incorporating PA technique into incremental gradient framework.

Incremental gradient methods featuring both scalability and fast convergence property have been receiving considerable attention as an efficient approach to mitigating the ever growing dataset problem. As these methods only calculate gradients associated with a randomly picked data sample in each iteration as stochastic gradient methods (Bottou (2010), Xiao (2010), Ghadimi and Lan (2012)), they have comparable low per-iteration computation cost. More importantly, by well exploiting the finite sum structure of the loss function which stochastic methods do not have, these incremental methods are able to achieve linear convergence rate as full gradient methods (Nesterov and Nesterov (2004)). For example, SAG (Roux et al. (2012a)) utilizes the average of the stored past gradients, one for each data. SVRG (Johnson and Zhang (2013), Xiao and Zhang (2014)) adopts a multi-stage scheme to progressively control the variance of the stochastic gradient. Both methods have linear convergence rate for strongly convex problem, but the theoretical convergence result for general convex loss is still unclear. SAGA (Defazio et al. (2014a), Defazio et al. (2014b)) has both sublinear convergence guarantee for general convex loss and linear convergence for strongly convex loss. It is a midpoint of SAG and SVRG by taking both update pattern from them in its iteration. There are also other incremental methods like FINITO (Defazio et al.) and MISO (Mairal (2014)), which consume more memory, in which they not only store the gradient, but also the variable. S2GD (Konečný and Richtárik (2013)) is a method very similar to SVRG with the difference only in stage length. SDCA (Shalev-Shwartz and Zhang (2013)) is a dual incremental method.

In this paper, we shall investigate the potential to incorporate incremental gradient methods with proximal average technique. We will show that, by solving a surrogate problem, the proposed method can achieve linear convergence for strongly convex problem and sublinear convergence rate for general convex problem. Meanwhile, it enjoys the simplicity advantage brought about by proximal average technique. Also, our method is provided with convergence analysis for both strongly convex and general convex loss functions, which is another advantage over ADMM-based incremental gradient methods. For instance, SDCA-ADMM only has convergence results for strongly convex loss, while the convergence analysis of SAG-ADMM only applies when the loss is general convex.

The remainder of this paper is organized as follows: Sections 2 introduces the techniques including incremental gradient method and proximal average, and gives the formulation of the problem. In Section 3, we propose our method for strongly convex loss problem and general convex loss with the corresponding convergence rate. Section 4 describes the related algorithms and compares them with our method. Section 5 shows the experimental results including efficiency verification for strongly convex problem and general convex problem on two real datasets. Finally, Section 6 concludes the paper.

2. Problem Formulation and Background

In the following, we denote the gradients of the differentiable function l_i and l at x as $l_i'(x)$ and $\nabla l(x)$, respectively. $||x||_2^2$ and $||x||_1$ denote the l_2 and l_1 norm of vector x correspondingly. $\langle l_i'(x), y\rangle$ is the inner product of $l_i'(x)$ and y. The superscript $(\cdot)^T$ stands for the transpose of $(\cdot)$.

2.1. Problem Formulation

Let us consider the following regularized empirical risk minimization problem:

$$\min_{x\in\mathbb{R}^d} F(x) = l(x) + r(x) = \frac{1}{n}\sum_{i=1}^{n} l_i(x) + \sum_{k=1}^{K} \alpha_k r_k(x) \text{ with } \sum_{k=1}^{K}\alpha_k = 1,\ \alpha_k \geq 0, \tag{1}$$

where $l_i(x)$ is the loss taken at data sample (ξ_i, y_i) with index i. ξ_i is the data vector, and y_i is its label. For example, the hinge loss used in SVM defines $l_i(x)$ as $\max\{0, 1 - y_i\xi_i^T x\}$ and the logistic loss for logistic regression defines $l_i(x)$ as $\log(1+\exp(-y_i\xi_i^T x))$. $r(x)$ is composite penalty, for example graph-guided fused lasso and overlapping group lasso. We hide the constant balancing the loss and the regularizer in the loss as (Yu (2013)) and (Zhong and Kwok (2014a)) so that $r(x)$ is a convex combination of the K components $r_k(x)$. We make the following assumptions to the loss and the penalty functions.

Assumption 1 *The loss $l_i(x)$ is convex if, for any $x, y \in \mathbb{R}^d$,*

$$l_i(y) - l_i(x) - \langle l_i'(x), y - x\rangle \geq 0. \tag{2}$$

In addition, if there is a $\mu > 0$ such that:

$$l_i(y) - l_i(x) - \langle l_i'(x), y - x \rangle \geq \frac{\mu}{2}||y - x||_2^2, \tag{3}$$

then $l_i(x)$ is strongly convex. Also $l_i(x)$ is smooth with L Lipschitz continuous gradient, i.e.

$$l_i(y) - l_i(x) - \langle l_i'(x), y - x \rangle \leq \frac{L}{2}||y - x||_2^2, \tag{4}$$

for all $y, x \in \mathbb{R}^d$.

The next assumption is required by proximal average technique (Yu (2013), Zhong and Kwok (2014a)). We first introduce the notation related to proximal step:

$$M_{r_k}^{\eta}(x) = \min_y \frac{1}{2\eta}||x - y||_2^2 + r_k(y), \tag{5}$$

and

$$P_{r_k}^{\eta}(x) = \arg\min_y \frac{1}{2\eta}||x - y||_2^2 + r_k(y). \tag{6}$$

Assumption 2 *Each $r_k(x)$ is convex and Lipschitz continues with constant M_k, i.e.*

$$|r_k(x) - r_k(y)| \leq M_{r_k}||x - y||_2, \text{ for all } x, y \in \mathbb{R}^d. \tag{7}$$

In addition, the proximal map $P_{r_k}^{\eta}$ for each r_k individually is simple to compute (e.g. has closed form solution) for any $\eta > 0$.

2.2. Incremental Gradient Method

The incremental gradient methods proposed recently make improvement on stochastic gradient methods provided that the training data is finite. Generally, at each iteration, these methods approximate the full gradient by a combination of a random gradient evaluated at the latest variable with past gradients. SAG utilizes a table to record past gradients one for each data sample. SVRG uses a single full gradient evaluated periodically. Both of the methods have linear convergence for strongly convex and smooth problem. SAGA shares part of update pattern from both SAG and SVRG, and has theoretical guarantee for both general convex and strongly convex problem. In addition, SAGA also admits proximal operator for simple nonsmooth regularizer as SVRG.

Let $l_i'(\phi_i^t)(i = 1, 2, ..., n)$ be the gradient of data i kept in the gradient table up to iteration t. SAGA, like SAG, updates the random i_t-th gradient with $l_{i_t}'(x^t)$ while keeping other terms unchanged:

$$l_i'(\phi_i^{t+1}) = \begin{cases} l_i'(x^t), & i = i_t \\ l_i'(\phi_i^t), & i \neq i_t. \end{cases} \tag{8}$$

Then, SAGA approximates the gradient for iteration t in a similar manner as SVRG:

$$g^t = l_{i_t}'(\phi_i^{t+1}) - l_{i_t}'(\phi_i^t) + \frac{1}{n}\sum_{i=1}^{n} l_i'(\phi_i^t). \tag{9}$$

Conditioned on information up to the t-th iteration, g^t is an unbiased estimation of the full gradient in expectation. According to (Johnson and Zhang (2013), Xiao and Zhang (2014)), such approximate gradients have the reduced variance, which would lead to speedup over stochastic methods.

2.3. Proximal Average

Proximal average (Bauschke et al. (2008), Yu (2013)), first introduced by (Yu (2013)) to deal with composite regularizer, admits a compact calculation when each single component satisfies Assumption 2. With proximal average, one actually uses a surrogate regularizer $\hat{r}(x)$.

Definition 1 *(Proximal Average Bauschke et al. (2008), Yu (2013)) The proximal average of r is the unique semicontinuous convex function $\hat{r}(x)$ such that $M^{\eta}_{\hat{r}(x)} = \sum_{k=1}^{K} \alpha_k M^{\eta}_{r_k}$. The corresponding proximal map of the proximal average $\hat{r}(x)$ is*

$$P^{\eta}_{\hat{r}(x)}(x) = \sum_{k=1}^{K} \alpha_k P^{\eta}_{r_k}(x). \tag{10}$$

The next lemma shows that the approximation of $\hat{r}(x)$ can be controlled arbitrarily close to $r(x)$ by the step size η.

Lemma 2 *(Yu (2013)) Under Assumption 2, we have $0 \leq r(x) - \hat{r}(x) \leq \frac{\eta \bar{M}^2}{2}$, where $\bar{M}^2 = \sum_{k=1}^{K} \alpha_k M_k^2$.*

3. Incremental Gradient Method with Proximal Average (PA-SAGA)

This section describes the proposed algorithm, denoted as PA-SAGA, based on incremental gradient decent SAGA and proximal average techniques. The procedure is shown in Algorithm 1.

3.1. Algorithm Description

The proposed PA-SAGA initializes the step size with the constant η, the variable with x_0, and the gradient table with $l_i'(x_0)$ for each item. Step 2-3 updates the gradient table, followed by Step 4 calculating the gradient. Step 5 and 6 can be combined into

$$x_{k+1} = \sum_{k=1}^{K} \alpha_k P^{\eta}_{r_k}(x_k - \eta g^t), \tag{11}$$

which uses proximal average to compute the next step. Compared with SAGA, we replace its proximal mapping which can be applied to simple regularizers only with proximal average. According to the property of proximal average, PA-SAGA approximates the original component regularizer with the proximal average as a surrogate. The approximation can be controlled arbitrarily close by the step size parameter η.

Algorithm 1 PA-SAGA

```
Input: $\eta$ (learning rate); $x_0$ (initial value); $l'_i(\phi_i^0), \phi_i^0 = x_0, i = 1, ..., n$ (initial table of gradients).
for $t = 1, 2, ...$ do
    Randomly pick $i_t \in \{1, 2, ..., n\}$;
    Update the derivative table as in equation (8);
    Calculate $g^t$ by equation (9);
    $w_{k+1} = x_k - \eta g^t$;
    $x_{k+1} = \sum_{k=1}^{K} \alpha_k P_{r_k}^{\eta}(w_{k+1})$;
end for
```

3.2. Convergence Analysis

In this section, we discuss the convergence property of PA-SAGA. We denote $\hat{F}(x) = l(x) + \hat{r}(x)$ as the surrogate problem implicitly solved by PA-SAGA. The next two theorems describe the convergence for strongly convex problem and general convex problem, respectively.

Theorem 3 *Under Assumption 1 with $l_i(x)$ μ-strongly convex and Assumption 2, let $\hat{x}^*$ be the optimal point of the surrogate problem. Denote a Lyapunov function T^t as:*

$$T^t = Q^t + \big(c_1 + \frac{c_2}{\eta}\big)||x^t - \hat{x}^*||_2^2 + c_2\big(\hat{F}(x^t) - \hat{F}(\hat{x}^*)\big), \tag{12}$$

$$Q^t = \frac{1}{n}\sum_{i=1}^{n} l_i(\phi_i^t) - l(\hat{x}^*) - \frac{1}{n}\sum_{i=1}^{n}\left\langle l'_i(\hat{x}^*), \phi_i^t - \hat{x}^*\right\rangle, \tag{13}$$

where t is the iteration number. After $(1 - \frac{1}{\kappa})(\log \frac{T^0}{\epsilon})$ iterations, we then have

$$\mathbb{E}\Big[F(x^t) - F(\hat{x}^*)\Big] \leq 2\epsilon. \tag{14}$$

In addition, there exits some $\beta \geq 1$ and possible choices of the parameters c_1, c_2, κ, η appeared in the proof are as follows: $\eta < \min(\frac{1}{2L}, \frac{2\epsilon}{M^2}, \frac{1}{2n\mu})$, $c_1 = \frac{1}{2\eta n}\frac{L}{L-\mu}$, $c_2 = c_1\eta\Big(\frac{1}{2\eta L\beta} - 1\Big)$, $\frac{1}{\kappa} = \frac{2\eta\mu}{1+\frac{1}{2\eta L}}$.

Theorem 4 *Under Assumption 1 with $l_i(x)$ general convex and Assumption 2, let $\hat{x}^*$ be the optimal point of the surrogate problem. With the same definition of Q^t as in Theorem 3, after $t \geq \frac{1}{c_2\epsilon}\Big(Q^0 + \big(c_1 + \frac{c_2}{2\eta}\big)||x^0 - \hat{x}^*||_2^2\Big)$ iterations, we then have*

$$\mathbb{E}\big[F(\bar{x}^t) - F(\hat{x}^*)\big] \leq 2\epsilon, \tag{15}$$

where $\bar{x}^t = \frac{1}{t}\sum_{i=1}^{t} x^i$. In addition, possible choices of the parameters c_1, c_2, η appeared in the proof are as follows: $\eta < \min(\frac{1}{2L}, \frac{2\epsilon}{M^2})$, $c_1 = \frac{1}{2\eta n}$, $c_2 = \frac{1}{2n}\Big(\frac{1}{2\eta L\beta} - 1\Big)$.

We defer the proof to the appendix. We represent all parameters by the step size η which controls the approximation by Lemma 2. The convergence rate of strongly convex case is related to $\frac{1}{\kappa}$, i.e. it converges faster when $\frac{1}{\kappa}$ is larger, which depends on L, μ and ϵ as given the data size n. Please note that, for an ill-conditioned problem where $\frac{L}{u} = n$, $\frac{1}{2n\mu}$ can be converted to $\frac{1}{2L}$. Thus, the convergence speed is related to L and ϵ. The convergence speed for the general convex case depends on c_2, i.e. the larger c_2 is, the faster it converges. Given the dataset size, the convergence speed is again related to L and ϵ.

Like the other incremental methods, the above convergence only reflects training loss (Suzuki (2014), Zhong and Kwok (2014b), Roux et al. (2012b)). The generalization performance is unknown partly because of the assumption of the finite training set size. Our experiments on testing loss show empirical results of the generalization performance.

Furthermore, our algorithm will converge to the optimal point of the surrogate function. We show the convergence rate by measuring the loss with respect to the objective function value at $\hat{x}^*$ ($F(\hat{x}^*)$), which is different from usual convention that measure with $F(x^*)$. Nevertheless, considering the over-fitting issue, a relative good approximation is potentially able to achieve satisfactory generalization performance. As a good approximation to the original problem, it is expected that our method will have satisfactory generalization performance. Indeed, the experimental results in Section 5 have verified this in terms of classification error and test loss on the test set of two real datasets.

4. Related Work

This section describes more recently proposed related algorithms targeting at problem (1). Stochastic ADMM methods include SGD-ADMM (Ouyang et al. (2013)) and RDA-ADMM (Suzuki (2013)). These methods have $O(\frac{1}{\sqrt{T}})$ convergence rate for general convex case and $O(\frac{\log T}{T})$ for strongly convex case. SADMM (Azadi and Sra (2014)) utilizes nonuniform averaging of the iterate variable and the accelerated stochastic gradient method for smooth loss function to accelerate the speed. It has $O(\frac{1}{T})$ convergence rate for strongly convex loss and $O(\frac{1}{T^2}) + O(\frac{1}{T}) + O(\frac{1}{\sqrt{T}})$ for smooth loss function.

SA-ADMM (Zhong and Kwok (2014b)) and SDCA-ADMM (Suzuki (2014)) are incremental gradient ADMM methods. Briefly, SA-ADMM (Zhong and Kwok (2014b)) uses SAG with ADMM and shows $O(\frac{1}{T})$ convergence rate for general convex problem and empirically it has fast convergence for strongly convex case. SDCA-ADMM solves problem that can be expressed in the form of

$$\min_x \frac{1}{n} \sum_{i=1}^{n} l_i(\xi_i^T x) + r(B^T x).$$

It takes Fenchel dual to convert the original problem to its dual form with a linear constraint. Then, it solves the dual problem by ADMM method. SDCA-ADMM has both single update and mini-batch update form, both of which are easy to be parallelized. Since its convergence analysis requires the regularizer to be locally strongly convex, for regularizers such as graph lasso, it perturbs the regularizer with a small l_2 norm counterpart. However, even without perturbation to guarantee locally strongly convex, it is empirically observed that it converges fast.

PA-APG (Yu (2013)) first introduces proximal average technique into regularized risk minimization. It uses proximal average with proximal FISTA (Beck and Teboulle (2009)) method. By virtue of the convergence analysis of FISTA (Beck and Teboulle (2009)) not relying on the optimal point x^*, PA-APG converges to the optimal of the original problem with the well chosen η. However, PA-APG is a batch method that does not well suit large scale problem. PA-ASGD (Zhong and Kwok (2014a)) incorporates proximal average with the accelerated stochastic gradient method. It has $O(\frac{1}{T^2}) + O(\frac{\log T}{T^2}) + O(\frac{1}{\sqrt{T}})$ for strongly convex problem and $O(\frac{1}{T^2}) + O(\frac{1}{T^{\frac{3}{2}}}) + O(\frac{1}{\sqrt{T}}) + O(\frac{1}{T})$ for general convex problem. The slowest term $O(\frac{1}{\sqrt{T}})$ is from the variance of the approximate gradient.

5. Experiments

This section evaluated the performance of the proposed method in comparison with two incremental gradient ADMM: SA-ADMM (Zhong and Kwok (2014b)) and SDCA-ADMM (Suzuki (2014)) along with a proximal average based stochastic gradient PA-ASGD (Zhong and Kwok (2014a)). We did not compare with batch gradient proximal average method as Zhong and Kwok (2014a) has already shown that it was less efficient than PA-ASGD. We utilized the real dataset '20 Newsgroup' and 'a9a', which were also used by Suzuki (2014). 80% of the data were randomly sampled for training to study the convergence and efficiency of the algorithms, while the rest are considered as testing data used to study the generalization performance. We investigated the algorithms on general convex loss problem and strongly convex loss problem, respectively.

5.1. General Convex Loss Function Problem

Let us consider the general convex loss problem, in which the smoothed hinge loss is:

$$l_i(u) = \begin{cases} 0, & y_i u \geq 1 \\ \frac{1}{2} - y_i u, & y_i u \leq 0 \\ \frac{1}{2}(1 - y_i u)^2, & \text{otherwise,} \end{cases} \tag{16}$$

where $u = \xi_i^T x$, (ξ_i, y_i) is the i-th data sample. We utilized the graph guided fused lasso

$$\lambda\big(||x||_1 + \sum_{\{i,j\}\in E} |x_i - x_j|\big) \tag{17}$$

as the composite regularizer. We constructed the graph as in Suzuki (2014) and set λ at 0.001. The proximal map for $||x||_1$ is simply soft thresholding. The proximal map for $|x_i - x_j|$ is

$$[P_{r_k}^{\eta}]_s = \begin{cases} x_s - sign(x_i - x_j)\min\{\eta, \frac{|x_i - x_j|}{2}\}, & s \in \{i, j\} \\ x_s, & \text{otherwise} \end{cases} \tag{18}$$

as given in (Yu (2013), Zhong and Kwok (2014a)). We reported the empirical risk, which is the training loss, against the number of iterations for both datasets. As for the generalization performance, we reported the classification error measured on testing set against the number of iterations and CPU time for the '20 Newsgroup' dataset. We plotted the testing loss against the number of iteration and CPU time for 'a9a' dataset.

As shown in Figure 1, our method decreased the training loss almost the same as the other two ADMM-based incremental gradient methods and had the similar generalization performance in terms of classification error on '20 Newsgroup' dataset. On the 'a9a' dataset (see Figure 2), the proposed method achieved the similar performance as SDCA-ADMM method in both training and generalization performance, while it is a bit inferior to SA-ADMM. In both datasets, our method is more efficient than PA-ASGD. Also, the classification error on '20 Newsgroup' testing set and test loss on 'a9a' testing set indicated that our solution obtained by the surrogate to regularizer is able to achieve satisfactory generalization performance.

5.2. Strongly Convex Loss Function Problem

For strongly convex case, we utilized the logistic loss with the large margin graph lasso regularizer as in (Zhong and Kwok (2014a)), i.e.

$$\lambda\big(||x||_2^2 + \sum_{\{i,j\}\in E} |x_i - x_j|\big). \tag{19}$$

We processed the logistic loss and the l_2 norm together to ensure the strong convexity of the loss part. Note that, in this case, the l_2 norm term can neither be incorporated into $l_i(\xi_i^T x)$, nor into $||Ax||_1$ form, thus SDCA-ADMM is unable to handle this case. We only compared with the other two methods. We reported the training loss, classification error for this case on '20 Newsgroup' dataset.

In this case, according to the experiment result (see Figure 3), our method has comparable convergence and efficiency performance with SA-ADMM in terms of both training loss and classification error against both of number of iterations and CPU time, and is better than PA-ASGD.

5.3. Discussion

We would like to point out that, as a proximal average method, our method has generally better performance than stochastic gradient-based method PA-ASGD in term of all performance metrics we have tried so far. As an incremental gradient-based method, the proposed method has comparable performance with SDCA-ADMM and SA-ADMM, but the merit of the proposed method is two-fold: (1) SDCA-ADMM's convergence analysis relies on the local strongly convexity of the loss function. In addition, SDCA-ADMM requires the dual problem to be in structure for ADMM to be applied to, which causes a stricter problem format and therefore limits its application domain. For example, in the case studied in Sub-section 5.2, SDCA-ADMM cannot work at all because the dual parts do not fit into the structure for ADMM, despite each dual of their primal correspondences is easy to take. By contrast, the proposed method has given the convergence analysis for both of general convex loss and strongly convex loss problems. Further, the format of objective function in our method is more general than SDCA-ADMM; (2) SA-ADMM lacks convergence analysis for strongly convex loss problem, but the proposed one does.

6. Conclusion

In this paper, we have proposed a new incremental gradient method for empirical risk minimization regularized by composite regularizer. As a proximal average technique-based method, it is more efficient and faster than its existing batch and stochastic counterpart. Compared with popular ADMM-based incremental gradient, it has comparable performance, yet enjoys more compact update form and simpler theoretical analysis by virtue of the proximal average technique. Experimental results on two real datasets have shown its efficiency and satisfactory generalization performance.

Appendix

The appendix contains proof sketch for Theorem 3 and Theorem 4, which is a combination of the proof in (Defazio et al. (2014b)) and (Yu (2013)). We proceed the proof by two steps. We first prove that, for the proximal average approximation $\hat{F}(x) = l(x) + \hat{r}(x)$ and its global optimal value $\hat{F}(\hat{x}^*)$, $\hat{F}(x^t) - \hat{F}(\hat{x}^*)$ converges linearly in expectation. Then, we conclude the proof using Lemma 2, which shows the surrogate $\hat{F}(x)$ and the original $F(x)$ can be arbitrarily close. Note that the step size η controls the approximation, and we express the other parameters in terms of η. Also, to safely leave out terms appeared in the proof, it requires a careful verification of the relationship among the parameters. This is different from (Defazio et al. (2014b)) as our η has the additional approximation constraint.

Proof Sketch of Theorem 3

With the parameters chosen in the theorems, for strongly convex case, we can get

$$\begin{aligned}\mathbb{E}\big[\hat{F}(x^t) - \hat{F}(\hat{x}^*)\big] &\leq \mathbb{E}\big[T^t\big] \leq \big(1 - \frac{1}{\kappa}\big)^t T^0 \\ &= \big(1 - \frac{1}{\kappa}\big)^t \Big[Q^0 + \big(c_1 + \frac{c_2}{\eta}\big)||x^0 - \hat{x}^*||_2^2 + c_2\big(\hat{F}(x^0) - \hat{F}(\hat{x}^*)\big)\Big].\end{aligned} \tag{20}$$

Hence, as long as $t \geq \log \frac{T^0}{\epsilon} / \log(1 - \frac{1}{\kappa})$, we have

$$\mathbb{E}\big[\hat{F}(x^t) - \hat{F}(\hat{x}^*)\big] \leq \epsilon. \tag{21}$$

By Lemma 2, we have

$$\begin{aligned}\mathbb{E}\big[F(x^t) - F(\hat{x}^*)\big] &= \mathbb{E}\big[\big(F(x^t) - \hat{F}(x^t)\big) + \big(\hat{F}(x^t) - \hat{F}(\hat{x}^*)\big) + \big(\hat{F}(\hat{x}^*) - F(\hat{x}^*)\big)\big] \\ &\leq \epsilon + \epsilon + 0.\end{aligned} \tag{22}$$

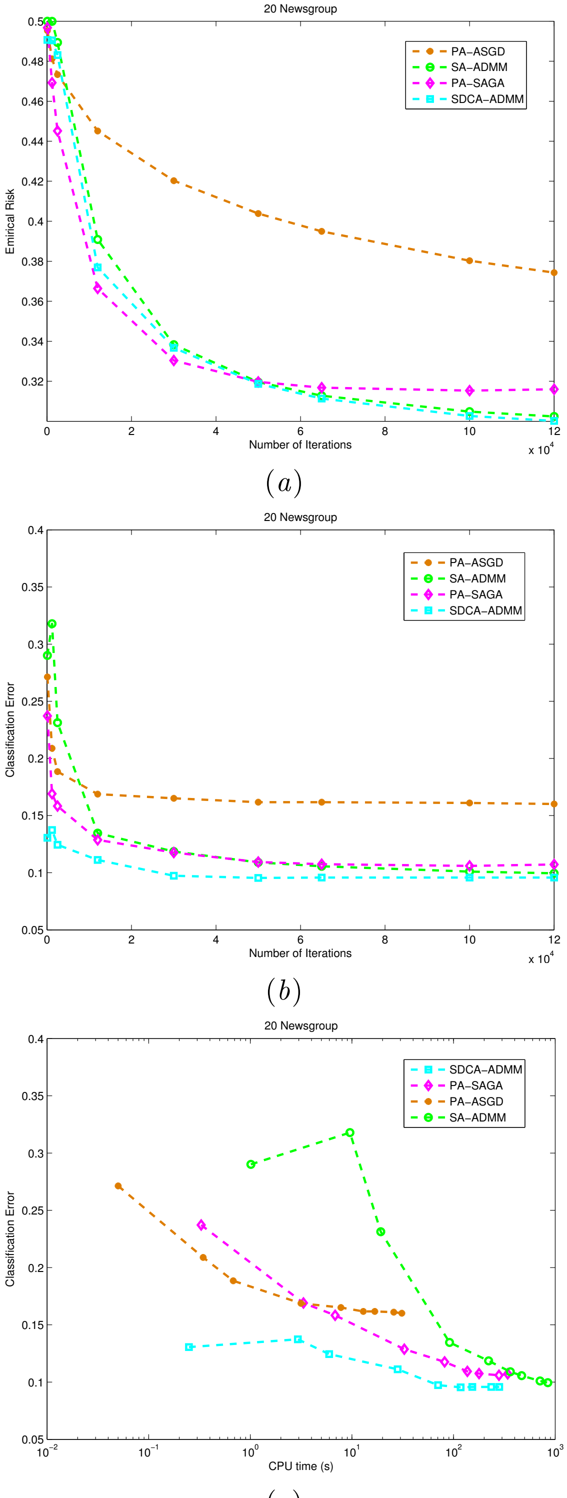
Thus, as long as $t \geq \log \frac{T^0}{\epsilon} / \log(1 - \frac{1}{\kappa})$, we get $\mathbb{E}\big[F(x^t) - F(\hat{x}^*)\big] \leq 2\epsilon$.

Proof Sketch of Theorem 4

With the parameters chosen in the theorems, for general convex case, we have:

$$\mathbb{E}c_2\Big[\hat{F}(\bar{x}^t) - \hat{F}(\hat{x}^*)\Big] \leq \frac{1}{t}\Big(Q^0 + \big(c_1 + \frac{c_2}{2\eta}\big)||x^0 - \hat{x}^*||_2^2\Big), \tag{23}$$

By Lemma 2, we then obtain:$\mathbb{E}\big[F(\bar{x}^t) - F(\hat{x}^*)\big] \leq \epsilon + \epsilon + 0$. Thus, as long as $t \geq \frac{1}{c_2\epsilon}\Big(Q^0 + \big(c_1 + \frac{c_2}{2\eta}\big)||x^0 - \hat{x}^*||_2^2\Big)$, $\mathbb{E}\big[F(\bar{x}^t) - F(\hat{x}^*)\big] \leq 2\epsilon$.

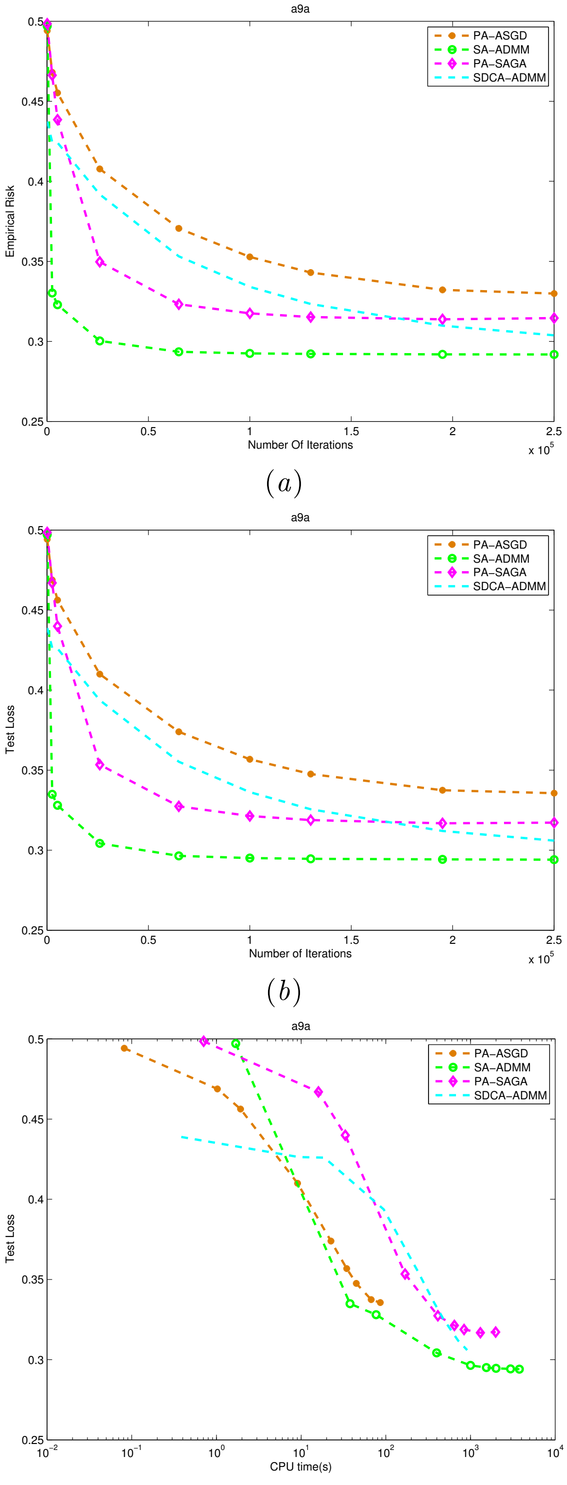


(a)

(b)

(c)

Figure 1: Smooth hinge loss with the graph lasso on 20 Newsgroup. Up: Training loss against number of iterations; Middle: Classification error against number of iterations; Bottom: Classification error against time.

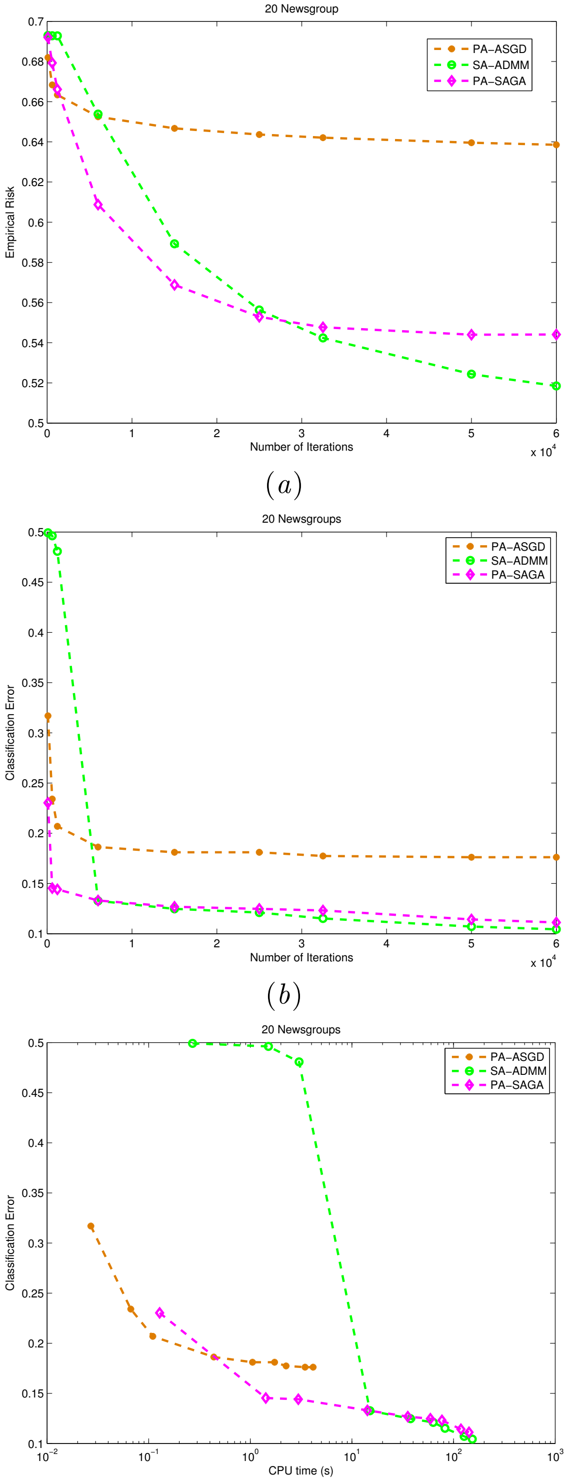


(a)

(b)

(c)

Figure 2: Smooth hinge loss with the graph lasso on a9a. Up: Training loss against number of iterations; Middle: Test loss against number of iterations; Bottom: Test loss against time.



(a)

(b)

(c)

Figure 3: Logistic loss with the large margin graph lasso on 20 Newsgroup. Up: Training loss against number of iterations; Middle: Classification error against number of iterations; Bottom: Classification error against time.

Acknowledgments

This work was supported by the Faculty Research Grant of Hong Kong Baptist University under Projects: FRG2/14-15/075 and FRG1/14-15/041, and the National Science Foundation of China under Grant: 61272366.

References

Samaneh Azadi and Suvrit Sra. Towards an optimal stochastic alternating direction method of multipliers. In *Proceedings of The 31st International Conference on Machine Learning*, pages 620–628, 2014.

Heinz H Bauschke, Rafal Goebel, Yves Lucet, and Xianfu Wang. The proximal average: basic theory. *SIAM Journal on Optimization*, 19(2):766–785, 2008.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *arXiv preprint arXiv:1407.0202*, 2014a.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *NIPS*, 2014b.

Aaron J Defazio, ANUEDU AU, Tibério S Caetano, NICTA COM AU, and Justin Domke. Finito: A faster, permutable incremental gradient method for big data problems.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440. ACM, 2009.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

Jakub Konečnỳ and Peter Richtárik. Semi-stochastic gradient descent methods. *arXiv preprint arXiv:1312.1666*, 2013.

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

Julien Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *arXiv preprint arXiv:1402.4419*, 2014.

Julien Mairal, Rodolphe Jenatton, Francis R Bach, and Guillaume R Obozinski. Network flow algorithms for structured sparsity. In *Advances in Neural Information Processing Systems*, pages 1558–1566, 2010.

Yurii Nesterov and IU E Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004.

Hua Ouyang, Niao He, Long Tran, and Alexander Gray. Stochastic alternating direction method of multipliers. In *Proceedings of the 30th International Conference on Machine Learning*, pages 80–88, 2013.

Nicolas L Roux, Mark Schmidt, and Francis R Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems*, pages 2663–2671, 2012a.

Nicolas L. Roux, Mark Schmidt, and Francis R. Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2663–2671. Curran Associates, Inc., 2012b.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *The Journal of Machine Learning Research*, 14(1):567–599, 2013.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of The 30th International Conference on Machine Learning*, pages 71–79, 2013.

Taiji Suzuki. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 392–400, 2013.

Taiji Suzuki. Stochastic dual coordinate ascent with alternating direction method of multipliers. In *Proceedings of The 31st International Conference on Machine Learning*, pages 736–744, 2014.

Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *arXiv preprint arXiv:1403.4699*, 2014.

Yao-Liang Yu. Better approximation and faster algorithm using the proximal average. In *Advances in Neural Information Processing Systems*, pages 458–466, 2013.

Lei Yuan, Jun Liu, and Jieping Ye. Efficient methods for overlapping group lasso. In *Advances in Neural Information Processing Systems*, pages 352–360, 2011.

Leon Wenliang Zhong and James T Kwok. Accelerated stochastic gradient method for composite regularization. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 1086–1094, 2014a.

Wenliang Zhong and James Kwok. Fast stochastic alternating direction method of multipliers. In *Proceedings of The 31st International Conference on Machine Learning*, pages 46–54, 2014b.