Neurocomputing 273 (2018) 516–525

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Stochastic Primal-Dual Proximal ExtraGradient descent for compositely regularized optimization

Tianyi Lin[a], Linbo Qiao[b,*], Teng Zhang[c], Jiashi Feng[d], Bofeng Zhang[b]

[a] Department of Industrial Engineering and Operations Research, UC Berkeley, USA
[b] College of Computer, National University of Defense Technology, Changsha, China
[c] Department of Management Science and Engineering, Stanford University, USA
[d] Department of ECE, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 30 December 2016
Revised 12 May 2017
Accepted 17 July 2017
Available online 24 August 2017

Communicated by Tianbao Yang

Keywords:
Compositely regularized optimization
Stochastic Primal-Dual Proximal ExtraGradient descent

ABSTRACT

We consider a wide range of regularized stochastic minimization problems with two regularization terms, one of which is composed with a linear function. This optimization model abstracts a number of important applications in artificial intelligence and machine learning, such as fused Lasso, fused logistic regression, and a class of graph-guided regularized minimization. The computational challenges of this model are in two folds. On one hand, the closed-form solution of the proximal mapping associated with the composed regularization term or the expected objective function is not available. On the other hand, the calculation of the full gradient of the expectation in the objective is very expensive when the number of input data samples is considerably large. To address these issues, we propose a stochastic variant of extra-gradient type methods, namely Stochastic Primal-Dual Proximal ExtraGradient descent (SPDPEG), and analyze its convergence property for both convex and strongly convex objectives. For general convex objectives, the uniformly average iterates generated by SPDPEG converge in expectation with $O(1/\sqrt{t})$ rate. While for strongly convex objectives, the uniformly and non-uniformly average iterates generated by SPDPEG converge with $O(\log(t)/t)$ and $O(1/t)$ rates, respectively. The order of the rate of the proposed algorithm is known to match the best convergence rate for first-order stochastic algorithms. Experiments on fused logistic regression and graph-guided regularized logistic regression problems show that the proposed algorithm performs very efficiently and consistently outperforms other competing algorithms.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

In this paper, we are interested in solving a class of convex optimization problems with both non-composite and composite regularization terms:

$$\min_{x\in\mathcal{X}} \ \mathbb{E}_\xi[l(x,\xi)] + r_1(x) + r_2(Fx), \tag{1}$$

where $\mathcal{X}\subset\mathbb{R}^d$ is a convex compact set with diameter D_x, the regularization terms $r_1:\mathbb{R}^d\to\mathbb{R}$ and $r_2:\mathbb{R}^l\to\mathbb{R}$ are both convex but possibly nonsmooth, and r_2 is composed with a possibly non-diagonal penalty matrix $F\in\mathbb{R}^{l\times d}$ specifying the desired structured sparsity pattern in x. We denote $l(\cdot,\cdot):\mathbb{R}^d\times\Omega\to\mathbb{R}$ as a convex and smooth loss function of a rule x for a sample data $\{\xi_i=(a_i,b_i)\}$, and define the corresponding expectation as $l(x)=\mathbb{E}_\xi[l(x,\xi)]$.

The above formulation covers quite a few popular models arising from statistics and machine learning, such as Lasso [1] obtained by setting $l(x,\xi_i)=\frac{1}{2}\left\|a_i^\top x-b_i\right\|^2$ and $r_1(x)=\lambda\|x\|_1$ and $r_2=0$, and linear SVM [2] obtained by letting $l(x,\xi_i)=\max\left(0,1-b_i\cdot a_i^\top x\right)$ and $r_1(x)=(\lambda/2)\|x\|_2^2$ and $r_2=0$, where $\lambda>0$ is a parameter. More importantly, we can accommodate problem (1) with more complicated structures by imposing the non-trivial regularization term $r_2(Fx)$, such as fused Lasso [3], fused logistic regression and graph-guided regularized minimization [4].

The standard algorithm applied to solve problem (1) is proximal gradient descent [5]. However, there are two main difficulties: (1) computing the exact proximal gradient is intractable since the closed-form solution to the proximal mapping of $r_1(x)+r_2(Fx)$, or even single $r_2(Fx)$ is in usually unavailable; (2) the computational complexity of the full gradient $\nabla l(x)$ rapidly increases as the size of samples grows, and is hence prohibitively expensive for modern data-intensive applications.

A common way to suppress the former one is to introduce a new auxiliary variable z with $z=Fx$ and reformulate problem

* Corresponding author.
E-mail address: qiao.linbo@nudt.edu.cn (L. Qiao).

http://dx.doi.org/10.1016/j.neucom.2017.07.066
0925-2312/© 2017 Elsevier B.V. All rights reserved.

(1) as a linearly constrained convex problem with respect to two variables x and z as follows:

$$\min_{x\in\mathcal{X}} \mathbb{E}_\xi[l(x,\xi)] + r_1(x) + r_2(z), \quad \text{s.t.} \quad Fx - z = 0. \tag{2}$$

Then one can resort to Linearized Alternating Direction Method of Multipliers (LADMM) [6,7]. Very recently, Lin et al. [8] have explored the efficiency of the extra-gradient descent [9,10], and further showed the hybrid Extra-Gradient ADM (EGADM) is very efficient on moderate size problems. However, these methods are computationally expensive due to the computation of the full gradient in each iteration.

To address the computational issue, several stochastic ADMM algorithms [11–15] have been proposed. The idea is to draw a mini-batch of samples and then compute a noisy sub-gradient of $l(x) + r_1(x)$ on the mini-batch in each iteration. However, for problem (1) with non-smooth regularization (which is actually common in practice), these sub-gradient type alternating direction methods may be slow and unstable [16].

In this work, we propose a Stochastic Primal-Dual Proximal Extra-Gradient Descent (SPDPEG), which inherits the advantages of EGADM and stochastic methods. Basically, the proposed method computes two noisy gradients of l at the kth iteration by randomly drawing two data samples ξ_1^{k+1} and ξ_2^{k+1}, and then performs extra-gradient descent along the noisy gradients. We demonstrate that the proposed algorithm is very efficient and stable in solving problem (1) with possible non-smooth terms at large scale.

Our contribution: We propose a novel Stochastic Primal-Dual Proximal Extra-Gradient Descent (SPDPEG). SPDPEG is efficient in solving large-scale problems with composite and nonsmooth regularizations. We demonstrate its theoretical convergence for both convex and strongly convex objectives. For convex objectives, SPDPEG has the convergence rate of $O(1/\sqrt{t})$ in expectation with the uniformly average iterates. This convergence rate is known to be the best possible for minimizing general convex objective using first-order noisy oracle [17]. When the objective to be optimized is strongly convex, SPDPEG converges at the rates of $O(\log(t)/t)$ and $O(1/t)$ in expectation with the uniformly and non-uniformly average iterates, respectively. This matches the convergence rate of stochastic ADMM with a significantly stronger robustness in terms of the numerical performance, as confirmed by encouraging experiments on fused logistic regression and graph-guided regularized minimization tasks.

Related work: The first line of related work are various stochastic alternating direction methods [11–15,18–20] developed to solve problem (2). They fall into two camps: (1) to compute the noisy sub-gradient of $l + r_1$ on a mini-batch of data samples and perform sub-gradient descent [11–14,18]; (2) to approximate problem (2) using the finite-sum loss and perform variance-reduced gradient descent or dual coordinate ascent [15,19,20].

For the first group of algorithms, drawing a noisy sub-gradient may lead to the unstable numerical performance, especially on large-scale problems. In the experimental section, we compare our algorithm against SGADM [18] and demonstrate the significant improvement.

For the second group of algorithms, it is not always feasible to use the finite-sum loss since we know nothing about the underlying distribution of data. In specific, Zhong and Kwok [19] proposed a Stochastic Averaged Gradient-based ADM (SAG-ADM) whose iteration complexity is $O(1/t)$. However, SAG-ADM needs to store a few variables and incurs a very high memory cost. Suzuki [20] proposed a linearly convergent Stochastic Dual Coordinate Ascent ADM (SDCA-ADM). However, a stronger assumption on r_1 and r_2 such as strong convexity and smoothness is imposed. Zheng and Kwok [15] proposed a Stochastic Variance-Reduce Gradient-based ADM (SVRG-ADM) for convex and non-convex problems. However, SVRG-ADM only focuses on the finite-sum problem. In contrast, our SPDPEG approach can be applied to solve problem (2) in very general form.

Very recently, a stochastic variant of hybrid gradient method, namely SPDHG [21], has been proposed to solve a class of compositely regularized minimization problems with very special regularization. In specific, $r_1 \equiv 0$ and $r_2(x) = \max_{y\in\mathcal{Y}} \langle y, x\rangle$ (see Assumption 3 in [21]). However, such assumption is very strong and does not hold for many compositely regularized minimization problems. This motivates us to consider problem (2) and develop SPDPEG approach.

The second line of related works is various extra-gradient methods. This idea is not new and originally proposed by Korpelevich for solving saddle-point problems and variational inequalities [9,10]. The convergence and iteration complexity of extra-gradient methods are established in [22] and [23], respectively. There are also some variants of extra-gradient methods. Solodov and Svaiter proposed a hybrid proximal extra-gradient method [24], whose iteration complexity is established by Monteiro and Svaiter in [25–27]. Bonettini and Ruggiero studied a generalized extragradient method for total variation based image restoration problem [28]. To the best of our knowledge, this is the first time that a stochastic primal-dual variant of extra-gradient type methods is introduced to solve problem (1).

2. Problem set-up and methods

We make the following assumptions that are common in optimization literature and usually hold in practice throughout the paper:

Assumption 2.1. The optimal set of problem (1) is nonempty.

Assumption 2.2. $l(\cdot)$ is continuously differentiable with Lipschitz continuous gradient. That is, there exists a constant $L > 0$ such that

$$\|\nabla l(x_1) - \nabla l(x_2)\| \le L\|x_1 - x_2\|, \forall x_1, x_2 \in \mathcal{X}.$$

Assumption 2.2 holds for many problems in machine learning. For example, the following least squares and logistic functions are two standard ones:

$$l(x,\xi_i) = \frac{1}{2}\left\|a_i^\top x - b_i\right\|^2 \quad \text{or} \quad l(x,\xi_i) = \log\left(1 + \exp\left(-b_i \cdot a_i^\top x\right)\right),$$

where $\xi_i = (a_i, b_i)$ is a single data sample.

Assumption 2.3. The regularization functions r_1 and r_2 are both continuous but possibly non-smooth; the associated proximal mapping for each individual regularization admits a closed-form solution, *i.e.*,

$$\text{prox}_{r_i}(x) = \underset{y}{\text{argmin}}\ r_i(y) + \frac{1}{2}\|y - x\|_2^2 \tag{3}$$

can be calculated in a closed form for $i = 1, 2$.

Remark 2.4. We remark that Assumption 2.3 is reasonable for a class of optimization problems regularized by ℓ_1-norm or nuclear norm, such as fused Lasso, fused logistic regression, and graph-guided regularized minimization problems. The proximal mapping of ℓ_1-norm can be computed as follows:

$$\left[\text{prox}_{\|\cdot\|_1}(x)\right]_i = \underset{y}{\text{argmin}}\ \|y\|_1 + \frac{1}{2}\|y - x_i\|_2^2 = \begin{cases} \text{sign}(x_i)(|x_i| - 1) & |x_i| > 1, \\ 0 & |x_i| \le 1. \end{cases}$$

We clarify that the proximal mapping of $r(x)$ and that of $r(Fx)$ are totally different and have different properties. For example, the proximal mapping of $\|x\|_1$ admits a closed-form solution but the proximal mapping of $\|Fx\|_1$ does not admit in general when F is non-diagonal. We only assume that the proximal mapping of $r(x)$ admits a closed-form solution in Assumption 2.3 but expect to address the case of $r(Fx)$ whose proximal mapping does not admit a closed-form solution in general.

Assumption 2.5. The gradient of the objective function $l(x)$ is easy to estimate. Any stochastic gradient estimation $\nabla l(\cdot, \xi)$ for $\nabla l(\cdot)$ at x satisfies

$$\mathbb{E}_\xi[\nabla l(x, \xi)] = \nabla l(x),$$

and

$$\mathbb{E}_\xi\left[\|\nabla l(x, \xi) - \nabla l(x)\|^2\right] \le \sigma^2,$$

where $\sigma > 0$ is a constant number.

Assumption 2.6. $l(\cdot)$ is μ-*strongly convex* at x. In other words, there exists a constant $\mu > 0$ such that

$$l(y) - l(x) - (y - x)^\top \nabla l(x) \ge \frac{\mu}{2}\|y - x\|^2, \forall y \in \mathcal{X}.$$

We remark that our algorithm works even without Assumption 2.6. However, the lower iteration complexity will be obtained with Assumption 2.6.

We introduce the Stochastic Primal-Dual Proximal ExtraGradient (SPDPEG) method, and further discuss the choice of step-size. We define the augmented Lagrangian function for problem (2) as

$$\mathcal{L}_\gamma(z, x, \lambda) = r_2(z) + r_1(x) + \phi(z, x, \lambda) + \frac{\gamma}{2}\|Fx - z\|^2,$$

where $\lambda \in \mathbb{R}^p$ is the dual variable associated with $Fx = z$. ϕ is defined as

$$\phi(z, x, \lambda) = l(x) - \langle \lambda, Fx - z\rangle.$$

The SPDPEG algorithm is based on the *primal-dual update scheme* where (z, x) is primal variable and λ is a dual variable, and can be seen as an inexact augmented Lagrangian method. The details are presented in Algorithm 1.

Algorithm 1 Stochastic Primal-Dual Proximal ExtraGradient (SPDPEG).

Initialize: x^0, z^0, and λ^0.
for $k = 0, 1, 2, \ldots$ do
  choose two data samples ξ_1^{k+1} and ξ_2^{k+1} randomly;
  update z^{k+1} according to Eq. 4;
  update $\bar{x}^{k+1}$ according to Eq. 5;
  update $\bar{\lambda}^{k+1}$ according to Eq. 6;
  update x^{k+1} according to Eq. 7;
  update λ^{k+1} according to Eq. 8;
end for
Output: $\tilde{z}^t = \sum_{k=0}^{t} \alpha^{k+1} z^{k+1}$, $\tilde{x}^t = \sum_{k=0}^{t} \alpha^{k+1} \bar{x}^{k+1}$, and $\tilde{\lambda}^t = \sum_{k=0}^{t} \alpha^{k+1} \bar{\lambda}^{k+1}$.

We provide details on following four important issues: how to solve the primal and dual sub-problems easily, how to apply the noisy gradient and perform extra-gradient descent, how to choose step-size, and how to determine the weights for the non-uniformly average iterates.

1. *Update for z:* The first sub-problem in Algorithm 1 is to minimize the augmented Lagrangian function $\mathcal{L}_\gamma$ with respect to z, *i.e.*,
$$z^{k+1} := \operatorname*{argmin}_z \ \mathcal{L}_\gamma(z, x^k; \lambda^k), \tag{4}$$
which is equivalent to computing the proximal mapping of r_2 and hence admits a closed-form solution from Assumption 2.3.
2. *Stochastic gradient:* According to Assumption 2.5, ϕ is known to be easy for gradient estimation with respect to x, and the stochastic gradient estimation $G(z, x, \lambda; \xi)$ is defined as
$$G(z, x, \lambda; \xi) = \nabla l(x, \xi) - F^\top \lambda.$$
To update x, the SPDPEG algorithm takes a proximal extra-gradient step using a stochastic gradient estimation $G(z, x, \lambda; \xi)$ and different step-sizes, *i.e.*,
$$\bar{x}^{k+1} := \mathrm{prox}_{c^{k+1} r_1}\left(x^k - c^{k+1} G\left(y^{k+1}, x^k, \lambda^k; \xi_1^{k+1}\right)\right), \tag{5}$$
$$\bar{\lambda}^{k+1} := \lambda^k - \gamma\left(Fx^k - z^{k+1}\right), \tag{6}$$
and
$$x^{k+1} := \mathrm{prox}_{c^{k+1} r_1}\left(x^k - c^{k+1} G\left(y^{k+1}, \bar{x}^{k+1}, \bar{\lambda}^{k+1}; \xi_2^{k+1}\right)\right), \tag{7}$$
$$\lambda^{k+1} := \lambda^k - \gamma\left(F\bar{x}^{k+1} - z^{k+1}\right). \tag{8}$$
3. *Step-size c^{k+1}:* The choice of step-size c^{k+1} depends on whether the objective function is strongly convex or not. The rate of convergence varies with respect to different step-size rules. Moreover, a sequence of vanishing step-sizes is necessary since we do not adopt any technique of variance reduction in the proposed algorithm.
4. *Non-uniformly average iterates:* Azadi and Sra [13] showed that non-uniform average iterates generated by stochastic algorithms converge with fewer iterations. Inspired by this work, through non-uniformly averaging the iterates of the SPDPEG algorithm and adopting a slightly modified step-size, we manage to establish an accelerated convergence rate of $O(1/t)$ in expectation.

3. Main result

In this section, we present the main result in this paper. For general convex objectives, the uniformly average iterates generated by the SPDPEG algorithm converge in expectation with $O(1/\sqrt{t})$ rate. While for strongly convex objectives, the uniformly and non-uniformly average iterates generated converge in expectation with $O(\log(t)/t)$ and $O(1/t)$ rates, respectively. **The computational complexity are $O(d/\sqrt{t})$, $O(d\log(t)/t)$ and $O(d/t)$ since the per-iteration complexity is the computational cost of the noisy gradient on ξ_1 and ξ_2 and the proximal mapping**, where d is the dimension of decision variable. The main theoretic results with respect to different settings are summarized as follows:

1. Assuming that l is a general convex objective function, the step-size is $c^{k+1} = \frac{1}{\sqrt{k+1}+\tilde{L}}$, and the weight of the iterates is $\alpha^{k+1} = \frac{1}{t+1}$, the proposed SPDPEG algorithm converges with the $O(1/\sqrt{t})$ rate in expectation.
2. Assuming that l is a μ-strongly convex objective function, the step-size is $c^{k+1} = \frac{2}{\mu(k+1)+2\tilde{L}}$, and the weight of the iterates is $\alpha^{k+1} = \frac{1}{t+1}$, the proposed SPDPEG algorithm converges with the $O(\log(t)/t)$ rate in expectation.
3. Assuming that l is a μ-strongly convex objective function, the step-size is $c^{k+1} = \frac{4}{\mu(k+2)+4\tilde{L}}$, the weight of the iterates is $\alpha^{k+1} = \frac{2(k+3)}{(t+1)(t+6)}$, and the dual variables are bounded by $D_\lambda > 0$ (this assumption is standard and also adopted in [13]), the proposed SPDPEG algorithm converges with the $O(1/t)$ rate in expectation.

In the above, $\tilde{L}$ is defined as

$$\tilde{L} = \max\left\{8\gamma\sigma_{\max}(F^\top F) + \mu, \sqrt{8L^2 + \gamma\sigma_{\max}(F^\top F)} + \mu\right\},$$

where $\sigma_{\max}(F^\top F)$ denotes the largest eigenvalue of $F^\top F$, and $\mu = 0$ when l is a general convex objective function.

We present the main theoretic result for uniformly average iterates under general convex objective functions in the following theorem.

Theorem 3.1. *Consider the SPDPEG algorithm with uniformly average iterates. For any optimal solution (z^*, x^*), it holds that*

$$\left|\mathbb{E}[l(\tilde{x}^t)] + \mathbb{E}[r_1(\tilde{x}^t)] + \mathbb{E}[r_2(\tilde{z}^t)] - l(x^*) - r_1(x^*) - r_2(z^*)\right| = O(1/\sqrt{t}), \tag{9}$$

$$\left\|F\mathbb{E}[\tilde{x}^t] - \mathbb{E}[\tilde{z}^t]\right\| = O(1/\sqrt{t}). \tag{10}$$

Note that this implies that the SPDPEG algorithm converges in expectation with the $O(1/\sqrt{t})$ rate in terms of both the objective error and constraint violation.

We present the main theoretic result for uniformly average iterates under a strongly convex objective function in the following theorem.

Theorem 3.2. *Consider the SPDPEG algorithm with uniformly average iterates. For any optimal solution (z^*, x^*), it holds that*

$$\left|\mathbb{E}[l(\tilde{x}^t)] + \mathbb{E}[r_1(\tilde{x}^t)] + \mathbb{E}[r_2(\tilde{z}^t)] - l(x^*) - r_1(x^*) - r_2(z^*)\right| = O(\log(t)/t), \tag{11}$$

$$\left\|F\mathbb{E}[\tilde{x}^t] - \mathbb{E}[\tilde{z}^t]\right\| = O(\log(t)/t). \tag{12}$$

Note that this implies that the SPDPEG algorithm converges in expectation with the $O(\log(t)/t)$ rate in terms of both the objective error and constraint violation.

We present the main theoretic result for non-uniformly average iterates under a strongly convex objective function in the following theorem.

Theorem 3.3. *Consider the SPDPEG algorithm with non-uniformly average iterates. For any optimal solution (z^*, x^*), it holds that*

$$\left|\mathbb{E}[l(\tilde{x}^t)] + \mathbb{E}[r_1(\tilde{x}^t)] + \mathbb{E}[r_2(\tilde{z}^t)] - l(x^*) - r_1(x^*) - r_2(z^*)\right| = O(1/t), \tag{13}$$

$$\left\|F\mathbb{E}[\tilde{x}^t] - \mathbb{E}[\tilde{z}^t]\right\| = O(1/t). \tag{14}$$

Note that this implies that the SPDPEG algorithm converges in expectation with the $O(1/t)$ rate in terms of both the objective error and constraint violation.

4. Proof

We first prove the key technical lemma which is very important to the proof of Theorems 3.1–3.3.

Lemma 4.1. *The sequence $\left\{z^{k+1}, \bar{x}^{k+1}, \bar{\lambda}^{k+1}, x^{k+1}, \lambda^{k+1}\right\}$ generated by the SPDPEG algorithm satisfies the following inequality:*

$$\begin{aligned}
&r_1(x) + r_2(z) - r_1(\bar{x}^{k+1}) - r_2(z^{k+1}) + \begin{pmatrix} z - z^{k+1} \\ x - \bar{x}^{k+1} \\ \lambda - \bar{\lambda}^{k+1} \end{pmatrix}^\top \\
&\quad\times \begin{pmatrix} \bar{\lambda}^{k+1} \\ G\left(z^{k+1}, \bar{x}^{k+1}, \bar{\lambda}^{k+1}; \xi_2^{k+1}\right) \\ F\bar{x}^{k+1} - z^{k+1} \end{pmatrix} \\
&\geq \frac{1}{2c^{k+1}}\left\|x - x^{k+1}\right\|^2 - \frac{1}{2c^{k+1}}\left\|x - x^k\right\|^2 - 4c^{k+1}\left\|\delta^{k+1}\right\|^2 \\
&\quad - 4c^{k+1}\left\|\bar{\delta}^{k+1}\right\|^2 \\
&\quad - \frac{1}{2\gamma}\left\|\lambda - \lambda^k\right\|^2 + \frac{1}{2\gamma}\left\|\lambda - \lambda^{k+1}\right\|^2 \\
&\quad + \left[\frac{1}{2\gamma} - 4c^{k+1}\sigma_{\max}(F^\top F)\right]\left\|\lambda^k - \bar{\lambda}^{k+1}\right\|^2 \\
&\quad + \left[\frac{1}{2c^{k+1}} - \frac{\gamma\sigma_{\max}(F^\top F)}{2} - 4c^{k+1}L^2\right]\left\|x^k - \bar{x}^{k+1}\right\|^2 \\
&\quad + \frac{1}{2c^{k+1}}\left\|x^{k+1} - \bar{x}^{k+1}\right\|^2,
\end{aligned} \tag{15}$$

where δ^{k+1} and $\bar{\delta}^{k+1}$ are respectively denoted by

$$\begin{aligned}
\delta^{k+1} &= \nabla l(x^k, \xi_1^{k+1}) - \nabla l(x^k) \quad \text{and} \\
\bar{\delta}^{k+1} &= \nabla l(\bar{x}^{k+1}, \xi_2^{k+1}) - \nabla l(\bar{x}^{k+1}).
\end{aligned} \tag{16}$$

Proof. The first-order optimality condition for updating z^{k+1} is given by

$$r_2(z) - r_2(z^{k+1}) + \left\langle z - z^{k+1}, \bar{\lambda}^{k+1}\right\rangle \geq 0. \tag{17}$$

For $\bar{x}^{k+1}, x^{k+1} \in \mathcal{X}$ and any $x \in \mathcal{X}$, the first-order optimality condition for updating $\bar{x}^{k+1}$ and x^{k+1} are given respectively by

$$r_1(x) - r_1(\bar{x}^{k+1}) + \left\langle x - \bar{x}^{k+1}, \frac{\bar{x}^{k+1} - x^k}{c^{k+1}} + G\left(z^{k+1}, x^k, \lambda^k; \xi_1^{k+1}\right)\right\rangle \geq 0, \tag{18}$$

$$r_1(x) - r_1(x^{k+1}) + \left\langle x - x^{k+1}, \frac{x^{k+1} - x^k}{c^{k+1}} + G\left(z^{k+1}, \bar{x}^{k+1}, \bar{\lambda}^{k+1}; \xi_2^{k+1}\right)\right\rangle \geq 0. \tag{19}$$

Setting $x = x^{k+1}$ in (18) and $x = \bar{x}^{k+1}$ in (19), and summing two resulting inequalities yields that

$$\begin{aligned}
&\frac{1}{c^{k+1}}\left\|x^{k+1} - \bar{x}^{k+1}\right\|^2 \\
&\leq \left\langle x^{k+1} - \bar{x}^{k+1}, G\left(z^{k+1}, x^k, \lambda^k; \xi_1^{k+1}\right) - G\left(z^{k+1}, \bar{x}^{k+1}, \bar{\lambda}^{k+1}; \xi_2^{k+1}\right)\right\rangle \\
&\leq \left\|x^{k+1} - \bar{x}^{k+1}\right\|\left\|G\left(z^{k+1}, x^k, \lambda^k; \xi_1^{k+1}\right) - G\left(z^{k+1}, \bar{x}^{k+1}, \bar{\lambda}^{k+1}; \xi_2^{k+1}\right)\right\|,
\end{aligned}$$

which implies that

$$\begin{aligned}
\left\|x^{k+1} - \bar{x}^{k+1}\right\| &\leq c^{k+1}\left\|G\left(z^{k+1}, x^k, \lambda^k; \xi_1^{k+1}\right)\right. \\
&\quad \left. - G\left(z^{k+1}, \bar{x}^{k+1}, \bar{\lambda}^{k+1}; \xi_2^{k+1}\right)\right\|.
\end{aligned} \tag{20}$$

Therefore, we get

$$\begin{aligned}
&r_1(x^{k+1}) - r_1(\bar{x}^{k+1}) + \left\langle x^{k+1} - \bar{x}^{k+1}, G\left(z^{k+1}, \bar{x}^{k+1}, \bar{\lambda}^{k+1}; \xi_2^{k+1}\right)\right\rangle \\
&\geq \left\langle x^{k+1} - \bar{x}^{k+1}, G\left(z^{k+1}, \bar{x}^{k+1}, \bar{\lambda}^{k+1}; \xi_2^{k+1}\right) - G\left(z^{k+1}, x^k, \lambda^k; \xi_1^{k+1}\right)\right\rangle \\
&\quad - \left\langle x^{k+1} - \bar{x}^{k+1}, \frac{\bar{x}^{k+1} - x^k}{c^{k+1}}\right\rangle \\
&\geq -c^{k+1}\left\|G\left(z^{k+1}, \bar{x}^{k+1}, \bar{\lambda}^{k+1}; \xi_2^{k+1}\right) - G\left(z^{k+1}, x^k, \lambda^k; \xi_1^{k+1}\right)\right\|^2 \\
&\quad - \frac{1}{2c^{k+1}}\left\|x^{k+1} - x^k\right\|^2 + \frac{1}{2c^{k+1}}\left\|x^{k+1} - \bar{x}^{k+1}\right\|^2 + \frac{1}{2c^{k+1}}\left\|\bar{x}^{k+1} - x^k\right\|^2.
\end{aligned} \tag{21}$$

where the first inequality is obtained by letting $x = x^{k+1}$ in (18) and the second inequality follows from (20). Furthermore, we have

$$\begin{aligned}
&\left\|G\left(z^{k+1}, \bar{x}^{k+1}, \bar{\lambda}^{k+1}; \xi_2^{k+1}\right) - G\left(z^{k+1}, x^k, \lambda^k; \xi_1^{k+1}\right)\right\|^2 \\
&= \left\|\bar{\delta}^{k+1} + \nabla l(\bar{x}^{k+1}) - F^\top\bar{\lambda}^{k+1} - \left[\delta^{k+1} + \nabla l(x^k) - F^\top\lambda^k\right]\right\|^2 \\
&\leq 4\left\|\delta^{k+1}\right\|^2 + 4\left\|\bar{\delta}^{k+1}\right\|^2 + 4L^2\left\|x^k - \bar{x}^{k+1}\right\|^2
\end{aligned} \tag{22}$$

$$+ 4\sigma_{\max}(F^\top F)\left\|\lambda^k - \bar{\lambda}^{k+1}\right\|^2,$$

where δ^{k+1} and $\bar{\delta}^{k+1}$ are defined in (16). By substituting (22) into (21), and then summing the resulting inequality and (19), we have

$$\begin{aligned}
&r_1(x) - r_1(\bar{x}^{k+1}) + \left\langle x - \bar{x}^{k+1}, G\left(z^{k+1}, \bar{x}^{k+1}, \bar{\lambda}^{k+1}; \xi_2^{k+1}\right)\right\rangle \qquad (23)\\
&\geq -4c^{k+1}\left\|\delta^{k+1}\right\|^2 - 4c^{k+1}\left\|\bar{\delta}^{k+1}\right\|^2 - 4c^{k+1}\sigma_{\max}(F^\top F)\left\|\lambda^k - \bar{\lambda}^{k+1}\right\|^2\\
&\quad - 4c^{k+1}L^2\left\|x^k - \bar{x}^{k+1}\right\|^2 - \frac{1}{2c^{k+1}}\left\|x^{k+1} - x^k\right\|^2\\
&\quad + \frac{1}{2c^{k+1}}\left\|x^{k+1} - \bar{x}^{k+1}\right\|^2\\
&\quad + \frac{1}{2c^{k+1}}\left\|\bar{x}^{k+1} - x^k\right\|^2 - \left\langle x - x^{k+1}, \frac{x^{k+1} - x^k}{c^{k+1}}\right\rangle\\
&= -4c^{k+1}\left\|\delta^{k+1}\right\|^2 - 4c^{k+1}\left\|\bar{\delta}^{k+1}\right\|^2 - 4c^{k+1}\sigma_{\max}(F^\top F)\left\|\lambda^k - \bar{\lambda}^{k+1}\right\|^2\\
&\quad + \frac{1}{2c^{k+1}}\left\|x^{k+1} - \bar{x}^{k+1}\right\|^2 + \left[\frac{1}{2c^{k+1}} - 4c^{k+1}L^2\right]\left\|x^k - \bar{x}^{k+1}\right\|^2\\
&\quad - \frac{1}{2c^{k+1}}\left\|x - x^k\right\|^2 + \frac{1}{2c^{k+1}}\left\|x - x^{k+1}\right\|^2.
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
&\left\langle \lambda - \bar{\lambda}^{k+1}, F\bar{x}^{k+1} - z^{k+1}\right\rangle \qquad (24)\\
&= \frac{1}{\gamma}\left\langle \lambda - \lambda^{k+1} + \lambda^{k+1} - \bar{\lambda}^{k+1}, \lambda^k - \lambda^{k+1}\right\rangle\\
&= -\frac{1}{2\gamma}\left\|\lambda - \lambda^k\right\|^2 + \frac{1}{2\gamma}\left\|\lambda - \lambda^{k+1}\right\|^2 - \frac{1}{2\gamma}\left\|\lambda^{k+1} - \bar{\lambda}^{k+1}\right\|^2\\
&\quad + \frac{1}{2\gamma}\left\|\lambda^k - \bar{\lambda}^{k+1}\right\|^2\\
&\geq -\frac{1}{2\gamma}\left\|\lambda - \lambda^k\right\|^2 + \frac{1}{2\gamma}\left\|\lambda - \lambda^{k+1}\right\|^2 + \frac{1}{2\gamma}\left\|\lambda^k - \bar{\lambda}^{k+1}\right\|^2\\
&\quad - \frac{\gamma\sigma_{\max}(F^\top F)}{2}\left\|x^k - \bar{x}^{k+1}\right\|^2,
\end{aligned}$$

where the last inequality holds since

$$\lambda^{k+1} - \bar{\lambda}^{k+1} = \gamma\left(F\bar{x}^{k+1} - z^{k+1}\right) - \gamma\left(Fx^k - z^{k+1}\right) = \gamma\left(F\bar{x}^{k+1} - Fx^k\right).$$

Finally, combining (17), (23) and (24) yields (15). □

4.1. Proof of Theorem 3.1

Lemma 4.2. *Suppose that* $\left\{z^{k+1}, \bar{x}^{k+1}, \bar{\lambda}^{k+1}, x^{k+1}, \lambda^{k+1}\right\}$ *are generated by the SPDPEG algorithm, and* α^{k+1} *and* c^{k+1} *are defined in the main paper. For any optimal solution* (z^*, x^*, λ^*), *it holds that*

$$\begin{aligned}
&l(x^*) + r_1(x^*) + r_2(z^*) - \mathbb{E}[l(\bar{x}^{k+1})] - \mathbb{E}[r_1(\bar{x}^{k+1})] - \mathbb{E}[r_2(z^{k+1})] \qquad (25)\\
&\quad + \mathbb{E}\left[\begin{pmatrix} z^* - z^{k+1}\\ x^* - \bar{x}^{k+1}\\ \lambda - \bar{\lambda}^{k+1}\end{pmatrix}^\top \begin{pmatrix} \bar{\lambda}^{k+1}\\ -F^\top\bar{\lambda}^{k+1}\\ F\bar{x}^{k+1} - z^{k+1}\end{pmatrix}\right]\\
&\geq \frac{\sqrt{k+1} + \tilde{L}}{2}\mathbb{E}\left\|x^* - x^{k+1}\right\|^2 - \frac{\sqrt{k+1} + \tilde{L}}{2}\mathbb{E}\left\|x^* - x^k\right\|^2 - \frac{8\sigma^2}{\sqrt{k+1}}\\
&\quad - \frac{1}{2\gamma}\mathbb{E}\left\|\lambda - \lambda^k\right\|^2 + \frac{1}{2\gamma}\mathbb{E}\left\|\lambda - \lambda^{k+1}\right\|^2.
\end{aligned}$$

Proof. By the definition of $\tilde{L}$, we have $\frac{1}{2\gamma} - 4c^{k+1}\sigma_{\max}(F^\top F) \geq 0$ and $\frac{1}{2c^{k+1}} - \frac{\gamma\sigma_{\max}(F^\top F)}{2} - 4c^{k+1}L^2 \geq 0$. Plugging them into (15) yields that

$$\begin{aligned}
&r_1(x) + r_2(z) - r_1(\bar{x}^{k+1}) - r_2(z^{k+1}) + \begin{pmatrix} z - z^{k+1}\\ x - \bar{x}^{k+1}\\ \lambda - \bar{\lambda}^{k+1}\end{pmatrix}^\top\\
&\quad \times \begin{pmatrix} \bar{\lambda}^{k+1}\\ G\left(z^{k+1}, \bar{x}^{k+1}, \bar{\lambda}^{k+1}; \xi_2^{k+1}\right)\\ F\bar{x}^{k+1} - z^{k+1}\end{pmatrix}\\
&\geq \frac{1}{2c^{k+1}}\left\|x - x^{k+1}\right\|^2 - \frac{1}{2c^{k+1}}\left\|x - x^k\right\|^2 - 4c^{k+1}\left\|\delta^{k+1}\right\|^2\\
&\quad - 4c^{k+1}\left\|\bar{\delta}^{k+1}\right\|^2\\
&\quad - \frac{1}{2\gamma}\left\|\lambda - \lambda^k\right\|^2 + \frac{1}{2\gamma}\left\|\lambda - \lambda^{k+1}\right\|^2.
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
&\left(x - \bar{x}^{k+1}\right)^\top G\left(z^{k+1}, \bar{x}^{k+1}, \bar{\lambda}^{k+1}; \xi_2^{k+1}\right)\\
&= \left(x - \bar{x}^{k+1}\right)^\top \nabla l(\bar{x}^{k+1}) + \left(x - \bar{x}^{k+1}\right)^\top \bar{\delta}^{k+1}\\
&\quad + \left(x - \bar{x}^{k+1}\right)^\top\left[-F^\top\bar{\lambda}^{k+1}\right]\\
&\leq l(x) - l(\bar{x}^{k+1}) + \left(x - \bar{x}^{k+1}\right)^\top\left[-F^\top\bar{\lambda}^{k+1}\right] + \left(x - \bar{x}^{k+1}\right)^\top\bar{\delta}^{k+1}.
\end{aligned}$$

Therefore, we conclude that

$$\begin{aligned}
&l(x) + r_1(x) + r_2(z) - l(\bar{x}^{k+1}) - r_1(\bar{x}^{k+1}) - r_2(z^{k+1})\\
&\quad + \begin{pmatrix} z - z^{k+1}\\ x - \bar{x}^{k+1}\\ \lambda - \bar{\lambda}^{k+1}\end{pmatrix}^\top \begin{pmatrix} \bar{\lambda}^{k+1}\\ -F^\top\bar{\lambda}^{k+1}\\ F\bar{x}^{k+1} - z^{k+1}\end{pmatrix}\\
&\geq \frac{\sqrt{k+1} + \tilde{L}}{2}\left\|x - x^{k+1}\right\|^2 - \frac{\sqrt{k+1} + \tilde{L}}{2}\left\|x - x^k\right\|^2\\
&\quad - \frac{4}{\sqrt{k+1} + \tilde{L}}\left(\left\|\delta^{k+1}\right\|^2 + \left\|\bar{\delta}^{k+1}\right\|^2\right)\\
&\quad - \frac{1}{2\gamma}\left\|\lambda - \lambda^k\right\|^2 + \frac{1}{2\gamma}\left\|\lambda - \lambda^{k+1}\right\|^2 - \left(x - \bar{x}^{k+1}\right)^\top\bar{\delta}^{k+1}.
\end{aligned}$$

Since x^k and $\bar{x}^{k+1}$ are independent of ξ_1^{k+1} and ξ_2^{k+1}, respectively, we take the expectation on both sides of above inequality conditioning on ξ_2^{k+1}, then ξ_1^{k+1}, and then $\{\xi_1^j, \xi_2^j\}_{j\leq k}$. Finally, we set $(z, x) = (z^*, x^*)$, and conclude (25). □

We are ready to prove Theorem 3.1. For any $\lambda \in \mathbb{R}^p$, we have

$$\begin{aligned}
&l(x^*) + r_1(x^*) + r_2(z^*) - \mathbb{E}[l(\tilde{x}^t)] - \mathbb{E}[r_1(\tilde{x}^t)] - \mathbb{E}[r_2(\tilde{z}^t)]\\
&\quad + \lambda^\top\left(F\mathbb{E}[\tilde{x}^t] - \mathbb{E}[\tilde{z}^t]\right)\\
&= l(x^*) + r_1(x^*) + r_2(z^*) - \mathbb{E}[l(\tilde{x}^t)] - \mathbb{E}[r_1(\tilde{x}^t)] - \mathbb{E}[r_2(\tilde{z}^t)]\\
&\quad + \mathbb{E}\left[\begin{pmatrix} z^* - \tilde{z}^t\\ x^* - \tilde{x}^t\\ \lambda - \tilde{\lambda}^t\end{pmatrix}^\top \begin{pmatrix} \tilde{\lambda}^t\\ -F^\top\tilde{\lambda}^t\\ F\tilde{x}^t - \tilde{z}^t\end{pmatrix}\right]\\
&\geq \frac{1}{t+1}\sum_{k=0}^{t}\Big\{l(x^*) + r_1(x^*) + r_2(z^*) - \mathbb{E}[l(\bar{x}^{k+1})] - \mathbb{E}[r_1(\bar{x}^{k+1})]\\
&\quad - \mathbb{E}[r_2(z^{k+1})]\\
&\quad + \mathbb{E}\left[\begin{pmatrix} z^* - z^{k+1}\\ x^* - \bar{x}^{k+1}\\ \lambda - \bar{\lambda}^{k+1}\end{pmatrix}^\top \begin{pmatrix} \bar{\lambda}^{k+1}\\ -F^\top\bar{\lambda}^{k+1}\\ F\bar{x}^{k+1} - z^{k+1}\end{pmatrix}\right]\Big\}\\
&\geq \frac{1}{t+1}\sum_{k=0}^{t}\Bigg[\frac{\sqrt{k+1} + \tilde{L}}{2}\mathbb{E}\left\|x^* - x^{k+1}\right\|^2\\
&\quad - \frac{\sqrt{k+1} + \tilde{L}}{2}\mathbb{E}\left\|x^* - x^k\right\|^2 - \frac{8\sigma^2}{\sqrt{k+1}}\\
&\quad - \frac{1}{2\gamma}\mathbb{E}\left\|\lambda - \lambda^k\right\|^2 + \frac{1}{2\gamma}\mathbb{E}\left\|\lambda - \lambda^{k+1}\right\|^2\Bigg]
\end{aligned}$$

$$\geq -\frac{\tilde{L}}{2(t+1)}\left\|x^*-x^0\right\|^2-\frac{D_x^2+16\sigma^2}{2\sqrt{t+1}}-\frac{1}{2\gamma(t+1)}\left\|\lambda-\lambda^0\right\|^2, \tag{26}$$

where the first inequality holds due to the convexity of l, r_1 and r_2. Note that the optimality condition imply the following inequality

$$\begin{aligned}0 \geq\ & l(x^*)+r_1(x^*)+r_2(z^*)-\mathbb{E}[l(\tilde{x}^t)]-\mathbb{E}[r_1(\tilde{x}^t)]\\ &-\mathbb{E}[r_2(\tilde{z}^t)]+(\lambda^*)^\top\left(F\mathbb{E}[\tilde{x}^t]-\mathbb{E}[\tilde{z}^t]\right).\end{aligned} \tag{27}$$

Now, define $\rho := \|\lambda^*\|+1$. By using Cauchy–Schwarz inequality in (27), we obtain

$$\begin{aligned}0 \leq\ & \mathbb{E}[l(\tilde{x}^t)]+\mathbb{E}[r_1(\tilde{x}^t)]+\mathbb{E}[r_2(\tilde{z}^t)]-l(x^*)-r_1(x^*)-r_2(z^*)\\ &+\rho\left\|F\mathbb{E}[\tilde{x}^t]-\mathbb{E}[\tilde{z}^t]\right\|.\end{aligned} \tag{28}$$

By setting $\lambda=-\rho\left(F\mathbb{E}[\tilde{x}^t]-\mathbb{E}[\tilde{z}^t]\right)/\left\|F\mathbb{E}[\tilde{x}^t]-\mathbb{E}[\tilde{z}^t]\right\|$ in (26), and noting that $\|\lambda\|=\rho$, we obtain

$$\begin{aligned}&\mathbb{E}[l(\tilde{x}^t)]+\mathbb{E}[r_1(\tilde{x}^t)]+\mathbb{E}[r_2(\tilde{z}^t)]-l(x^*)-r_1(x^*)-r_2(z^*)\\ &\quad+\rho\left\|F\mathbb{E}[\tilde{x}^t]-\mathbb{E}[\tilde{z}^t]\right\|\\ &\leq \frac{\tilde{L}D_x^2}{2(t+1)}+\frac{D_x^2+16\sigma^2}{2\sqrt{t+1}}+\frac{\rho^2+\left\|\lambda^0\right\|^2}{\gamma(t+1)}.\end{aligned} \tag{29}$$

We now define the function

$$v(\eta)=\min\{l(x)+r_1(x)+r_2(z)|Fx-z=\eta, x\in\mathcal{X}\}.$$

It is easy to verify that v is convex, $v(0)=l(x^*)+r_1(x^*)+r_2(z^*)$, and $\lambda^*\in\partial v(0)$. Therefore, from the convexity of v, it holds that

$$v(\eta)\geq v(0)+\langle\lambda^*,\eta\rangle\geq l(x^*)+r_1(x^*)+r_2(z^*)-\|\lambda^*\|\|\eta\|. \tag{30}$$

Let $\bar{\eta}=F\mathbb{E}[\tilde{x}^t]-\mathbb{E}[\tilde{z}^t]$, we have

$$\begin{aligned}&\mathbb{E}[l(\tilde{x}^t)]+\mathbb{E}[r_1(\tilde{x}^t)]+\mathbb{E}[r_2(\tilde{z}^t)]\\ &\quad\geq l(\mathbb{E}[\tilde{x}^t])+r_1(\mathbb{E}[\tilde{x}^t])+r_2(\mathbb{E}[\tilde{z}^t])\geq v(\bar{\eta}).\end{aligned}$$

Therefore, combining (28), (29) and (30), we get

$$\begin{aligned}&-\|\lambda^*\|\|\bar{\eta}\|\\ &\quad\leq \mathbb{E}[l(\tilde{x}^t)]+\mathbb{E}[r_1(\tilde{x}^t)]+\mathbb{E}[r_2(\tilde{z}^t)]-l(x^*)-r_1(x^*)-r_2(z^*)\\ &\quad\leq \frac{\tilde{L}D_x^2}{2(t+1)}+\frac{D_x^2+16\sigma^2}{2\sqrt{t+1}}+\frac{\rho^2+\left\|\lambda^0\right\|^2}{\gamma(t+1)}-\rho\|\bar{\eta}\|,\end{aligned}$$

which implies (9) and (10).

4.2. Proof of Theorem 3.2

Lemma 4.3. *Let $\left\{z^{k+1},\bar{x}^{k+1},\bar{\lambda}^{k+1},x^{k+1},\lambda^{k+1}\right\}$ be generated by the SPDPEG Algorithm, and α^{k+1} and c^{k+1} be defined in the main paper. For any optimal solution (z^*, x^*), it holds that*

$$\begin{aligned}&l(x^*)+r_1(x^*)+r_2(z^*)-\mathbb{E}[l(\bar{x}^{k+1})]-\mathbb{E}[r_1(\bar{x}^{k+1})]-\mathbb{E}[r_2(z^{k+1})]\\ &\quad+\mathbb{E}\left[\begin{pmatrix}z^*-z^{k+1}\\ x^*-\bar{x}^{k+1}\\ \lambda-\bar{\lambda}^{k+1}\end{pmatrix}^\top\begin{pmatrix}\bar{\lambda}^{k+1}\\ -F^\top\bar{\lambda}^{k+1}\\ F\bar{x}^{k+1}-z^{k+1}\end{pmatrix}\right]\\ &\geq \frac{\mu(k+2)+2\tilde{L}}{4}\mathbb{E}\left\|x^*-x^{k+1}\right\|^2-\frac{\mu(k+1)+2\tilde{L}}{4}\mathbb{E}\left\|x^*-x^k\right\|^2\\ &\quad-\frac{16\sigma^2}{\mu(k+1)}\\ &\quad-\frac{1}{2\gamma}\mathbb{E}\left\|\lambda-\lambda^k\right\|^2+\frac{1}{2\gamma}\mathbb{E}\left\|\lambda-\lambda^{k+1}\right\|^2.\end{aligned} \tag{31}$$

Proof. Since $\frac{1}{2\gamma}-4c^{k+1}\sigma_{\max}(F^\top F)\geq 0$ and $\frac{1}{2c^{k+1}}-\frac{\gamma\sigma_{\max}(F^\top F)}{2}-4c^{k+1}L^2\geq 0$ and $c^{k+1}<\frac{1}{\mu}$, we conclude from (15) that

$$\begin{aligned}&r_1(x)+r_2(z)-r_1(\bar{x}^{k+1})-r_2(z^{k+1})\\ &\quad+\begin{pmatrix}z-z^{k+1}\\ x-\bar{x}^{k+1}\\ \lambda-\bar{\lambda}^{k+1}\end{pmatrix}^\top\begin{pmatrix}\bar{\lambda}^{k+1}\\ G\left(z^{k+1},\bar{x}^{k+1},\bar{\lambda}^{k+1};\xi_2^{k+1}\right)\\ F\bar{x}^{k+1}-z^{k+1}\end{pmatrix}\\ &\geq \frac{1}{2c^{k+1}}\left\|x-x^{k+1}\right\|^2-\frac{1}{2c^{k+1}}\left\|x-x^k\right\|^2-4c^{k+1}\left\|\delta^{k+1}\right\|^2\\ &\quad-4c^{k+1}\left\|\bar{\delta}^{k+1}\right\|^2\\ &\quad-\frac{1}{2\gamma}\left\|\lambda-\lambda^k\right\|^2+\frac{1}{2\gamma}\left\|\lambda-\lambda^{k+1}\right\|^2+\frac{\mu}{2}\left\|\bar{x}^{k+1}-x^{k+1}\right\|^2.\end{aligned}$$

Moreover, we have

$$\begin{aligned}&\left(x-\bar{x}^{k+1}\right)^\top G\left(z^{k+1},\bar{x}^{k+1},\bar{\lambda}^{k+1};\xi_2^{k+1}\right)\\ &=\left(x-\bar{x}^{k+1}\right)^\top\nabla l(\bar{x}^{k+1})+\left(x-\bar{x}^{k+1}\right)^\top\bar{\delta}^{k+1}\\ &\quad+\left(x-\bar{x}^{k+1}\right)^\top\left[-A^\top\bar{\lambda}^{k+1}\right]\\ &\leq l(x)-l(\bar{x}^{k+1})-\frac{\mu}{2}\left\|x-\bar{x}^{k+1}\right\|^2+\left(x-\bar{x}^{k+1}\right)^\top\left[-A^\top\bar{\lambda}^{k+1}\right]\\ &\quad+\left(x-\bar{x}^{k+1}\right)^\top\bar{\delta}^{k+1}.\end{aligned}$$

Therefore, we conclude that

$$\begin{aligned}&l(x)+r_1(x)+r_2(z)-l(\bar{x}^{k+1})-r_1(\bar{x}^{k+1})-r_2(z^{k+1})\\ &\quad+\begin{pmatrix}z-z^{k+1}\\ x-\bar{x}^{k+1}\\ \lambda-\bar{\lambda}^{k+1}\end{pmatrix}^\top\begin{pmatrix}\bar{\lambda}^{k+1}\\ -F^\top\bar{\lambda}^{k+1}\\ F\bar{x}^{k+1}-z^{k+1}\end{pmatrix}\\ &\geq \frac{\mu(k+1)+2\tilde{L}}{4}\left\|x-x^{k+1}\right\|^2-\frac{\mu(k+1)+2\tilde{L}}{4}\left\|x-x^k\right\|^2\\ &\quad-\frac{8}{\mu(k+1)+\tilde{L}}\left(\left\|\delta^{k+1}\right\|^2+\left\|\bar{\delta}^{k+1}\right\|^2\right)\\ &\quad-\frac{1}{2\gamma}\left\|\lambda-\lambda^k\right\|^2+\frac{1}{2\gamma}\left\|\lambda-\lambda^{k+1}\right\|^2-\left(x-\bar{x}^{k+1}\right)^\top\bar{\delta}^{k+1}\\ &\quad+\frac{\mu}{2}\left\|x-\bar{x}^{k+1}\right\|^2+\frac{\mu}{2}\left\|\bar{x}^{k+1}-x^{k+1}\right\|^2\\ &\geq \frac{\mu(k+2)+2\tilde{L}}{4}\left\|x-x^{k+1}\right\|^2-\frac{\mu(k+1)+2\tilde{L}}{4}\left\|x-x^k\right\|^2\\ &\quad-\frac{8}{\mu(k+1)+2\tilde{L}}\left(\left\|\delta^{k+1}\right\|^2+\left\|\bar{\delta}^{k+1}\right\|^2\right)\\ &\quad-\frac{1}{2\gamma}\left\|\lambda-\lambda^k\right\|^2+\frac{1}{2\gamma}\left\|\lambda-\lambda^{k+1}\right\|^2-\left(x-\bar{x}^{k+1}\right)^\top\bar{\delta}^{k+1}.\end{aligned}$$

Since x^k and $\bar{x}^{k+1}$ are independent of ξ_1^{k+1} and ξ_2^{k+1} respectively, we take the expectation on both sides of above inequality conditioning on ξ_2^{k+1}, then ξ_1^{k+1} and then $\{\xi_1^j,\xi_2^j\}_{j\leq k}$. Finally, we set $(z,x)=(z^*,x^*)$, and conclude (31). □

We are ready to prove Theorem 3.2. For any $\lambda\in\mathbb{R}^p$, we have

$$\begin{aligned}&l(x^*)+r_1(x^*)+r_2(z^*)-\mathbb{E}[l(\tilde{x}^t)]-\mathbb{E}[r_1(\tilde{x}^t)]\\ &\quad-\mathbb{E}[r_2(\tilde{z}^t)]+\lambda^\top\left(F\mathbb{E}[\tilde{x}^t]-\mathbb{E}[\tilde{z}^t]\right)\\ &\geq \frac{1}{t+1}\sum_{k=0}^{t}\Bigg\{l(x^*)+r_1(x^*)+r_2(z^*)-\mathbb{E}[l(\bar{x}^{k+1})]\\ &\quad-\mathbb{E}[r_1(\bar{x}^{k+1})]-\mathbb{E}[r_2(z^{k+1})]\\ &\quad+\mathbb{E}\left[\begin{pmatrix}z^*-z^{k+1}\\ x^*-\bar{x}^{k+1}\\ \lambda-\bar{\lambda}^{k+1}\end{pmatrix}^\top\begin{pmatrix}\bar{\lambda}^{k+1}\\ -F^\top\bar{\lambda}^{k+1}\\ F\bar{x}^{k+1}-z^{k+1}\end{pmatrix}\right]\Bigg\}\\ &\geq \frac{1}{t+1}\sum_{k=0}^{t}\Bigg[\frac{\mu(k+2)+2\tilde{L}}{4}\mathbb{E}\left\|x^*-x^{k+1}\right\|^2\end{aligned}$$

$$
\begin{aligned}
&- \frac{\mu(k+1)+2\tilde{L}}{4}\mathbb{E}\left\|x^*-x^k\right\|^2 - \frac{16\sigma^2}{\mu(k+1)} \\
&- \frac{1}{2\gamma}\mathbb{E}\left\|\lambda-\lambda^k\right\|^2 + \frac{1}{2\gamma}\mathbb{E}\left\|\lambda-\lambda^{k+1}\right\|^2\Bigg] \\
\geq& -\frac{\mu+2\tilde{L}}{4(t+1)}\left\|x^*-x^0\right\|^2 - \frac{16\sigma^2\log(t+1)}{\mu(t+1)} \\
&- \frac{1}{2\gamma(t+1)}\left\|\lambda-\lambda^0\right\|^2.
\end{aligned}
$$

where the first inequality holds due to the convexity of l, r_1 and r_2. By the same argument as Theorem 3.1, we conclude (11) and (12).

4.3. Proof of Theorem 3.3

Lemma 4.4. *Let $\left\{z^{k+1}, \bar{x}^{k+1}, \bar{\lambda}^{k+1}, x^{k+1}, \lambda^{k+1}\right\}$ be generated by the SPDPEG Algorithm, and α^{k+1} and c^{k+1} be defined in the main paper. For any optimal solution (z^*, x^*), it holds that*

$$
\begin{aligned}
&l(x^*)+r_1(x^*)+r_2(z^*)-\mathbb{E}[l(\bar{x}^{k+1})]-\mathbb{E}[r_1(\bar{x}^{k+1})]-\mathbb{E}[r_2(z^{k+1})] \\
&\quad+\mathbb{E}\left[\begin{pmatrix} z^*-z^{k+1} \\ x^*-\bar{x}^{k+1} \\ \lambda-\bar{\lambda}^{k+1}\end{pmatrix}^\top \begin{pmatrix} \bar{\lambda}^{k+1} \\ -F^\top\bar{\lambda}^{k+1} \\ F\bar{x}^{k+1}-z^{k+1}\end{pmatrix}\right] \\
&\geq \frac{\mu(k+4)+4\tilde{L}}{8}\mathbb{E}\left\|x^*-x^{k+1}\right\|^2 - \frac{\mu(k+2)+4\tilde{L}}{8}\mathbb{E}\left\|x^*-x^k\right\|^2 \\
&\quad-\frac{32\sigma^2}{\mu(k+2)} \\
&\quad-\frac{1}{2\gamma}\mathbb{E}\left\|\lambda-\lambda^k\right\|^2+\frac{1}{2\gamma}\mathbb{E}\left\|\lambda-\lambda^{k+1}\right\|^2. \qquad (32)
\end{aligned}
$$

Proof. By the same argument as Lemma 4.3, we conclude from (15) that

$$
\begin{aligned}
&r_1(x)+r_2(z)-r_1(\bar{x}^{k+1})-r_2(z^{k+1}) \\
&\quad+\begin{pmatrix} z-z^{k+1} \\ x-\bar{x}^{k+1} \\ \lambda-\bar{\lambda}^{k+1}\end{pmatrix}^\top \begin{pmatrix} \bar{\lambda}^{k+1} \\ G\left(z^{k+1},\bar{x}^{k+1},\bar{\lambda}^{k+1};\xi_2^{k+1}\right) \\ F\bar{x}^{k+1}-z^{k+1}\end{pmatrix} \\
&\geq \frac{1}{2c^{k+1}}\left\|x-x^{k+1}\right\|^2-\frac{1}{2c^{k+1}}\left\|x-x^k\right\|^2-4c^{k+1}\left\|\delta^{k+1}\right\|^2 \\
&\quad-4c^{k+1}\left\|\bar{\delta}^{k+1}\right\|^2 \\
&\quad-\frac{1}{2\gamma}\left\|\lambda-\lambda^k\right\|^2+\frac{1}{2\gamma}\left\|\lambda-\lambda^{k+1}\right\|^2+\frac{\mu}{2}\left\|\bar{x}^{k+1}-x^{k+1}\right\|^2,
\end{aligned}
$$

and

$$
\begin{aligned}
&\left(x-\bar{x}^{k+1}\right)^\top G\left(z^{k+1},\bar{x}^{k+1},\bar{\lambda}^{k+1};\xi_2^{k+1}\right) \\
&\leq l(x)-l(\bar{x}^{k+1})-\frac{\mu}{2}\left\|x-\bar{x}^{k+1}\right\|^2+\left(x-\bar{x}^{k+1}\right)^\top\left[-A^\top\bar{\lambda}^{k+1}\right] \\
&\quad+\left(x-\bar{x}^{k+1}\right)^\top\bar{\delta}^{k+1}.
\end{aligned}
$$

Therefore, we conclude that

$$
\begin{aligned}
&l(x)+r_1(x)+r_2(z)-l(\bar{x}^{k+1})-r_1(\bar{x}^{k+1})-r_2(z^{k+1}) \\
&\quad+\begin{pmatrix} z-z^{k+1} \\ x-\bar{x}^{k+1} \\ \lambda-\bar{\lambda}^{k+1}\end{pmatrix}^\top \begin{pmatrix} \bar{\lambda}^{k+1} \\ -F^\top\bar{\lambda}^{k+1} \\ F\bar{x}^{k+1}-z^{k+1}\end{pmatrix} \\
&\geq \frac{\mu(k+4)+4\tilde{L}}{8}\left\|x-x^{k+1}\right\|^2-\frac{\mu(k+2)+4\tilde{L}}{8}\left\|x-x^k\right\|^2 \\
&\quad-\frac{16}{\mu(k+2)+4\tilde{L}}\left(\left\|\delta^{k+1}\right\|^2+\left\|\bar{\delta}^{k+1}\right\|^2\right) \\
&\quad-\frac{1}{2\gamma}\left\|\lambda-\lambda^k\right\|^2+\frac{1}{2\gamma}\left\|\lambda-\lambda^{k+1}\right\|^2-\left(x-\bar{x}^{k+1}\right)^\top\bar{\delta}^{k+1}.
\end{aligned}
$$

Since x^k and $\bar{x}^{k+1}$ are independent of ξ_1^{k+1} and ξ_2^{k+1} respectively, we take the expectation on both sides of above inequality conditioning on ξ_2^{k+1}, then ξ_1^{k+1} and then $\{\xi_1^j, \xi_2^j\}_{j\leq k}$. Finally, we set $(z, x) = (z^*, x^*)$, and conclude (32). □

We are ready to prove Theorem 3.3. For any $\lambda \in \mathbb{R}^p$, we have

$$
\begin{aligned}
&l(x^*)+r_1(x^*)+r_2(z^*)-\mathbb{E}[l(\tilde{x}^t)]-\mathbb{E}[r_1(\tilde{x}^t)] \\
&\quad-\mathbb{E}[r_2(\tilde{z}^t)]+\lambda^\top\left(F\mathbb{E}[\tilde{x}^t]-\mathbb{E}[\tilde{z}^t]\right) \\
&\geq \frac{2}{(t+1)(t+6)}\sum_{k=0}^{t}(k+3)\Bigg\{l(x^*)+r_1(x^*)+r_2(z^*) \\
&\quad-\mathbb{E}[l(\bar{x}^{k+1})]-\mathbb{E}[r_1(\bar{x}^{k+1})]-\mathbb{E}[r_2(z^{k+1})] \\
&\quad+\mathbb{E}\left[\begin{pmatrix} z^*-z^{k+1} \\ x^*-\bar{x}^{k+1} \\ \lambda-\bar{\lambda}^{k+1}\end{pmatrix}^\top \begin{pmatrix} \bar{\lambda}^{k+1} \\ -F^\top\bar{\lambda}^{k+1} \\ F\bar{x}^{k+1}-z^{k+1}\end{pmatrix}\right]\Bigg\} \\
&\geq \frac{2}{(t+1)(t+6)}\sum_{k=0}^{t}(k+3)\Bigg[\frac{\mu(k+4)+4\tilde{L}}{8}\mathbb{E}\left\|x^*-x^{k+1}\right\|^2 \\
&\quad-\frac{\mu(k+2)+4\tilde{L}}{8}\mathbb{E}\left\|x^*-x^k\right\|^2 \\
&\quad-\frac{32\sigma^2}{\mu(k+2)}-\frac{1}{2\gamma}\mathbb{E}\left\|\lambda-\lambda^k\right\|^2+\frac{1}{2\gamma}\mathbb{E}\left\|\lambda-\lambda^{k+1}\right\|^2\Bigg] \\
&\geq -\frac{3\mu+2\tilde{L}}{4(t+1)(t+6)}\left\|x^*-x^0\right\|^2-\frac{96\sigma^2}{\mu(t+6)}-\frac{2\|\lambda\|^2+2D_\lambda^2}{\gamma(t+1)},
\end{aligned}
$$

where the first inequality holds due to the convexity of l, r_1 and r_2. By the same argument as Theorem 3.1, we conclude (13) and (14).

5. Experiment

We apply our proposed SPDPEG algorithm to solve following two popular problems: fused logistic regression (FLR) (33) and graph-guided regularized logistic regression (GGRLR) (34) [19], which are formulated as follows:

$$
\text{FLR:} \quad \min_x \; l(x)+\gamma\|x\|_1+\lambda\|Lx\|_1, \qquad (33)
$$

and

$$
\text{GGRLR:} \quad \min_x \; l(x)+\frac{\gamma}{2}\|x\|_2^2+\lambda\|Fx\|_1. \qquad (34)
$$

Here $l(x) = \frac{1}{N}\left[\sum_{i=1}^N l(x, \xi_i)\right]$, where $l(x, \xi_i)$ is the logistic loss on ξ_i and $\lambda > 0$ is a parameter. L and F are penalty matrices promoting the desired sparse structure of x. Specifically, $L \in \mathbb{R}^{(n-1)\times n}$ in problem (33) is specified as a matrix with all ones on the diagonal, negative ones on the super-diagonal and zeros elsewhere, and F in problem (34) is generated by sparse inverse covariance selection [29].

In the experiments, we compare our SPDPEG algorithm with the EGADM algorithm [8] and six existing stochastic ADMM-type algorithms[1]: SGADM [18], SADMM [11], OPG-ADMM [12], RDA-ADMM [12], and two adaptive SADMM (*i.e.*, SADMMdiag and SADMMfull)[14]. We exclude online ADMM [30] since [12] has shown that RDA-ADMM performs better than online ADMM. FSADMM [19] is also excluded since it requires storage of all gradients, which results in impractical performance in some complex applications [31].

[1] We use the implementation of SADMM, OPG-ADMM and RDA-ADMM provided by the authors and two adaptive ADMM according to [14].

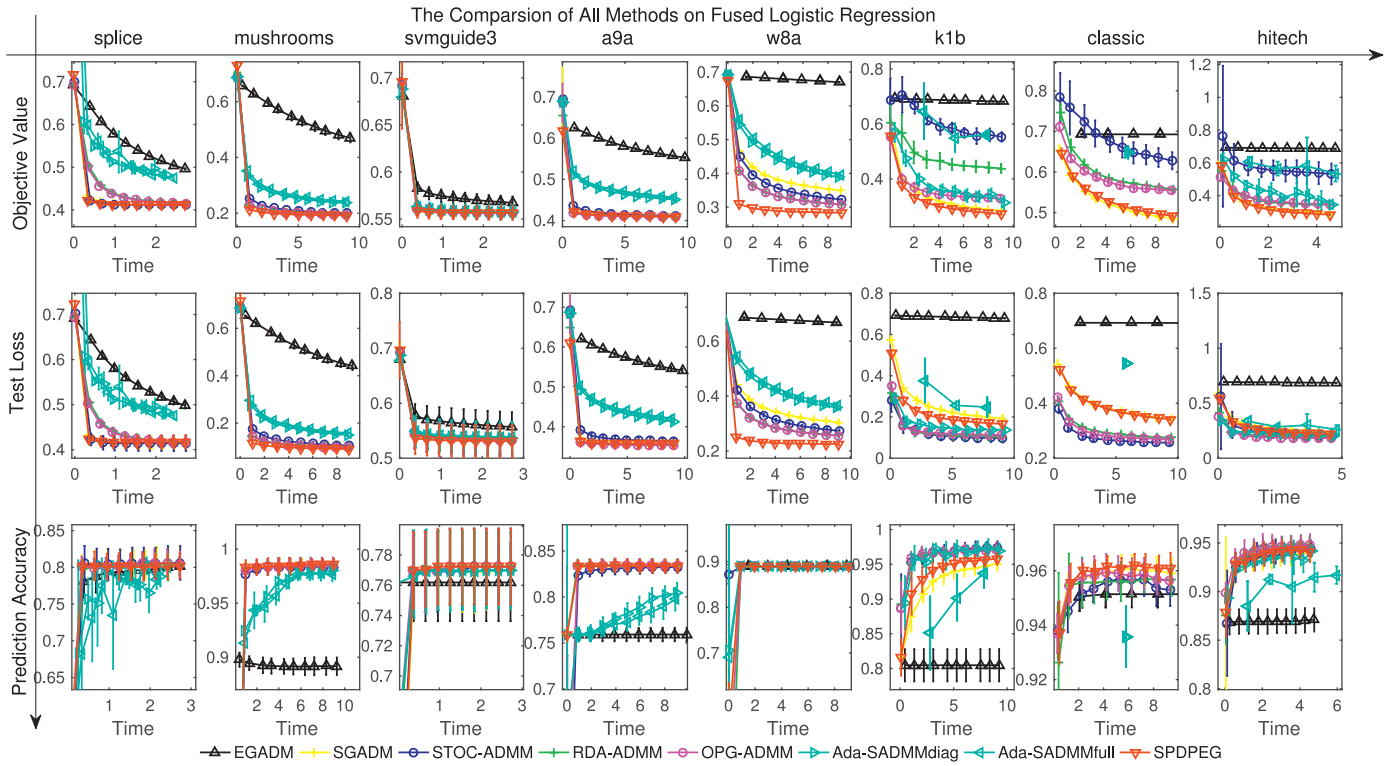


Fig. 1. Comparison of SPDPEG with SGADM, SADMM, RDA-ADMM, OPG-ADMM, SADMMdiag and SADMMfull on *fused logistic regression* task. **First row**: average objective values. **Second row**: average test losses. **Third row**: average prediction accuracies.

The experiments are conducted on five binary classification datasets: *splice, svmguide3, mushrooms, a9a*, and *w8a*[2] with large number of samples, *classic, hitech, k1b*[3] with high dimensionality. We set the parameters of SPDPEG exactly following our theory while using the cross validation to select the parameters for other algorithms. For each dataset, we calculate the lipschitz constant L as its classical upper bound $\hat{L} = 0.25 \max_{1 \le i \le n} \|a_i\|^2$. The regularization parameter $\lambda = 5 \times 10^{-3}$ and $\gamma = 5 \times 10^{-4}$ for problem (33), and $\lambda = 10^{-5}$ and $\gamma = 10^{-2}$ for problem (34). To reduce statistical variability, experimental results are repeated 5 rounds. Additionally, we use the metrics including objective value, test loss and prediction accuracy to compare our method with other methods. The "objective value" means the sum of the loss function and regularized terms evaluated on a training data sample, while the "test loss" means the value of the loss function evaluated on a test data sample. Specifically, we use objective function values on training datasets, test losses (*i.e.,* $l(x)$) on test datasets, and prediction accuracy on test datasets.

Fig. 1 shows the objective value, test loss and prediction accuracy as the functions of the time costs on the FLR task, where the objective function is convex but not necessarily strongly convex. We observe that our method mostly achieves the best performance, followed by six stochastic ADMM-type algorithms, all of which outperform EGADM by a large margin. We find that the prediction accuracy of the SPDPEG algorithm is competitive with other algorithms, which supports the use of extra-gradient in the SPDPEG algorithm. The performance of our SPDPEG algorithm on six datasets is the most stable and effective among all methods.

We further compare our algorithm with other algorithms on the GGRLR task, where the objective function is strongly convex. We use both uniformly and non-uniformly averaged iterates, noted as SPDPEG-SC1 (Uniformly Averaged) and SPDPEG-SC2 (Non-Uniformly Averaged). The experimental results presented in Fig. 2 show that our algorithm consistently outperforms other algorithms, and exhibits the advantage with non-uniformly averaged iterates over its counterpart with uniformly averaged iterates. This matches our analysis in the previous sections.

Table 1
Statistics of datasets.

Dataset	Number of samples	Dimensionality
splice	1000	60
svmguide3	1243	21
mushrooms	8124	112
a9a	32,561	123
w8a	64,700	300
hitech	2301	10,080
k1b	2340	21,839
classic	7094	41,681

6. Conclusions

In this paper, we proposed a novel algorithm, namely Stochastic Primal-Dual Proximal ExtraGradient (SPDPEG), to resolve stochastic minimization problems including two regularization terms, one of which is composed with a linear function $F(x)$, as shown in problem (1). Problem (1) is computationally difficult when the penalty matrix F is non-diagonal or the number of training samples is large.

Inspired by the nice efficiency of EGADM, we developed an ADM-type optimization scheme that employs proximal noisy extra-gradient descent to achieve reasonable numerical efficiency and stability. For general convex objectives, we showed that the

[2] https://www.csie.ntu.edu.tw/~cjlin/libsvm/
[3] https://www.shi-zhong.com/software/docdata.zip

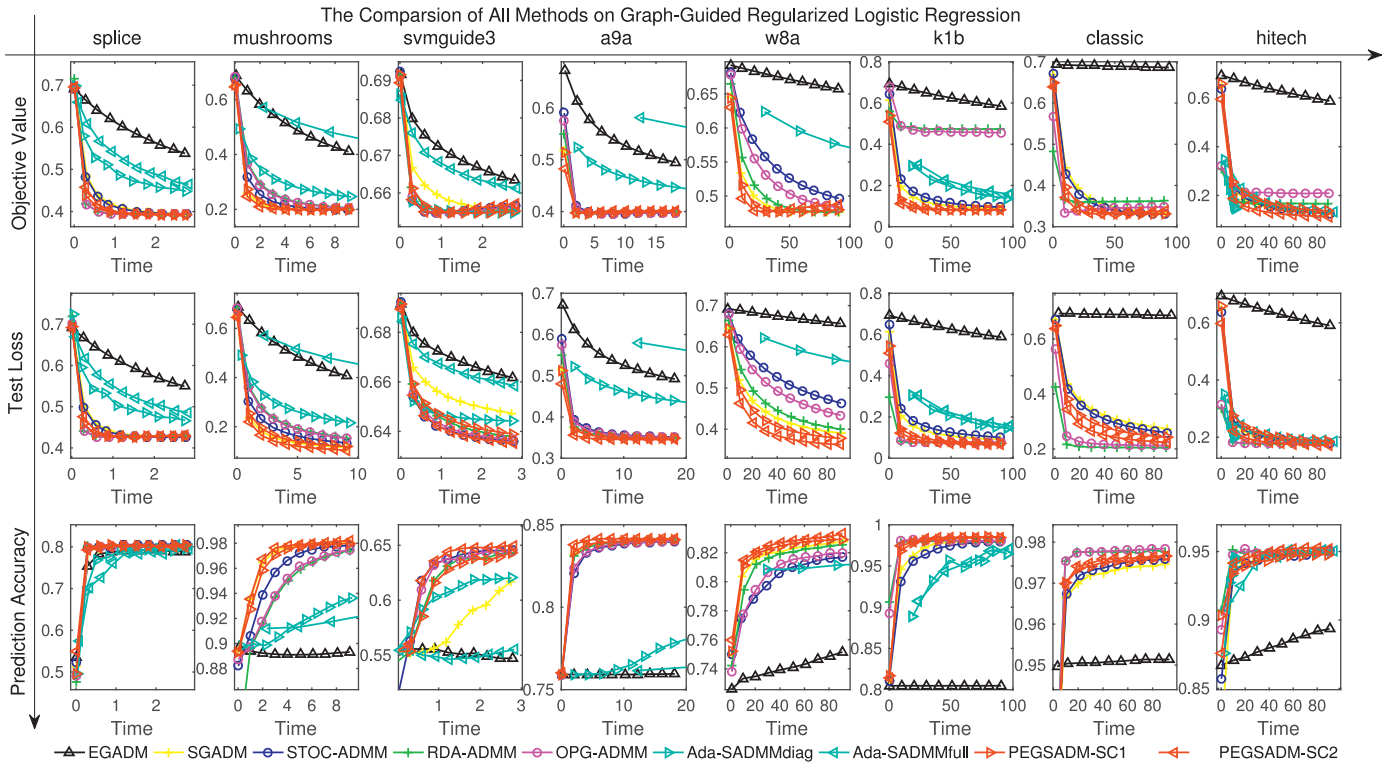


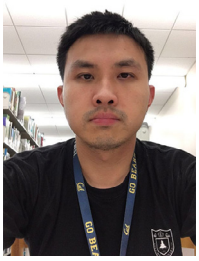
Fig. 2. Comparison of SPDPEG-SC1 (Uniformly Averaged) and SPDPEG-SC2 (Non-Uniformly Averaged) with SGADM, SADMM, RDA-ADMM, OPG-ADMM, SADMMdiag and SADMMfull on **graph-guided regularized logistic regression task. First row**: average objective values. **Second row**: average test losses. **Third row**: average prediction accuracies.

uniformly average iterates converge in expectation with the rate of $O(1/\sqrt{t})$; while for strongly convex objectives, the uniformly and non-uniformly average iterates generated by the SPDPEG algorithm were proven to converge in expectation with the $O(\log(t)/t)$ and $O(1/t)$ rates, respectively. It is worth mentioning that these rates are both known to be best possible for first-order stochastic optimization algorithms. The numerical experiments conducted on fused logistic regression and graph-guided regularized logistic regression problems demonstrated that our proposed algorithm consistently outperforms the other competing stochastic algorithms. A future research direction is to consider incorporating variance reduction techniques into the SPDPEG algorithm.

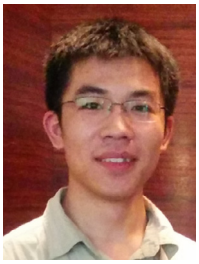
References

[1] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Methodol. (1996) 267–288.
[2] C. Cortes, V. Vapnik, Support vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[3] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, J. R. Stat. Soc. Ser. B Stat. Method. 67 (1) (2005) 91–108.
[4] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2nd edition, Springer, New York, 2009.
[5] N. Parikh, S. Boyd, Proximal algorithms, Found. Trends Optim. 1 (3) (2014) 127–239.
[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. 3 (1) (2011) 1–122.
[7] J. Yang, X. Yuan, Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization, Math. Comput. 82 (281) (2013) 301–329.
[8] T. Lin, S. Ma, S. Zhang, An extragradient-based alternating direction method for convex minimization, Found. Comput. Math. (2015) 1–25.
[9] G. Korpelevich, The extragradient method for finding saddle points and other problems, Ekonomika i Matematicheskie Metody 12 (1976) 747–756.
[10] G. Korpelevich, Extrapolation gradient methods and relation to modified Lagrangeans, Ekonomika i Matematicheskie Metody 19 (1983) 694–703.
[11] H. Ouyang, N. He, L. Tran, A. Gray, Stochastic alternating direction method of multipliers, in: Proceedings of the International Conference on Machine Learning, ICML, 2013, pp. 80–88.
[12] T. Suzuki, Dual averaging and proximal gradient descent for online alternating direction multiplier method, in: Proceedings of the International Conference on Machine Learning, ICML, 2013, pp. 392–400.
[13] S. Azadi, S. Sra, Towards an optimal stochastic alternating direction method of multipliers, in: Proceedings of the International Conference on Machine Learning, ICML, 2014, pp. 620–628.
[14] P. Zhao, J. Yang, T. Zhang, P. Li, Adaptive stochastic alternating direction method of multipliers, in: Proceedings of the International Conference on Machine Learning, ICML, 2015, pp. 69–77.
[15] S. Zheng, J.T. Kwok, Stochastic variance-reduced ADMM, arXiv preprint arXiv:1604.07070 (2016).
[16] J. Duchi, Y. Singer, Efficient online and batch learning using forward backward splitting, J. Mach. Learn. Res. 10 (2009) 2899–2934.
[17] A. Agarwal, P.L. Bartlett, P. Ravikumar, M.J. Wainwright, Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization, IEEE Trans. Inf. Theory 58 (5) (2012) 32–35.
[18] X. Gao, B. Jiang, S. Zhang, On the information-adaptive variants of the ADMM: an iteration complexity perspective, Optim. Online (2014).
[19] W. Zhong, J. Kwok, Fast stochastic alternating direction method of multipliers, in: Proceedings of the International Conference on Machine Learning, ICML, 2014, pp. 46–54.
[20] T. Suzuki, Stochastic dual coordinate ascent with alternating direction method of multipliers, in: Proceedings of the International Conference on Machine Learning, ICML, 2014, pp. 736–744.
[21] L. Qiao, T. Lin, Y.-G. Jiang, F. Yang, W. Liu, X. Lu, On stochastic primal-dual hybrid gradient approach for compositely regularized minimization, in: Proceedings of the European Conference on Artificial Intelligence, ECAI, vol.285, IOS Press, 2016, pp. 167–174.
[22] M.A. Noor, New extragradient-type methods for general variational inequalities, J. Math. Anal. Appl. 277 (2) (2003) 379–394.
[23] A. Nemirovski, Prox-method with rate of convergence o(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex–concave saddle point problems, SIAM J. Optim. 15 (1) (2004) 229–251.
[24] M.V. Solodov, B.F. Svaiter, A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator, Set Val. Anal. 7 (4) (1999) 323–345.
[25] R.D.C. Monteiro, B.F. Svaiter, On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean, SIAM J. Optim. 20 (6) (2010) 2755–2787.

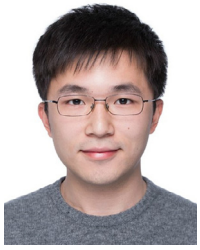
[26] R.D.C. Monteiro, B.F. Svaiter, Complexity of variants of Tseng's modified FB splitting and Korpelevich's methods for hemi-variational inequalities with applications to saddle-point and convex optimization problems, SIAM J. Optim. 21 (4) (2011) 1688–1720.
[27] R.D.C. Monteiro, B.F. Svaiter, Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers, SIAM J. Optim. 23 (1) (2013) 475–507.
[28] S. Bonettini, V. Ruggiero, An alternating extragradient method for total variation-based image restoration from poisson data, Inv. Prob. 27 (9) (2011) 095001.
[29] K. Scheinberg, S. Ma, D. Goldfarb, Sparse inverse covariance selection via alternating linearization methods, in: Proceedings of the Neural Information Processing Systems, NIPS, 2010, pp. 2101–2109.
[30] H. Wang, A. Banerjee, Online alternating direction method, in: Proceedings of the International Conference on Machine Learning, ICML, 2012, pp. 1119–1126.
[31] R. Johnson, T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction, in: Proceedings of the Neural Information Processing Systems, NIPS, 2013, pp. 315–323.



Tianyi Lin is a first-year Ph.D. student at UC Berkeley in Operations Research. His research is at the intersection of optimization, statistics and stochastic calculus. In specific, he focuses on the convergence property of several algorithms when applied to non-convex problems, and its statistical interpretation. He is also interested in applying the techniques of stochastic calculus in real scenarios. He earned a BASci in Mathematics Science, majoring in Pure Mathematics and Applied Mathematics from the Nanjing University, and a MS in Pure Mathematics and Statistics from University of Cambridge.



Linbo Qiao received the M.S. and B.S. degrees in computer science and technology from the National University of Defense Technology (NUDT), Changsha, China, in 2012 and 2010, respectively. Now, he is currently pursuing the Ph.D. degree in computer science and technology in NUDT. He worked as a Research Assistant with the Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, Hong Kong, from May 2014 to Oct. 2014. His research interests include structured sparse learning, online and distributed optimization.



Teng Zhang is currently a Ph.D. student in Management Science and Engineering at Stanford University. He received a B.Sc. degree in mathematics from Fudan University, Shanghai, China. His current research interests lie in the area of optimization and machine learning.

Jiashi Feng is currently an Assistant Professor in the Department of Electrical and Computer Engineering at National University of Singapore. He received his Ph.D. from National University of Singapore in 2014. Before joining NUS as a faculty, he was a postdoc research follow at UC Berkeley. His research areas include computer vision and machine learning. In particular, he is interested in object recognition, detection, segmentation, robust learning and deep learning.

Bofeng Zhang received the Ph.D. degree in computer science from the National University of Defense Technology (NUDT), Changsha, China, in 2007. He is currently an associate professor with College of Computer, National University of Defense Technology. His current research interests include computer networks, information security, data mining and machine learning algorithms.