Adaptive Proximal Average Approximation for Composite Convex Minimization

Li Shen,† Wei Liu,‡ Junzhou Huang,♮ Yu-Gang Jiang,♯ Shiqian Ma§
†School of Mathematics, South China University of Technology, Guangzhou, China
‡Tencent AI Lab, Shenzhen, China
♮Department of Computer Science and Engineering, University of Texas at Arlington, USA
♯ School of Computer Science, Fudan University, Shanghai, China
§Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, China
shen.li@mail.scut.edu.cn, wliu@ee.columbia.edu, sqma@se.cuhk.edu.hk

Abstract

We propose a fast first-order method to solve multi-term nonsmooth composite convex minimization problems by employing a recent proximal average approximation technique and a novel adaptive parameter tuning technique. Thanks to this powerful parameter tuning technique, the proximal gradient step can be performed with a much larger stepsize in the algorithm implementation compared with the prior PA-APG method (Yu 2013), which is the core to enable significant improvements in practical performance. Moreover, by choosing the approximation parameter adaptively, the proposed method is shown to enjoy the $\mathcal{O}(\frac{1}{k})$ iteration complexity theoretically without needing any extra computational cost, while the PA-APG method incurs much more iterations for convergence. The preliminary experimental results on overlapping group Lasso and graph-guided fused Lasso problems confirm our theoretic claim well, and indicate that the proposed method is almost five times faster than the state-of-the-art PA-APG method and therefore suitable for higher-precision required optimization.

Introduction

Let $\mathbb{X}$ be a finite-dimensional linear space endowed with the inner product $\langle \cdot, \cdot \rangle$ and its induced norm $\| \cdot \|$. Here, we are interested in solving the following multi-term nonsmooth composite convex minimization problem

$$F^* := \min_{x \in \mathbb{X}} F(x) = f(x) + g(x) \tag{1}$$

with $g(x) = \sum_{i=1}^N \alpha_i g_i(x)$, where $\alpha_i \geq 0$ satisfying $\sum_{i=1}^N \alpha_i = 1$, $g_i : \mathbb{X} \to [-\infty, +\infty]$ is a proper, closed convex function, and $f : \mathbb{X} \to (-\infty, +\infty)$ is a continuously differentiable and gradient Lipschitz convex function with modulus L_f, *i.e.*,

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\|, \qquad \forall x, y \in \mathbb{X}.$$

Moreover, we assume that $\mathcal{Q}_i = \operatorname{dom} g_i^*$ is a bounded convex set for all $i = 1, \cdots, N$, in which g_i^* denotes the Fenchel conjugate of g_i with the following definition

$$g_i^*(x) = \sup_{u_i \in \mathcal{Q}_i} \left\{ \langle u_i, x \rangle - g_i(u_i) \right\}. \tag{2}$$

Notice that the boundedness assumption about $\mathcal{Q}_i$ is actually equivalent to the global Lipschitz continuousness of g_i used in (Yu 2013, Assumption 1) according to (Borwein and Vanderwerff 2010, Proposition 4.4.6).

This multi-term nonsmooth composite convex minimization problem (1) covers a large number of important applications in machine learning, such as overlapping group Lasso (Zhao, Rocha, and Yu 2009; Mairal et al. 2010), graph-guided fused Lasso (Chen et al. 2012; Kim and Xing 2009), graph-guided logistic regression (Ouyang et al. 2013), and other types of regularized risk minimization problems (Teo et al. 2010). The regularization term $g(x) = \sum_{i=1}^N \alpha_i g_i(x)$ often carries some important structure information about the structure of the problem itself or data, such as the structured sparsity (Bach et al. 2011; 2012) and nonnegativity. However, the involved vital multi-term nonsmooth components make the optimization problem (1) too complicated to be solved even if N is small. For the special case $N = 0, 1$, the most popular first-order methods are the accelerated gradient-type methods enjoying the $\mathcal{O}(\frac{1}{K^2})$ optimal iteration complexity (Nesterov 2013b), which were first proposed by Nesterov (Nesterov 1983) for $N = 0$ and then popularized for $N = 1$ by Beck and Teboulle (Beck and Teboulle 2009a) and Nesterov (Nesterov 2013a). Beck and Teboulle's method is called "FISTA" while Nesterov's method in (Nesterov 2013a) is called "APG". When N is larger, one feasible method is the subgradient-type method (Nemirovsky, Yudin, and Dawson 1982; Polyak 1977) with an extremely slow iteration complexity $\mathcal{O}(\frac{1}{\sqrt{K}})$. To opt out of this dilemma, Nesterov proposed the smoothed accelerated proximal gradient (S-APG) method (Nesterov 2005a; 2005b) for dealing with nonsmooth minimization involving multi-term nonsmooth functions. To make the smoothed accelerated proximal gradient method to achieve the iteration complexity $\mathcal{O}(\frac{1}{K})$, the smoothing parameter must be taken as small as $\mathcal{O}(\frac{1}{K})$. However, the small smoothing parameter leads to a small iteration stepsize, which has a negative effect on practical optimization performance. To make the smoothed accelerated proximal gradient method much more appealing, some adaptive smoothing algorithms (Boţ and Hendrich 2015; Tran-Dinh 2015) were proposed based on Nesterov's smoothing technique. However, another drawback is that the smoothing technique may compromise the important structure contained in those nonsmooth terms. Moreover, as N grows, the approximation error is linearly increasing with respect to N. Regarding this issue, Yu (Yu 2013) utilized the proximal average approximation technique that exploits the proximal mapping information of each nonsmooth term to decrease the approximation error, and then proposed the proximal average based accelerated proximal gradient (PA-APG) method which shows promising performance compared with the smoothed accelerated proximal gradient method in (Nesterov 2005b).

To let the PA-APG method enjoy the $\mathcal{O}(\frac{1}{K})$ iteration complexity, PA-APG has to suffer from the same issue that the approximation parameter must be as small as $\mathcal{O}(\frac{1}{K})$. This issue will make PA-APG impractical if we need to attain higher-precision optimization, which has been demonstrated in our experiments. To tackle this difficulty, we unite an easy-to-use adaptive approximation technique and the proximal approximation technique to propose an **A**daptive **P**roximal **A**verage based **A**ccelerated **P**roximal **G**radient (APA-APG) method, which still enjoys the $\mathcal{O}(\frac{1}{K})$ iteration complexity without increasing any extra computational cost. In contrast, Yu's PA-APG method in (Yu 2013) incurs much more iterations to reach convergence. It should be emphasized that such a combination is nontrivial and also very efficient in enhancing the optimization performance. To accomplish the $\mathcal{O}(\frac{1}{K})$ iteration complexity, we first derive the dual formulation of one proximal average approximation function based on the convex analysis techniques (Rockafellar 2015), and then leverage the dual formulation and Danskin's min-max theory (Bertsekas 1999, Proposition B.25) to establish a much tighter lower estimation of the proximal average approximation function, which is crucial to proving the $\mathcal{O}(\frac{1}{K})$ complexity of the APA-APG method. At last, we evaluate our proposed APA-APG method by executing overlapping group Lasso and graph-guided fused Lasso tasks. All experimental results indicate that our APA-APG method is about five times faster than the PA-APG method in (Yu 2013).

The rest of this paper is organized as follows. In Section 2 we first give the definition of one proximal average function, then derive its dual formulation, and discover a much stronger property of the proximal average function. In Section 3 we present our proposed APA-APG method along with its two variants denoted by APA-APG1 and APA-APG2, respectively. Importantly, we prove its $\mathcal{O}(\frac{1}{k})$ iteration complexity. In Section 4 we conduct experiments to evaluate APA-APG. Finally, we draw our conclusions in Section 5.

In the rest of the paper, we use the notations $\mathcal{A} := (\alpha_1, \cdots, \alpha_N)$, $\mathcal{B} := \mathrm{Diag}(\alpha_1, \cdots, \alpha_N)$ and $\mathcal{C} := \mathcal{B} - \mathcal{A}^*\mathcal{A}$. Let $\mathcal{I}_N$ denote the identity matrix with dimension $\mathbb{R}^{N\times N}$. It is easy to check that $0 \preceq \mathcal{C} \preceq \mathcal{I}_N$. Let $\mathcal{Q} = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_N$ be the Cartesian product of $\mathcal{Q}_i$ for $i = 1, \cdots, N$. We express $\mathrm{Prox}_g^\gamma(x) := \arg\min_y g(y) + \frac{1}{2\gamma}\|y - x\|^2$ as the proximal mapping of $g(y)$ specified by the parameter $\gamma > 0$. Let $\mathrm{Diam}(\mathcal{Q}) := \max_{x\in\mathcal{Q}} \|x\|$ be the diameter of set $\mathcal{Q}$, and $\mathcal{C}^{\frac{1}{2}}\mathcal{Q} := \{\mathcal{C}^{1/2}u \mid u \in \mathcal{Q}\}$.

Proximal Average Function and Its Dual Formulation

In this section, we first present the definition of one proximal average function and then give its equivalent dual formulation which is vital for us to establish the main results. In addition, more properties and applications about proximal average functions can be found in (Bauschke et al. 2008; Hare 2009; Yu et al. 2015; Zhong and Kwok 2014a; 2014b).

Definition 1 *(Bauschke et al. 2008, Definition 4.1) The proximal average function of $g(x) = \sum_{i=1}^N \alpha_i g_i(x)$ with parameter $\gamma > 0$ is defined via the following optimization problem*

$$g_\gamma(x) := \inf_{y_i} \Big\{ \sum_{i=1}^N \alpha_i g_i(y_i) + \frac{1}{2\gamma}\sum_{i=1}^N \alpha_i\|y_i\|^2 - \frac{1}{2\gamma}\|x\|^2 : \sum_{i=1}^N \alpha_i y_i = x \Big\}. \tag{3}$$

The following proposition states a key property of the proximal mapping $\mathrm{Prox}_{g_\gamma}^\gamma(x)$ in (Bauschke et al. 2008).

Proposition 1 *(Bauschke et al. 2008, Theorem 6.7)* $\mathrm{Prox}_{g_\gamma}^\gamma(x) = \alpha_1 \mathrm{Prox}_{g_1}^\gamma(x) + \cdots + \alpha_N \mathrm{Prox}_{g_N}^\gamma(x)$.

The following dual formulation plays a critical role in establishing the main results of this paper. We first claim that $\mathcal{Q}_i$ is a compact convex set for all $i = 1, \cdots, N$. The details can be found in the proof of Lemma 1 in the supplemental material.

Lemma 1 *Suppose that $\alpha_i \geq 0$ satisfying $\sum_{i=1}^N \alpha_i = 1$ and $g_\gamma(x)$ is the proximal average of $g(x)$ with approximation parameter $\gamma > 0$. Then, for any $\gamma, \gamma_1, \gamma_2 > 0$, the following statements hold:*

$$g_\gamma(x) = \sup_{u\in\mathcal{Q}} \Big\{ \sum_{i=1}^N \alpha_i\langle u_i, x\rangle - \sum_{i=1}^N \alpha_i g_i^*(u_i) - \frac{\gamma}{2}\|u\|_\mathcal{C}^2 \Big\}, \tag{4}$$

$$g_{\gamma_1}(x) \leq g_{\gamma_2}(x) + \frac{\gamma_2 - \gamma_1}{2}\|u_{\gamma_1}^*(x)\|_\mathcal{C}^2, \tag{5}$$

$$\frac{\gamma_1}{2}\|u_{\gamma_1}^*(x)\|_\mathcal{C}^2 + g_{\gamma_1}(x) \leq g(x) \leq g_{\gamma_2}(x) + \frac{\gamma_2}{2}\|u_0^*(x)\|_\mathcal{C}^2, \tag{6}$$

where $u_\gamma^*(x) \in \mathcal{U}_\gamma^*(x)$ and $\mathcal{U}_\gamma^*(x)$ denotes the optimal solution set of (4) *with some parameter $\gamma \geq 0$. Moreover, $\|u_\gamma^*(x)\|_\mathcal{C}^2$ is constant over the set $\mathcal{U}_\gamma^*(x)$ for any $\gamma > 0$.*

Remark 1 **(1)** *The proximal average function $g_\gamma(x)$ is a nonsmooth lower semicontinuous convex function (Bauschke et al. 2008, Corollary 5.2). When $N = 1$, matrix $\mathcal{C} = 0$ and $g(x) = g_\gamma(x)$.*

(2) *One can easily find that the proximal average approximation is tighter than the moreau envelope smoothing function according to the dual formulation* (4) *and the fact $0 \preceq \mathcal{C} \preceq \mathcal{I}_N$.*

The following lemma ensures a much stronger theoretical property of $g_\gamma(x)$ than the general convexity proven in (Bauschke et al. 2008, Corollary 5.2). This lower bound estimation of the proximal average function $g_\gamma(x)$ is crucial for us to establish the $\mathcal{O}(\frac{1}{k})$ iteration complexity of the proposed APA-APG algorithm. Due to the page limit, we defer the proof of Lemma 2 to the supplemental material.

Lemma 2 *Let $g_\gamma(x)$ be the function defined by* (4)*. Then, it holds that $\partial g_\gamma(y) = \mathcal{A}\mathcal{U}^*_\gamma(y)$ and for any $\xi \in \partial g_\gamma(y)$,*

$$g_\gamma(x) \geq g_\gamma(y) + \langle \xi, x - y\rangle + \frac{\gamma}{2}\|u^*_\gamma(x) - u^*_\gamma(y)\|^2_{\mathcal{C}}. \quad (7)$$

Adaptive Proximal Average Approximation Method

By virtue of the basic property of proximal average functions, we solve the following approximated convex minimization (8) to obtain an appropriate approximate solution to the initial problem (1):

$$\min_x \; F_\gamma(x) = f(x) + g_\gamma(x). \quad (8)$$

In what follows, we integrate the accelerated proximal gradient and adaptive proximal average approximation techniques to tackle problem (8), establishing the upper bound estimation of $F(x^{k+1}) - F^*$. We name the proposed adaptive proximal average approximation method as **A**daptive **P**roximal **A**verage based **A**ccelerated **P**roximal **G**radient (APA-APG), and also present two variants of the APA-APG algorithm, which we denote as APA-APG1 (see Algorithm 1) and APA-APG2 (see Algorithm 2), respectively.

Algorithm 1 APA-APG1 Algorithm

```
Parameters: Choose $\gamma_1 > 0, a > 0$, and an initial point $x_0$. Let $\widehat{x}_0 = x_0$.
for $k = 0, 1, \cdots$ do
    Set $\tau_k = \frac{1}{k+a}$ and $\gamma_{k+1} = \min(\frac{\gamma_1 a}{k+a}, \frac{1}{L_f})$;
    $\widehat{x}^k := (1 - \tau_k)x^k + \tau_k \widetilde{x}^k$;
    $x^{k+1} := \sum_{i=1}^N \alpha_i \mathrm{Prox}^{\gamma_{k+1}}_{g_{\gamma_{k+1}}}(\widehat{x}^k - \gamma_{k+1}\nabla f(\widehat{x}^k))$;
    $\widetilde{x}_{k+1} := \widetilde{x}_k + \tau_k^{-1}(x^{k+1} - \widehat{x}^k)$;
end for
```

Algorithm 2 APA-APG2 Algorithm

```
Parameters: Choose $\gamma_1 > 0, a > 0$, and an initial point $x_0$. Let $\widehat{x}_0 = x_0$.
for $k = 0, 1, \cdots$ do
    Set $\tau_k = \frac{1}{k+a}$ and $\gamma_{k+1} = \min(\frac{\gamma_1 a}{k+a}, \frac{1}{L_f})$;
    $\widehat{x}^k := (1 - \tau_k)x^k + \tau_k \widetilde{x}^k$;
    $x^{k+1} := \sum_{i=1}^N \alpha_i \mathrm{Prox}^{\gamma_{k+1}}_{g_{\gamma_{k+1}}}(\widehat{x}^k - \gamma_{k+1}\nabla f(\widehat{x}^k))$;
    $\widetilde{x}_{k+1} := \widetilde{x}_k + \tau_k^{-1}(2 - \gamma_{k+1}L_f)(x^{k+1} - \widehat{x}^k)$;
end for
```

Remark 2 **(1)** *Notice that the step of updating x^{k+1} in Algorithms 1 and 2 can be equivalently formulated as the following convex minimization problem according to Proposition 1*

$$x^{k+1} = \arg\min_x \; g_{\gamma_{k+1}}(x) + \frac{1}{2\gamma_{k+1}}\|x - G^k\|^2, \quad (9)$$

where $G^k = \big(\widehat{x}^k - \gamma_{k+1}\nabla f(\widehat{x}^k)\big)$.
(2) *In the above two proximal average approximation algorithms, the adaptive stepsize $\gamma_{k+1} = \min(\frac{1}{k+a}, \frac{1}{L_f})$ is used to make these algorithms much more practical and efficient. Also, $\gamma_{k+1} \to 0$ as $k \to \infty$, which implies that the APA-APG has a better approximation than PA-APG, where the approximation parameter $\gamma = \mathcal{O}(\min(\frac{1}{L_f}, \epsilon))$ is a pregiven constant.*
(3) *The difference between the two proposed algorithms only exists in the updating step on $\widetilde{x}_{k+1}$. In details, the updating step of $\widetilde{x}_{k+1}$ in APA-APG2 enjoys a larger weight on $(x^{k+1} - \widehat{x}^k)$. Moreover, the iteration complexity of APA-APG2 is easy to establish by adopting the same proof technique as APA-APG1. To avoid redundancy, we only give the detailed proof for APA-APG1.*

Lemma 3 *Let τ_k, γ_k be the parameters appearing in Algorithms 1. Then, the statement holds that*

$$(1 - \tau_k)\frac{\gamma_{k+1}}{\tau_k^2} = \frac{\gamma_k}{\tau_{k-1}^2},$$

$$-\sum_{k=1}^{\infty} \frac{(1 - \tau_k)(\gamma_{k+1} - \gamma_k + \gamma_{k+1}\tau_k)}{2}\frac{\gamma_{k+1}}{\tau_k^2} \leq \gamma_1^2 a.$$

Iteration Complexity Analysis

The following lemma gives some useful inequalities. For notational convenience, we define

$$\mathcal{Q}^k_\gamma(x) := f(\widehat{x}^k) + \langle \nabla f(\widehat{x}^k), x - \widehat{x}^k\rangle + g_\gamma(x). \quad (10)$$

Lemma 4 *For any $x \in \operatorname{dom} F_{\gamma_{k+1}}$, the point pair $\{x^k, \widehat{x}^k\}$ generated by Algorithm 1 and all $k \geq 1$, the following sandwich relation holds*

$$\begin{aligned} F_{\gamma_{k+1}}(x) \geq \mathcal{Q}^k_{\gamma_{k+1}}(x) \geq F_{\gamma_{k+1}}(x^{k+1}) &- \frac{L_f}{2}\|x^{k+1} - \widehat{x}^k\|^2 \\ &- (\gamma_{k+1})^{-1}\langle x^{k+1} - \widehat{x}^k, x - x^{k+1}\rangle \\ &+ \frac{\gamma_{k+1}}{2}\|u^*_{\gamma_{k+1}}(x^{k+1}) - u^*_{\gamma_{k+1}}(x)\|^2_{\mathcal{C}}. \end{aligned} \quad (11)$$

Remark 3 *This type of sandwich inequalities is commonly used for estimating the iteration complexities of accelerated proximal gradient (FISTA) algorithms (Beck and Teboulle 2009b; 2009a). Moreover, this bound is much tighter since we have explored the specific structure of the proximal average function $g_\gamma(x)$.*

The following two lemmas give an upper bound of $\Lambda^{k+1}(x, \gamma, \tau)$ and $\Gamma^{k+1}(x, \gamma, \tau)$, respectively (see their proofs in the supplemental material).

Lemma 5 *Let $\{x^k, \widehat{x}^k, \widetilde{x}^k\}$ be the sequence generated by Algorithm 1. For any $x \in \operatorname{Dom} F$, the following upper bound estimation for $\Lambda^{k+1}(x, \gamma_{k+1}, \tau_k)$ holds*

$$\Lambda^{k+1}(x, \gamma_{k+1}, \tau_k) \leq \frac{\tau_k^2}{2\gamma_{k+1}}(\|x - \widetilde{x}^k\|^2 - \|x - \widetilde{x}^{k+1}\|^2),$$

where $\Lambda^{k+1}(x,\gamma,\tau) = \frac{\tau}{\gamma}\langle x^{k+1}-\widehat{x}^k, x-\widehat{x}^k\rangle + \frac{1-\tau}{\gamma}\langle x^{k+1}-\widehat{x}^k, x^k-\widehat{x}^k\rangle - \frac{1-\gamma L_f}{2\gamma}\|x^{k+1}-\widehat{x}^k\|^2$.

Lemma 6 *Let x^* be an optimal solution of problem* (1) *and $\{x^{k+1}\}$ be the sequence generated by Algorithm 1 or 2. Then, $\Gamma^{k+1}(x^*,\gamma_{k+1},\tau_k)$ takes the lower bound estimation as follows*

$$\begin{aligned}&\Gamma^{k+1}(x^*,\gamma_{k+1},\tau_k)\\ &\geq \frac{(1-\tau_k)(\gamma_{k+1}-\gamma_k+\gamma_{k+1}\tau_k)}{2}\|u^*_{\gamma_{k+1}}(x^*)\|^2_{\mathcal{C}},\end{aligned} \quad (12)$$

where $\Gamma^{k+1}(x,\gamma,\tau) = \frac{\gamma(1-\tau)}{2}\|u^*_\gamma(x^{k+1})-u^*_\gamma(x)\|^2_{\mathcal{C}} + \frac{\gamma\tau}{2}\|u^*_\gamma(x)\|^2_{\mathcal{C}} - \frac{(\gamma_k-\gamma)(1-\tau)}{2}\|u^*_\gamma(x^{k+1})\|^2_{\mathcal{C}}$.

In the following lemma, we give an upper bound estimation for $F_{\gamma_{k+1}}(x^{k+1})-F^*$ (see its proof in the supplemental material).

Lemma 7 *Let $\{x^k\}$ be the sequence generated by Algorithm 1. It holds that*

$$\begin{aligned}F_{\gamma_{k+1}}(x^{k+1})-F^* \leq \frac{\tau_k^2}{\gamma_{k+1}}\Big\{&\frac{\gamma_1}{\tau_0^2}\big(F_{\gamma_1}(x^1)-F(x^*)\big)\\ &+\frac{1}{2}\|x^*-\widetilde{x}^1\|^2+(\gamma_1)^2 a\|\mathrm{Diam}_{\mathcal{C}^{\frac{1}{2}}\mathcal{Q}}\|^2\Big\}.\end{aligned}$$

Now, we are on the position to establish the $\mathcal{O}(\frac{1}{k})$ upper bound estimation for $F(x^{k+1})-F^*$. Based on (Nesterov 2005a; 2005b), the theorem below indicates that the proposed algorithm enjoys the $\mathcal{O}(\frac{1}{k})$ iteration complexity.

Theorem 1 *Let $\{x^k\}$ be the sequence generated by Algorithm 1 or 2. It holds that*

$$F(x^{k+1})-F^* \leq \mathcal{O}(\frac{1}{k}).$$

Experiments

In this section, we perform some experiments on two important problems in machine learning: overlapping group Lasso and graph-guided fused Lasso to verify the efficacy of our proposed APA-APG1 and APA-APG2 algorithms. Since the S-APG algorithm has been shown of less efficacy compared with PA-APG in (Nesterov 2005b), we only need to compare our proposed algorithms APA-APG1 and APA-APG2 with the state-of-the-art solver PA-APG. To be fair, all the compared algorithms start with the same initial points.

Overlapping Group Lasso

We first conduct the experiments on the overlapping group Lasso optimization problem (Zhao, Rocha, and Yu 2009; Jacob, Obozinski, and Vert 2009; Mairal et al. 2010)

$$\min \frac{1}{2\lambda K}\|Ax-b\|^2 + \sum_{i=1}^{K}\alpha_i\|x_{\mathcal{G}_i}\|,$$

where $A \in \mathbb{R}^{n\times d}$ is the sampling matrix in which the entries are sampled with $i.i.d.$ normal distribution; $x_j^* = (-1)^j \exp^{-(j-1)/100}$; $b = Ax+\xi$ in which ξ is the noise sampled from the zero mean and unit variance normal distribution; $\mathcal{G} = \{\mathcal{G}_1,\cdots,\mathcal{G}_K\}$ denotes the set of groups each of which is a subset of $\{1,\cdots,d\}$; $x_{\mathcal{G}_i} \in \mathbb{R}^d$ is a copy of x with $x_{\{1,\cdots,d\}\setminus\mathcal{G}_i} = 0$; λ is the regularization parameter for structured sparsity Moreover, the groups are defined as

$$\{\{1,\cdots,100\},\{91,\cdots,190\},\cdots,\{d-99,\cdots,d\}\}$$

with $d = 90K+10$. It means that in each group there exist ten overlapped components. Similar to (Yu 2013), we adopt the uniform weight $\alpha_i = \frac{1}{K}$, and set $\lambda = \frac{K}{5}$.

In the experiments, the sampling dimension n is fixed to $n = 4000$ and the numbers of groups are set to $K = 10$, $K = 20$, and $K = 40$, respectively. Figures 1-3 show the performance of PA-APG, APA-APG1, and APA-APG2 on the overlapping group Lasso problem with increasing precision parameters $\epsilon = 1.0e^{-4}$, $\epsilon = 1.0e^{-5}$, and $\epsilon = 1.0e^{-6}$, respectively. It is obvious that PA-APG is much sensitive with the number of the nonsmooth term K and required precision. Hence, PA-APG is not suitable for the problem with too many nonsmooth terms and highly required precision. Clearly, APA-APG1 and APA-APG2 are much faster than PA-APG during the first 1000 iterations. Table 1 lists the number of iterations needed by each solver under the same terminal condition, which implies that APA-APG1 and APA-APG2 exhibit great superiority over PA-APG and can work for high-precision required problems.

Graph-Guided Fused Lasso

Here, we perform the experiments on the graph-guided fused Lasso optimization problem (Chen et al. 2012; Kim and Xing 2009)

$$\min \frac{1}{2\lambda K}\|Ax-b\|^2 + \sum_{(i,j)\in E}\alpha_{ij}\|x_i-x_j\|$$

with $\alpha_{ij} \geq 0$ for all $(i,i) \in E$ and $\sum_{(i,j)\in E}\alpha_{ij} = 1$, where E is the graph edge set constructed by thresholding the correlation matrix. In the experiments, we fix $n = 4000$ and vary the dimension d from 500 to 3000 to construct different edge sets E. In the following experiments, we test the graph-guided fused Lasso problem with $d = 500, 1000, 2000$.

Figures 4-6 show the performance of PA-APG, APA-APG1, and APA-APG2 on the graph-guided fused Lasso problem with three types of precision parameters $\epsilon = 1.0e^{-4}$, $\epsilon = 1.0e^{-5}$, and $\epsilon = 1.0e^{-6}$, respectively. Apparently, the PA-APG algorithm proposed in (Yu 2013) is also less efficient than the proposed adaptive proximal average approximation algorithms APA-APG1 and APA-APG2. In addition, APA-APG2 is shown slightly better than APA-APG1 on the graph-guided fused Lasso problem from the shown performance in Figures 4-6. Moreover, according to the number of iterations shown in Table 2, we know that the PA-APG algorithm almost does not work for problems with $1.0e^{-6}$ highly required precision and larger number of nonsmooth terms. The performance on the graph-guided fused Lasso problem also implies great superiority of the proposed algorithms APA-APG1 and APA-APG2.

Table 1: Performance of the number of iterations for overlapping group Lasso.

Method	$K = 10$	$K = 20$	$K = 40$
Error	1e-4|1e-5|1e-6	1e-4|1e-5|1e-6	1e-4|1e-5|1e-6
PA-APG	124 | 710 | 1524	370 | 1456 | 5765	1531 | 5377 | 18087
APA-APG1	25 | 41 | 41	67 | 73 | 76	331 | 457 | 653
APA-APG2	25 | 41 | 41	67 | 73 | 76	261 | 335 | 1031

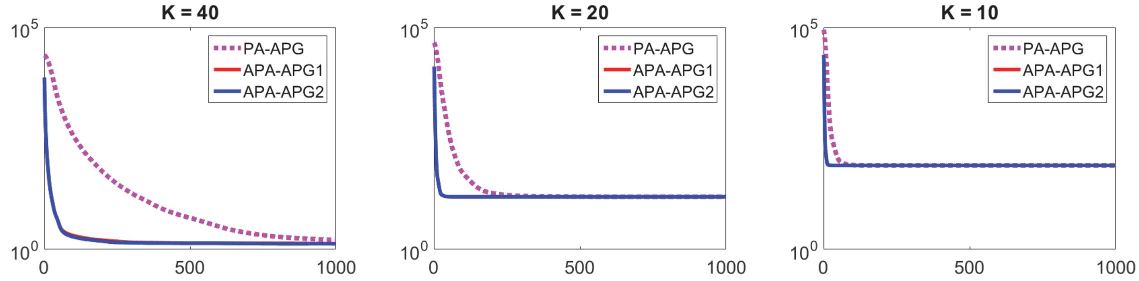

Figure 1: Objective value vs. iteration on overlapping group Lasso with accuracy $1.0e^{-4}$

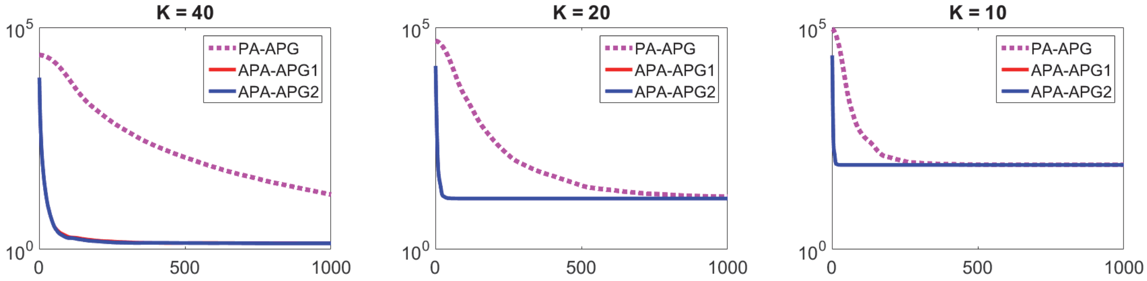

Figure 2: Objective value vs. iteration on overlapping group Lasso with accuracy $1.0e^{-5}$

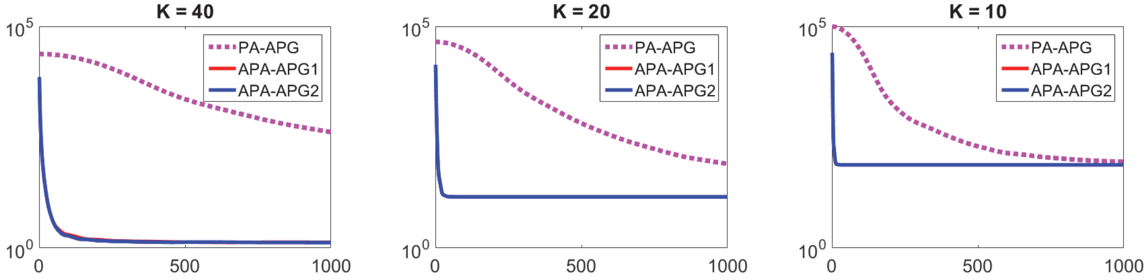

Figure 3: Objective value vs. iteration on overlapping group Lasso with accuracy $1.0e^{-6}$

Conclusions

In this paper, we proposed an adaptive proximal average approximation approach for multi-term nonsmooth composite convex minimization by employing the proximal average approximation and adaptive parameter tuning techniques. For the proximal average approximation function, we first gave its equivalent dual formulation and then established a tighter lower bound estimation for it. By exploiting the new structure of the proximal average approximation function, we achieved the $\mathcal{O}(\frac{1}{k})$ iteration complexity for the proposed adaptive proximal average approximation algorithm with an appropriate choice of the approximation parameter without increasing any extra computational cost. Moreover, we conducted the experiments on overlapping group Lasso and graph-guided fused Lasso problems, verifying the efficacy of the proposed adaptive proximal average approximation method. Specifically, we compared this method against the state-of-the-art method PA-APG (Yu 2013) which has been demonstrated much faster than the S-APG method (Nesterov 2005b). The experimental results showed that our proposed method exhibits great superiority to resolve multi-term com-

Table 2: Performance of the number of iterations for graph-guided fused Lasso.

Method	$d = 500$			$d = 1000$			$d = 2000$		
Error	1e-4	1e-5	1e-6	1e-4	1e-5	1e-6	1e-4	1e-5	1e-6
PA-APG	1426	7297	14624	2267	9834	44364	3426	13785	50472
APA-APG1	177	121	901	654	1820	1764	402 1074	4758	9341
APA-APG2	100	88	465	733	957	889	712	3527	4878

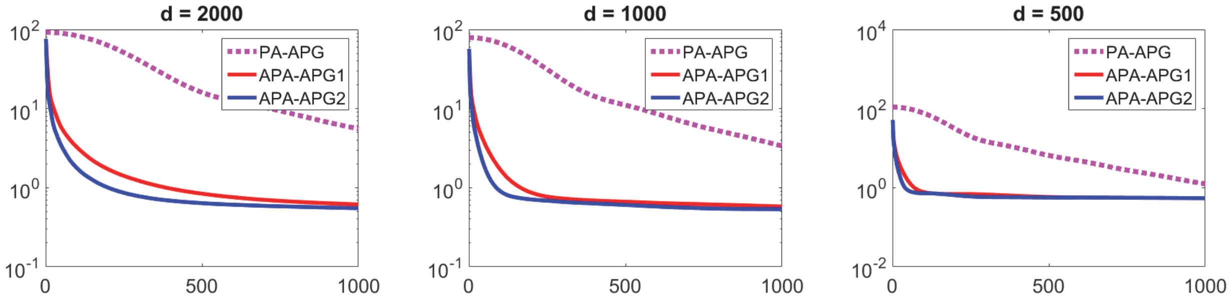

Figure 4: Objective value vs. iteration on graph-guided fused Lasso with accuracy $1.0e^{-4}$

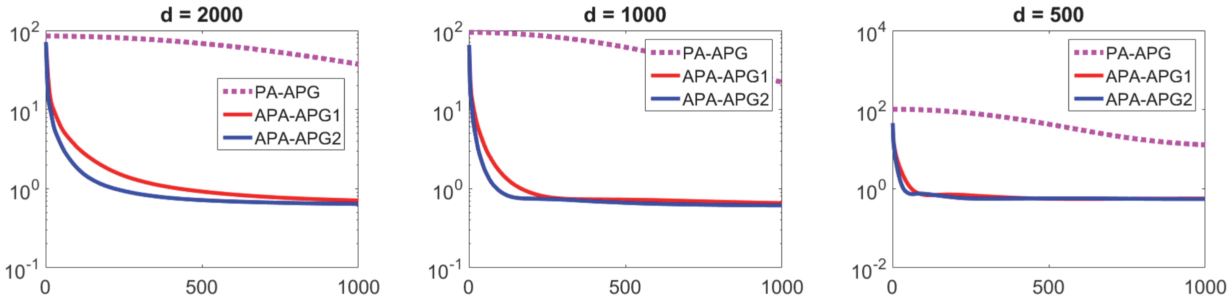

Figure 5: Objective value vs. iteration on graph-guided fused Lasso with accuracy $1.0e^{-5}$

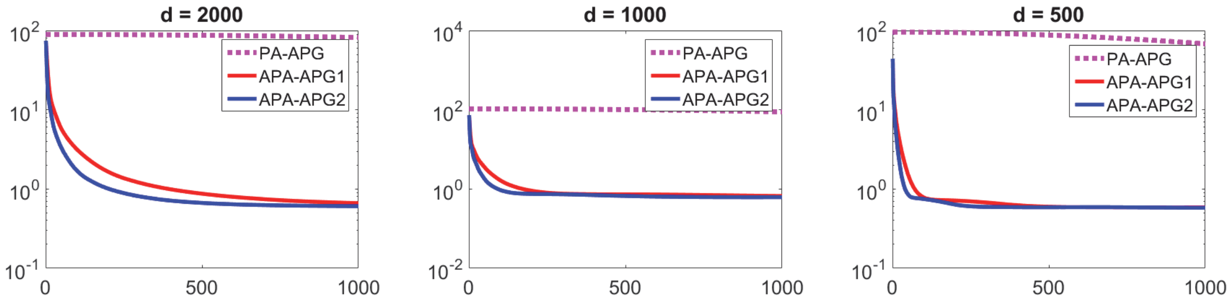

Figure 6: Objective value vs. iteration on graph-guided fused Lasso with accuracy $1.0e^{-6}$

posite convex minimization and also works well for higher-precision required optimization.

Acknowledgments

The authors would like to thank Yaoliang Yu from University of Waterloo for sharing us the code. Shiqian Ma is partially supported by the Hong Kong Research Grants Council General Research Fund (Grant No. 14205314). Li Shen is partially supported by the Open Project Program of the State Key Lab of CAD&CG (Grant No. A1613), Zhejiang University. Junzhou Huang is partially supported by U.S. NSF IIS-1423056, CMMI-1434401, CNS-1405985 and the NSF CAREER grant IIS-1553687.

References

Bach, F.; Jenatton, R.; Mairal, J.; Obozinski, G.; et al. 2011. Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning* 5.

Bach, F.; Jenatton, R.; Mairal, J.; Obozinski, G.; et al. 2012. Structured sparsity through convex optimization. *Statistical Science* 27(4):450–468.

Bauschke, H. H.; Goebel, R.; Lucet, Y.; and Wang, X. 2008. The proximal average: basic theory. *SIAM Journal on Optimization* 19(2):766–785.

Beck, A., and Teboulle, M. 2009a. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *Image Processing, IEEE Transactions on* 18(11):2419–2434.

Beck, A., and Teboulle, M. 2009b. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* 2(1):183–202.

Bertsekas, D. P. 1999. Nonlinear programming.

Borwein, J. M., and Vanderwerff, J. D. 2010. *Convex functions: constructions, characterizations and counterexamples*, volume 109. Cambridge University Press Cambridge.

Boţ, R. I., and Hendrich, C. 2015. A variable smoothing algorithm for solving convex optimization problems. *TOP* 23(1):124–150.

Chen, X.; Lin, Q.; Kim, S.; Carbonell, J. G.; Xing, E. P.; et al. 2012. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics* 6(2):719–752.

Hare, W. L. 2009. A proximal average for nonconvex functions: a proximal stability perspective. *SIAM Journal on Optimization* 20(2):650–666.

Jacob, L.; Obozinski, G.; and Vert, J. P. 2009. Group lasso with overlap and graph lasso. In *International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June*, 433–440.

Kim, S., and Xing, E. P. 2009. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genet* 5(8):e1000587.

Mairal, J.; Jenatton, R.; Bach, F. R.; and Obozinski, G. R. 2010. Network flow algorithms for structured sparsity. In *Advances in Neural Information Processing Systems*, 1558–1566.

Nemirovsky, A.-S.; Yudin, D.-B.; and Dawson, E.-R. 1982. Problem complexity and method efficiency in optimization.

Nesterov, Y. 1983. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, volume 27, 372–376.

Nesterov, Y. 2005a. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization* 16(1):235–249.

Nesterov, Y. 2005b. Smooth minimization of non-smooth functions. *Mathematical programming* 103(1):127–152.

Nesterov, Y. 2013a. Gradient methods for minimizing composite functions. *Mathematical Programming* 140(1):125–161.

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

Ouyang, H.; He, N.; Tran, L.; and Gray, A. 2013. Stochastic alternating direction method of multipliers. In *Proceedings of the 30th International Conference on Machine Learning*, 80–88.

Polyak, B. 1977. Subgradient methods: A survey of soviet research. In *Nonsmooth optimization: Proceedings of the IIASA workshop March*, 5–30.

Rockafellar, R. T. 2015. *Convex analysis*. Princeton university press.

Teo, C. H.; Vishwanthan, S.; Smola, A. J.; and Le, Q. V. 2010. Bundle methods for regularized risk minimization. *The Journal of Machine Learning Research* 11:311–365.

Tran-Dinh, Q. 2015. Adaptive smoothing algorithms for nonsmooth composite convex minimization. *arXiv preprint arXiv:1509.00106*.

Yu, Y.; Zheng, X.; Marchetti-Bowick, M.; and Xing, E. P. 2015. Minimizing nonconvex non-separable functions. In *AISTATS*.

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

Zhao, P.; Rocha, G.; and Yu, B. 2009. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics* 3468–3497.

Zhong, W., and Kwok, J. T. 2014a. Gradient descent with proximal average for nonconvex and composite regularization. In *AAAI*, 2206–2212.

Zhong, W., and Kwok, J. T.-Y. 2014b. Accelerated stochastic gradient method for composite regularization. In *AISTATS*, 1086–1094.