J Optim Theory Appl (2018) 179:676–695
https://doi.org/10.1007/s10957-018-1270-0

Stochastic Accelerated Alternating Direction Method of Multipliers with Importance Sampling

Chenxi Chen[1] · **Yunmei Chen**[1] ·
Yuyuan Ouyang[2] · **Eduardo Pasiliao**[3]

Received: 11 August 2017 / Accepted: 21 March 2018 / Published online: 4 April 2018

Abstract In this paper, we incorporate importance sampling strategy into accelerated framework of stochastic alternating direction method of multipliers for solving a class of stochastic composite problems with linear equality constraint. The rates of convergence for primal residual and feasibility violation are established. Moreover, the estimation of variance of stochastic gradient is improved due to the use of important sampling. The proposed algorithm is capable of dealing with the situation, where the feasible set is unbounded. The experimental results indicate the effectiveness of the proposed method.

Keywords Stochastic ADMM · Duality gap · Variance estimation · Importance sampling

Mathematics Subject Classification 90C06 · 90C25 · 90C30

✉ Chenxi Chen
chenc@ufl.edu

Yunmei Chen
yun@ufl.edu

Yuyuan Ouyang
yuyuano@clemson.edu

Eduardo Pasiliao
pasiliao@eglin.af.mil

[1] Department of Mathematics, University of Florida, Gainesville, FL, USA

[2] Department of Mathematical Sciences, Clemson University, Clemson, SC, USA

[3] Munitions Directorate, Air Force Research Laboratory, AFB, Eglin, FL, USA

1 Introduction

The alternating direction method of multipliers (ADMM) is a simple and popular method for solving affine equality constrained composite problems. Combined with variable splitting, ADMM can solve these problems efficiently. Nowadays, due to the explosion in size and complexity of datasets, it is increasingly important to be able to efficiently solve these problems with a large number of training samples [1]. And this motivates the current development of stochastic ADMM-type algorithms [2–5], that use stochastic gradient to reduce the computational cost for estimating the gradient over the entire data set at each iteration. The convergence rates in these works depend on the estimation of variance of stochastic gradients. Recently, several works on randomized optimization show that the estimation of the convergence rate of stochastic algorithms can be improved by using importance sampling strategy, including randomized Kaczmarz method [6], randomized coordinate descent [7], stochastic gradient methods [8], stochastic average gradient method [9], stochastic mirror descent and stochastic dual coordinate ascent [10], etc.

In this paper, we propose an accelerated stochastic ADMM with importance sampling, name it as SAI, aiming at improving the estimation of the variance of stochastic gradients, and the practical performance. We will provide the convergence analysis for solving (1) over bounded or unbounded feasible sets. To our best knowledge, there has not been any discussion on the rate of convergence for stochastic ADMMs in the case of unbounded feasible set.

The outline of this paper is as follows. Section 2 introduces the problem formulation and related works. Section 3, the core section of this paper, describes the idea of importance sampling for stochastic ADMM and presents our algorithm and the convergence results. Section 4 gives the convergence analysis. The numerical experiments are presented in Sect. 5. Section 6 concludes this paper.

2 Problem Formulation and Related Works

In this paper, we consider the following class of affine equality constrained stochastic composite optimization (AECSCO) problems:

$$\begin{aligned} \min_{x\in X, y\in Y} \ & f(x) + g(y) := \mathbb{E}_\xi[F(x,\xi)] + g(y), \\ s.t. \ \ & Kx + By = c, \end{aligned} \tag{1}$$

where ξ is a random variable following some distribution, $f(x)$ is the expectation of $F(x,\xi)$, and for each ξ, $F(x,\xi)$ is closed and convex, $X \subseteq \mathbb{R}^n, Y \subseteq \mathbb{R}^m$ are closed convex sets, $f : X \to \mathbb{R}$ is a proper, convex, smooth function, and $\|\nabla f(x_1) - \nabla f(x_2)\| \le L\|x_1 - x_2\|$, for any $x_1, x_2 \in X$. For stochastic gradients, $\|\nabla_x F(x,\xi)\| \le J(\xi)$, for any $x \in X$, and $\sigma = \mathbb{E}[J(\xi)]$. $g : Y \to \mathbb{R}$ is a proper, convex, l.s.c. and simple function, $K : X \to \mathbb{R}^l$, $B : Y \to \mathbb{R}^l$ are bounded linear operators, $c \in \mathbb{R}^l$. For problem (1), it is equivalent to a saddle point problem as follows (see [11])

$$\min_{x,y} \max_{\lambda \in W} \Big[f(x) + g(Kx) - \langle \lambda, Kx + By - c \rangle \Big], \tag{2}$$

where $\lambda \in W$ is the dual variable. In one special case of the problem (1), where $B = -I$ and $c = 0$, the AECSCO problem reduces to the following unconstrained stochastic composite optimization (USCO) problem:

$$\min_{x \in X} f(x) + g(Kx) := \mathbb{E}_{\xi}[F(x, \xi)] + g(Kx). \tag{3}$$

Problems (1) and (3) have a wide range of applications in various fields, including machine learning [12,13], image processing [14,15], quantitative finance [16,17]. In those applications $f(x)$ represents an empirical loss function, while $g(y)$ enforces certain structured regularization, such as overlapped group lasso, total variation-based smoothing. To cope with the computational challenge in solving large-scale data analysis problems with structured sparsity constraints, much effort has been invested for developing stochastic or online ADMMs [2–5]. The key idea of those algorithms is to incorporate the basic stochastic optimization techniques into ADMM to significantly reduce each iteration cost. The work in [3] is one of the most original works in developing stochastic ADMM to solve (1). It employs a stochastic gradient in the linear approximation of $f(x)$ and uses the linearized ADMM scheme to update x. In [3] the accuracy of the approximation solution (x_t, y_t) is evaluated by the summation of the primal residual and feasibility violation, i.e., $f(x_t) + g(y_t) - (f^* + g^*) + \rho \|Kx_t + By_t - c\|$. Ouyang [3] achieves convergence rates of $O\left(\frac{1}{\sqrt{t}}\right)$ for general convex functions f and $O\left(\frac{L}{t}\right) + O\left(\frac{\sigma}{\sqrt{t}}\right)$ for L-smooth convex functions f, where σ^2 is an uniform upper bound for the variance of stochastic gradient of $f(x)$. If f is strongly convex, the convergence rate can be improved to $O\left(\frac{\log t}{t}\right)$. In [4] two online variants of ADMM: online proximal gradient descent-type method (OPG-ADMM) and regularized dual averaging-type method are presented for solving (3). The scheme of the OPG-ADMM is the same as the stochastic ADMM in [3]. In [2] two accelerated stochastic ADMM algorithms are proposed for solving (1). The first one uses the same update rule as that in [3], but replaces the output of the averaging iterates in the algorithm in [3] by weighted averaging iterates, which converge at a rate of $O\left(\frac{1}{t}\right)$ for strongly convex f. The second algorithm, namely optimal stochastic ADMM (OS-ADMM), incorporates Nesterov's multi-step acceleration techniques [18,19] into the stochastic ADMM in [3]. [2] achieves the convergence rate of $O\left(\frac{L}{t^2}\right) + O\left(\frac{\sigma}{\sqrt{t}}\right)$. The convergence analysis for OS-ADMM algorithm requires the assumption that for each iteration t, $f(x_t) \geq f(x^*)$, and $g(y_t) \geq g(y^*)$. It is noticeable that the $O\left(\frac{\sigma}{\sqrt{t}}\right)$ is a leading term in the convergence rates of stochastic ADMMs for solving (1). Hence, an appropriate estimation of variance of stochastic gradient will benefit the convergence.

It is also worth to mention that recently several variance-reduction methods for stochastic ADMM are developed in [20–23]. Those variance-reduction methods are designed specifically for a special case of (1), namely, the case when $f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x)$ is a finite sum of functions $f_i(x) := F(x, \xi_i)$. When the random variable ξ in (1) is discrete and takes only a finite number of distinct values $\xi_1, \ldots, \xi_N$ with uniform distribution, then we obtain the aforementioned special case as $f(x) = \mathbb{E}_\xi[F(x, \xi)] = \frac{1}{N}\sum_{i=1}^{N} F(x, \xi_i)$. However, our main focus in this paper is different from that in [20–23] for two reasons. First, our goal is to consider a more general stochastic optimization problem (1) in which the random variable ξ is not necessarily a discrete random variable. While it is possible to approximate $\mathbb{E}_\xi[F(x, \xi)]$ by $\frac{1}{N}\sum_{i=1}^{N} F(x, \xi_i)$ (known as the sample average approximation (see, e.g., [24,25]), the extra sample approximation error should also be taken into consideration. Instead of using sample average approximation and solving a finite sum optimization problem, the works in [2–4] and our proposed method follow the core concept of the robust stochastic approximation algorithms originated from [26] (see [26] also for a comparison between the sample average approximation and the stochastic approximation techniques). Second, for the finite sum special case of (1), the variance-reduction methods in [20–23] need to compute the gradient of $f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x) := \frac{1}{N}\sum_{i=1}^{N} F(x, \xi_i)$ from time to time. In particular, those randomized algorithms are divided into T epochs, and each epoch has $O(N)$ iterations. In each epoch, a computation of the full gradient $\nabla f = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla_x F(x, \xi_i)$ is required, and consequently, it is necessary to enumerate all the available samples $\xi_i, i = 1, \ldots, N$ [23]. The convergence analysis is based on the number of epochs, i.e., the number of evaluations of full gradients ∇f. Therefore, such algorithms are not applicable for the cases when the samples ξ_i could not be obtained all together, e.g., the online optimization case in which samples are streamed in a sequential fashion.

2.1 Contributions

The contribution of this paper mainly consists of three aspects. First, an algorithm of accelerated stochastic ADMM with importance sampling is proposed for solving AECSCO and USCO problems. By incorporating Nesterov's multi-step acceleration method into the stochastic ADMM as the OS-ADMM in [2], for L-smooth f, the SAI algorithm can achieve the rate of convergence $O\left(\frac{L}{t^2} + \frac{\sigma}{\sqrt{t}}\right)$ and $O\left(\sqrt{\frac{L}{t^3}} + \sqrt{\frac{\sigma}{t^{3/2}}}\right)$, in terms of the primal residue and feasible violation, respectively. For smooth f, the analysis of OS-ADMM in [2] studies the convergence rate in terms of the summation of the primal residue $f(x)+g(y)-(f^*+g^*)$ and feasibility violation $\|Kx+By-c\|$. However, it should be noted that the estimate of the summation of the primal residue and feasibility violation does not apply immediately to the primal residue and feasibility violation separately. In particular, the relation $f(x) + g(y) - (f^* + g^*) + \|Kx + By - c\| < \epsilon$ for small $\epsilon > 0$ does not necessarily imply small feasibility violation $\|Kx+By-c\| < \epsilon$, since the approximate solution (x, y) may have smaller objective function value than the optimal solution (i.e., $f(x)+g(y)-(f^*+g^*) < 0$) while not satisfying the constraint $Kx + By = c$. Our analysis is based on the estimation of the duality gap. Importantly, we do not require the assumption in [2] that for each iteration

t, $f(x_t) \geq f(x^*)$, and $g(y_t) \geq g(y^*)$, which in general, is too strong to be satisfied. Second, by incorporating important sampling to the algorithm, instead of uniform sampling, SAI improves the estimation of the variance of stochastic gradients. Finally, we are able to solve (1) and (3) over an unbounded convex set X, and achieve the same convergence rates as that for a bounded X. It is worth to mention that the convergence analysis in the existing stochastic ADMM algorithms requires the compactness of the feasible sets.

2.2 Notations

We assume that the optimal solution of (1) exists and is denoted as (x^*, y^*), and the optimal solution of (3) is x^*. We will use the following notations in this paper: $D_X = \sup_{x,x' \in X} \|x - x'\|$, $D_W = \sup_{\lambda,\lambda' \in W} \|\lambda - \lambda'\|$, $D_{\lambda^*} = \|\lambda^*\|$, $D_{x^*} = \|x_1 - x^*\|$, $D_{y^*} = \|y_1 - y^*\|$. In addition, $\|K\| = \sup_{\|x\|=1, x\in X} \|Kx\|$.

3 Stochastic Accelerated ADMM with Importance Sampling

In this section, we will present an accelerated stochastic ADMM with importance sampling (SAI) for solving (1) and (3).

3.1 Importance Sampling for Stochastic ADMM

The idea of importance sampling lies on a basic equality:

$$\mathbb{E}[h(\xi)] = \int h(s)p(s)\mathrm{d}s = \int \frac{1}{w(s)} h(s) \cdot w(s)p(s)\mathrm{d}s, \tag{4}$$

where ξ is a random variable with probability density function (PDF) $p(s)$, and $w(s)$ is any nonnegative function sharing the same support with $p(s)$, called the weight function. It can be viewed that we distribute certain weights for samples. If $\int w(s)p(s)\mathrm{d}s = 1$, i.e., $\mathbb{E}[w(\xi)] = 1$, the function $w(s)p(s)$ is a probability density function for a new distribution, denoted as $D^{(w)}$, then (4) turns to be $\mathbb{E}[h(\xi)] = \mathbb{E}^{(w)}\left[\frac{h(\xi)}{w(\xi)}\right]$. Thus, by multiplicative refinement, a stochastic estimation of $\nabla f(x)$ can be obtained through a different distribution $D^{(w)}$, which owns a different but may lower variance by choosing appropriate $w(s)$. In the following, we will discuss how to determine $w(s)$ for the proposed algorithm in a similar way suggested by [10].

For the problems of interest in (1) and (3), if ξ is sampled from the origin distribution, $\delta = \nabla_x F(x, \xi) - \nabla f(x)$, then an uniform bound for the variance is given by

$$\mathbb{E}\big[\|\nabla_x F(x, \xi) - \nabla f(x)\|^2\big] \leq \mathbb{E}\big[J^2(\xi)\big] - \|\nabla f(x)\|^2, \tag{5}$$

Alternatively, drawing the sample ξ through the weighted distribution $D^{(w)}$, the variance of stochastic gradient can be bounded by

$$\mathbb{E}^{(w)}\left[\left\|\frac{\nabla_x F(x,\xi)}{w(\xi)} - \nabla f(x)\right\|^2\right] \le \mathbb{E}\left[\frac{J(\xi)}{w(\xi)}\right]^2 - \|\nabla f(x)\|^2. \tag{6}$$

Practically, it is hard or impossible to evaluate the true variance, and the alternative is to use an appropriate upper bound to estimate the variance. Therefore, as suggested in (6), we should determine $w(s)$ such that $\mathbb{E}\left[\frac{J(\xi)}{w(\xi)}\right]^2$ would be minimized. Thus, it can be easily verified that $w(\xi) = \frac{J(\xi)}{\mathbb{E}[J(\xi)]}$ minimizes $\mathbb{E}\left[\frac{J(\xi)}{w(\xi)}\right]^2$, and further we have, for $\delta = \frac{\nabla_x F(x,\xi)}{w(\xi)} - \nabla f(x)$,

$$\mathbb{E}^{(w)}\big[\|\delta\|^2\big] \le \mathbb{E}\big[J(\xi)\big]^2 - \|\nabla f(x)\|^2. \tag{7}$$

In this paper, we take $\mathbb{E}\,[J(\xi)]^2$ as an uniform upper bound of $\mathbb{E}^{(w)}\big[\|\delta\|^2\big]$, comparing (6) and (7), we can see that $\mathbb{E}[J(\xi)]^2$ is a better choice than $\mathbb{E}\big[J^2(\xi)\big]$ for estimating the variance, because of $\mathbb{E}\big[J^2(\xi)\big] - \mathbb{E}\big[J(\xi)\big]^2 = \mathrm{Var}\big[J(\xi)\big]$. And for problems in which stochastic gradients own higher variance, it will benefit more.

3.2 Stochastic Accelerated ADMM with Importance Sampling

In this section, we present our algorithm in Algorithm 1. The random sampling in this section is according to the weighted distribution $D^{(w)}$ with weight function $w(\xi) = \frac{J(\xi)}{\mathbb{E}[J(\xi)]}$, and the expectation is with respect to $D^{(w)}$.

In Algorithm 1, we initialize (x_1, y_1, λ_1), s.t. $Kx_1 + By_1 - c = 0, \lambda_1 = 0$, and $(x_1^{\text{ag}}, y_1^{\text{ag}}, \lambda_1^{\text{ag}}) = (x_1, y_1, \lambda_1)$. In each iteration, a sample ξ_t will be extracted from the weighted distribution $D^{(w)}$, and $\frac{\nabla_x F(x_t^{md}, \xi_t)}{w(\xi_t)}$ serves as a stochastic gradient for updating x_{t+1} in (10).

For AECSCO problems, the compactness of feasible sets X or W is not required. The following theorem gives the convergence result for AECSCO problems.

Theorem 3.1 *Suppose that the total number of iterations N is given as a priori, and the parameters are set to $a_t = \frac{2}{t+1}$, $b_t = c_t = \frac{\rho N}{t}$, $\rho_t = \frac{\rho t}{N}$, $\eta_t = \frac{t}{2L + c\sigma N^{3/2}}$, where $c > 0$, $\rho > 0$ are constants, then we have*

Algorithm 1 Stochastic Accelerated ADMM with Importance Sampling

```
1: Choose $x_1 \in X$, $y_1 \in Y$, s.t. $Kx_1 + By_1 - c = 0$, $\lambda_1 = 0$. Set $x_1^{ag} = x_1$, $y_1^{ag} = y_1$, $\lambda_1^{ag} = \lambda_1$.
2: For $t = 1, \ldots, N-1$, do
       $\text{Sample } \xi_t \sim \mathrm{D}^{(\mathrm{w})};$   (8)
       $x_t^{md} = (1 - a_t)x_t^{\mathrm{ag}} + a_t x_t;$   (9)
       $x_{t+1} = \underset{x \in X}{\operatorname{argmin}} \langle \frac{\nabla_x F(x_t^{md}, \xi_t)}{w(\xi_t)} - K^T \lambda_t, x\rangle + \frac{b_t}{2}\|Kx + By_t - c\|^2 + \frac{1}{2\eta_t}\|x - x_t\|^2;$   (10)
       $y_{t+1} = \underset{y \in Y}{\operatorname{argmin}}\, g(y) - \langle \lambda_t, By\rangle + \frac{c_t}{2}\|Kx_{t+1} + By - c\|^2;$   (11)
       $\lambda_{t+1} = \lambda_t - \rho_t(Kx_{t+1} + By_{t+1} - c);$   (12)
       $x_{t+1}^{\mathrm{ag}} = (1 - a_t)x_t^{\mathrm{ag}} + a_t x_{t+1};$   (13)
       $y_{t+1}^{\mathrm{ag}} = (1 - a_t)y_t^{\mathrm{ag}} + a_t y_{t+1};$   (14)
       $\lambda_{t+1}^{\mathrm{ag}} = (1 - a_t)\lambda_t^{\mathrm{ag}} + a_t \lambda_{t+1};$   (15)
       $set\ t \leftarrow t + 1.$   (16)
3: Output $(x_N^{ag}, y_N^{ag})$.
```

$$
\begin{aligned}
&\mathbb{E}\left[f(x_N^{\mathrm{ag}}) + g(y_N^{\mathrm{ag}}) - (f^* + g^*)\right] \\
&\quad \le \frac{2LD_{x^*}^2}{N(N-1)} + \frac{\rho\|K\|^2 D_{x^*}^2}{N-1} + \left(\frac{2}{3c} + 2cD_{x^*}^2\right)\frac{\sigma}{\sqrt{N}},
\end{aligned} \tag{17}
$$

$$
\begin{aligned}
&\mathbb{E}\left[\|Kx_N^{\mathrm{ag}} + By_N^{\mathrm{ag}} - c\|\right] \\
&\quad \le \frac{4D_{x^*}\sqrt{L}}{N - 1\sqrt{\rho N}} + \frac{2\sqrt{2}(\|K\|D_{x^*} + D_{\lambda^*}/\rho)}{N-1} + \frac{(4 + 4\sqrt{3}cD_{x^*})\sqrt{\sigma}}{\sqrt{3c\rho(N-1)\sqrt{N}}}.
\end{aligned} \tag{18}
$$

Moreover, if X is compact, then, for any $\beta > 0$,

$$
\operatorname{Prob}\left[f(x_N^{\mathrm{ag}}) + g(y_N^{\mathrm{ag}}) - (f^* + g^*) \ge V_1(N)\right] \le \exp\left\{-\frac{\beta^2}{9}\right\} + \exp\{-\beta\}, \tag{19}
$$

$$
\operatorname{Prob}\left[\|Kx_N^{\mathrm{ag}} + By_N^{\mathrm{ag}} - c\| \ge V_2(N)\right] \le \exp\left\{-\frac{\beta^2}{9}\right\} + \exp\{-\beta\}, \tag{20}
$$

where

$$V_1(N) = \frac{2LD_X^2}{N(N-1)} + \frac{\rho\|K\|^2 D_X^2}{N-1} + \left(\frac{2}{3c} + 2cD_X^2\right)\frac{\sigma}{\sqrt{N}}, \tag{21}$$

$$V_2(N) = \frac{\sqrt{2L}D_X}{N} + \frac{\sqrt{\rho}\|K\|D_X}{\sqrt{N}} + \sqrt{\left(\frac{2}{3c} + 2cD_X^2\right)\frac{\sigma}{\sqrt{N}}} + \sqrt{\frac{N}{\rho}}D_{\lambda^*}, \tag{22}$$

$$V_0(N) = \frac{\sqrt{6}}{3}\beta D_X \sigma N\sqrt{N-1} + \frac{4+\beta}{3c}\sigma N^{1/2}(N-1). \tag{23}$$

For USCO problems, we assume that feasible sets X is compact, $g(y)$ is finite-valued and Lipschitz continuous on Y. Hence, $W = \text{dom } g^*$ is bounded (see, e.g., Corollary 13.3.3 in [27]). The convergence result for USCO problems is given in the following theorem.

Theorem 3.2 *Suppose that the parameters are set as follows*

$$a_t = \frac{2}{t+1}, \quad c_t = \rho_t = \rho, \quad b_t = \frac{(t-1)\rho}{t}, \quad \eta_t = \frac{t}{2L + c\sigma t^{3/2}}, \tag{24}$$

where $c > 0$ is a constant. Then, for any $\beta > 0, t > 1$, we have

$$\mathbb{E}\left[f(x_t^{\text{ag}}) + g(Kx_t^{\text{ag}}) - (f^* + g^*)\right] \le P(t), \tag{25}$$

$$\text{Prob}\left[f(x_t^{\text{ag}}) + g(Kx_t^{\text{ag}}) - (f^* + g^*) \ge Q(t)\right] \le \exp\left\{-\frac{\beta^2}{9}\right\} + \exp\{-\beta\}, \tag{26}$$

where

$$P(t) = \frac{2LD_X^2}{t(t-1)} + \frac{\rho\|K\|^2 D_X^2}{t} + \frac{D_W^2}{\rho t} + \left(\frac{1}{c} + cD_X^2\right)\frac{\sigma}{\sqrt{t-1}}, \tag{27}$$

$$Q(t) = Q_0(t) + \frac{2LD_X^2}{t(t-1)} + \frac{\rho\|K\|^2 D_X^2}{t} + \frac{D_W^2}{\rho t}, \tag{28}$$

$$Q_0(t) = \beta\frac{2\sqrt{6}D_X\sigma}{3\sqrt{t-1}} + (4+\beta)\frac{\sigma}{c\sqrt{t}}. \tag{29}$$

4 Convergence Analysis

In this section, we give proof for Theorems 3.1 and 3.2. The convergence analysis will be mainly based on duality gap function. We will firstly give the definition and some results for duality gap function.

4.1 Duality Gap Function

Definition 4.1 Denote $Z = X \times Y \times W$, for $z = (x, y, \lambda)$, $z' = (x', y', \lambda')$, the gap function $G(z; z')$ is defined by the following

$$G(z; z') = [f(x) + g(y) - \langle \lambda', Kx + By - c\rangle] - [f(x') + g(y') - \langle \lambda, Kx' + By' - c\rangle], \tag{30}$$

For problems with compact feasible set Z, i.e., both of X, Y, W are compact, the duality gap function is defined as

$$d(z) := \sup_{z' \in Z} G(z; z') = G(x, y, \lambda; x', y', \lambda'), \tag{31}$$

For unbounded and closed W, the duality gap function is

$$d_W(v, z) := \sup_{\lambda' \in W} \left\{ G(x, y, \lambda; x^*, y^*, \lambda') + \langle \lambda', v\rangle \right\}. \tag{32}$$

where $z = (x, y, \lambda) \in Z$, $v \in W$. In particular, if $W = \mathbb{R}^l$, we will omit the subscript W, and simply use the notation $d(v, z)$.

It is easy to verify that z^* is a saddle point of (2), i.f.f. $G(z; z^*) \geq 0$ (see [11]). The following proposition describes the relation between duality gap functions and the convergence criterion.

Proposition 4.1 *Suppose that x, y, λ are random variables, $(x, y, \lambda) \in Z$, then for any closed W, we have*

$$d_W(Kx + By - c, z) = f(x) + g(y) - (f^* + g^*). \tag{33}$$

Moreover, if $W = \mathbb{R}^l$, and $\mathbb{E}[d(v, z)] < \infty$, for some $v \in W$, then

$$f(x) + g(y) - (f^* + g^*) = d(v, z), \ almost\ surely, \tag{34}$$
$$Kx + By - c = v, \ almost\ surely. \tag{35}$$

In addition, if $\mathbb{E}[d(v, z)] \leq \epsilon$, and $\mathbb{E}[\|v\|] \leq \delta$, we have

$$\mathbb{E}[f(x) + g(y) - (f^* + g^*)] \leq \epsilon, \tag{36}$$
$$\mathbb{E}[\|Kx + By - c\|] \leq \delta. \tag{37}$$

Proof Following the definition in (32) and the fact that $Kx^* + By^* - c = 0$, it is easy to obtain (33). In addition, if $W = \mathbb{R}^l$, we have

$$d(v, z) = f(x) + g(y) - (f^* + g^*) + \sup_{\lambda' \in \mathbb{R}^l} \langle \lambda', v - (Kx + By - c)\rangle, \tag{38}$$

and we can see that $d(v, z) = \infty$ if $v \neq Kx + By - c$. Hence $d(v, z) < \infty$ if and only if $v = Kx + By - c$. Therefore, if $\mathbb{E}[d(v, z)] < \infty$, it implies that Prob $[d(v, z) < \infty] =$

1, and thus Prob $[v = Kx + By - c] = 1$. Also, we have that $f(x) + g(y) - (f^* + g^*) = d(v, z)$, a.s.. Moreover, if $\mathbb{E}[d(v, z)] \le \epsilon$, $\mathbb{E}[\|v\|] \le \delta$, it leads to (34) and (35). □

Now we are ready to give an important estimate related to gap function.

Lemma 4.1 *For* $\{z_t^{\mathrm{ag}}\}_{t\ge1}$ *generated by Algorithm 1,* $z = (x, y, \lambda) \in Z$*, and* $\eta_t^{-1} - La_t > 0$, $c_t \ge \rho_t$*, then we have*

$$\begin{aligned}
&\mathbb{E}\Big[G(z; z_{t+1}^{\mathrm{ag}}) - (1 - a_t)G(z; z_t^{\mathrm{ag}})\Big] \\
&\quad\le \mathbb{E}\Bigg[\frac{a_t\sigma^2}{2\left(\eta_t^{-1} - La_t\right)} + a_t\left(\frac{c_t}{\rho_t} - 1\right)\langle \lambda_t - \lambda_{t+1}, Kx + By - c\rangle \\
&\qquad + \frac{a_t}{2\eta_t}[\|x - x_t\|^2 - \|x - x_{t+1}\|^2] + \frac{a_t}{2\rho_t}[\|\lambda - \lambda_t\|^2 - \|\lambda - \lambda_{t+1}\|^2] \\
&\qquad + a_t\left[\frac{b_t}{2}\|Kx + By_t - c\|^2 - \frac{c_t}{2}\|Kx + By_{t+1} - c\|^2\right] \\
&\qquad + \frac{a_t(c_t - b_t)}{2}\|Kx_{t+1} - Kx\|^2\Bigg].
\end{aligned} \tag{39}$$

Proof By the definition of $G(\cdot;\cdot)$ in (30), and in view of (13)–(15), we have

$$\begin{aligned}
&G(z; z_{t+1}^{\mathrm{ag}}) - (1 - a_t)G(z; z_t^{\mathrm{ag}}) \\
&\quad= \Big[f(x_{t+1}^{\mathrm{ag}}) - (1 - a_t)f(x_t^{\mathrm{ag}}) - a_t f(x)\Big] + \Big[g(y_{t+1}^{\mathrm{ag}}) - (1 - a_t)g(y_t^{\mathrm{ag}}) - a_t g(y)\Big] \\
&\qquad + a_t\Big[\langle \lambda_{t+1}, Kx + By - c\rangle - \langle \lambda, Kx_{t+1} + By_{t+1} - c\rangle\Big].
\end{aligned} \tag{40}$$

Since $y_{t+1}^{\mathrm{ag}} = (1 - a_t)y_t^{\mathrm{ag}} + a_t y_{t+1}$, we have

$$\begin{aligned}
g(y_{t+1}^{\mathrm{ag}}) - (1 - a_t)g(y_t^{\mathrm{ag}}) - a_t g(y) &\le a_t(g(y_{t+1}) - g(y)) \\
&\le -\left\langle \left(\frac{c_t}{\rho_t} - 1\right)(\lambda_t - \lambda_{t+1}) - \lambda_{t+1}, B(y_{t+1} - y)\right\rangle,
\end{aligned} \tag{41}$$

where the first inequality is by the convexity of $g(y)$, and the second inequality follows the optimal condition in (11).

In view of the optimal condition in (10), we can see that for any $x \in X$,

$$\left\langle \frac{\nabla_x F(x_t^{md}, \xi_t)}{w(\xi_t)} + \frac{x_{t+1} - x_t}{\eta_t} + b_t K^T(Kx_{t+1} + By_t - c) - \lambda_t, x_{t+1} - x\right\rangle \le 0. \tag{42}$$

Combining (2.16) in [11], (41) and (42) with (40), also noticing the fact that $Kx_{t+1} + By_t - c = \dfrac{\lambda_t - \lambda_{t+1}}{\rho_t} + B(y_t - y_{t+1})$, and $\delta_t = \dfrac{\nabla_x F(x_t^{md}, \xi_t)}{w(\xi_t)} - \nabla f(x_t^{md})$, it yields

$$\begin{aligned}
&G(z; z_{t+1}^{\mathrm{ag}}) - (1-a_t)G(z; z_t^{\mathrm{ag}}) \\
&\quad \le a_t\Big[\langle \delta_t, x - x_{t+1}\rangle + \frac{1}{\eta_t}\langle x_t - x_{t+1}, x_{t+1} - x\rangle + \langle \lambda_{t+1} - \lambda, Kx_{t+1} + By_{t+1} - c\rangle \\
&\qquad - \Big\langle \Big(\frac{b_t}{\rho_t} - 1\Big)(\lambda_t - \lambda_{t+1}), K(x_{t+1} - x)\Big\rangle - \Big\langle \Big(\frac{c_t}{\rho_t} - 1\Big)(\lambda_t - \lambda_{t+1}), B(y_{t+1} - y)\Big\rangle \\
&\qquad + b_t\,\langle B(y_{t+1} - y_t), K(x_{t+1} - x)\rangle + \frac{La_t}{2}\|x_{t+1} - x_t\|^2\Big]. \qquad (43)
\end{aligned}$$

By the fact that $\langle b - c, c - a\rangle = \frac{1}{2}\Big[\|a-b\|^2 - \|a-c\|^2 - \|b-c\|^2\Big]$, we can obtain

$$\begin{aligned}
&\frac{1}{\eta_t}\langle x_t - x_{t+1}, x_{t+1} - x\rangle + \langle \lambda_{t+1} - \lambda, Kx_{t+1} + By_{t+1} - c\rangle \\
&\quad = \frac{1}{2\eta_t}\Big[\|x - x_t\|^2 - \|x - x_{t+1}\|^2 - \|x_t - x_{t+1}\|^2\Big] \qquad (44) \\
&\qquad + \frac{1}{2\rho_t}\Big[\|\lambda - \lambda_t\|^2 - \|\lambda - \lambda_{t+1}\|^2 - \|\lambda_t - \lambda_{t+1}\|^2\Big].
\end{aligned}$$

Also, it is easy to see that

$$\begin{aligned}
&-\Big\langle \Big(\frac{b_t}{\rho_t} - 1\Big)(\lambda_t - \lambda_{t+1}), K(x_{t+1} - x)\Big\rangle - \Big\langle \Big(\frac{c_t}{\rho_t} - 1\Big)(\lambda_t - \lambda_{t+1}), B(y_{t+1} - y)\Big\rangle \\
&\quad \le \Big(\frac{c_t}{\rho_t} - 1\Big)\langle \lambda_t - \lambda_{t+1}, Kx + By - c\rangle + \frac{c_t - b_t}{\rho_t}\langle \lambda_t - \lambda_{t+1}, K(x_{t+1} - x)\rangle. \qquad (45)
\end{aligned}$$

$$\begin{aligned}
&b_t\langle B(y_{t+1} - y_t), K(x_{t+1} - x_t)\rangle \\
&\quad \le \frac{b_t}{2}[\|Kx + By_t - c\|^2 - \|Kx + By_{t+1} - c\|^2] + \frac{b_t}{2\rho_t^2}\|Kx_{t+1} + By_{t+1} - c\|^2. \qquad (46)
\end{aligned}$$

For the term of $\langle \delta_t, x - x_{t+1}\rangle$, since $\eta_t^{-1} > La_t$ and $\mathbb{E}[\|\delta_t\|^2] \le \sigma^2$, we have

$$\mathbb{E}\Big[\langle \delta_t, x - x_{t+1}\rangle\Big] \le \mathbb{E}\Big[\frac{\sigma^2}{2\Big(\eta_t^{-1} - La_t\Big)} + \frac{\Big(\eta_t^{-1} - La_t\Big)}{2}\|x_t - x_{t+1}\|^2\Big]. \qquad (47)$$

where the inequality follows from Cauchy's inequality and $\mathbb{E}[\langle \delta_t, x - x_t\rangle] = 0$. Summing up (43)–(47), it immediately yields (39). □

Lemma 4.2 *For any $z = (x, y, \lambda) \in Z$, $a_t = \frac{2}{t+1}$, $A_t = (1 - a_t)A_{t-1}$, $A_1 = 1$, then for $\forall t > 1$, we have the following estimate*

$$
\begin{aligned}
&\mathbb{E}\Big[\frac{1}{A_{t-1}} G(z_t^{\mathrm{ag}}; z)\Big] \\
&\le \ \mathbb{E}\Bigg[\sum_{i=1}^{t-1} \frac{a_i \sigma^2}{2A_i\left(\eta_i^{-1} - La_i\right)} + \sum_{i=1}^{t-1} \frac{a_i}{2A_i\eta_i}\left[\|x - x_i\|^2 - \|x - x_{i+1}\|^2\right] \\
&\quad + \sum_{i=1}^{t-1} \frac{a_i}{2A_i}\left[b_i\|Kx + By_i - c\|^2 - c_i\|Kx + By_{i+1} - c\|^2\right] \\
&\quad + \sum_{i=1}^{t-1} \frac{a_i}{2A_i\rho_i}\left[\|\lambda - \lambda_i\|^2 - \|\lambda - \lambda_{i+1}\|^2\right] \\
&\quad + \sum_{i=1}^{t-1} \frac{a_i(c_i - b_i)}{2A_i}\|Kx_{i+1} - Kx\|^2 \\
&\quad + \sum_{i=1}^{t-1} \frac{a_i(c_i - \rho_i)}{2A_i\rho_i}\langle \lambda_i - \lambda_{i+1}, Kx + By - c\rangle\Bigg].
\end{aligned}
\tag{48}
$$

Proof We divide (39) by A_i on both sides, and take the summation from $i = 1$ to $t-1$, since $a_1 = 1$, it follows

$$
\sum_{i=1}^{t-1}\left[\frac{1}{A_i} G(z_{i+1}^{\mathrm{ag}}; z) - \frac{1 - a_i}{A_i} G(z_i^{\mathrm{ag}}; z)\right] = \frac{1}{A_{t-1}} G(z_t^{\mathrm{ag}}; z). \tag{49}
$$

In view of (39) and (49), it yields (48). □

It is easy to see that $\{\xi_t\}_{t\ge1}$ is an independent sampling-sequence, also $\{\langle \delta_t, x_t - x\rangle\}_{t\ge1}$ is a martingale difference sequence. Hence, the well-known large-deviation theorem for martingale difference sequences can be applied to get the estimate of tail probability [28].

Lemma 4.3 *Suppose that* $\mathbb{E}_{|t-1}[\psi_t] = 0$ *and* $\mathbb{E}_{|t-1}[\exp\{\psi_t^2/\sigma_t^2\}] \le \exp\{4\}$, *then, for any* $\beta > 0$, *we have*

$$
\operatorname{Prob}\left\{\sum_{t=1}^{N} \psi_t \ge \beta \sqrt{\sum_{t=1}^{N} \sigma_t^2}\right\} \le \exp\{-\beta^2/9\}. \tag{50}
$$

The proof of Lemma 4.3 is similar to Lemma 2 in [28].

Lemma 4.4 *If* X *is compact with diameter* D_X, *and* $\mathbb{E}[\exp\{\delta_t^2/\sigma^2\}] \le \exp\{4\}$, *for* $t > 0$, *then for* $\beta > 0$,

$$
\begin{aligned}
&\operatorname{Prob}\left\{\sum_{i=1}^{t-1} \frac{a_i}{A_i}\langle \sigma_i, x_i - x^*\rangle \ge \beta\sigma D_X \sqrt{\sum_{i=1}^{t-1} \frac{a_i^2}{A_i^2}}\right\} \le \exp\left\{-\frac{\beta^2}{9}\right\}, \\
&\operatorname{Prob}\left\{\sum_{i=1}^{t-1} \frac{a_i\|\delta_i\|^2}{A_i(\eta_i^{-1} - La_i)} \ge (4 + \beta)\sum_{i=1}^{t-1} \frac{a_i\sigma^2}{2A_i(\eta_i^{-1} - La_i)}\right\} \le \exp\{-\beta\}.
\end{aligned}
\tag{51}
$$

Proof Denote $\psi_i := \frac{a_i}{A_i}\langle \sigma_i, x_i - x^*\rangle$ and $\sigma_i := \frac{a_i}{A_i}\sigma D_X$, in view of Lemma 3.4, we can see that for any $\beta > 0$,

$$\text{Prob}\left\{\sum_{i=1}^{t-1}\frac{a_i}{A_i}\langle \sigma_i, x_i - x^*\rangle \geq \beta\sigma D_X\sqrt{\sum_{i=1}^{t-1}\frac{a_i^2}{A_i^2}}\right\} \leq \exp\left\{-\frac{\beta^2}{9}\right\}. \tag{52}$$

Moreover, denote $M_i = \frac{a_i}{A_i(\eta_i^{-1} - La_i)}$, and $M = \sum_{i=1}^{t-1} M_i$, then

$$\begin{aligned}&\text{Prob}\left\{\sum_{i=1}^{t-1} M_i\|\delta_i\|^2 \geq (4+\beta)M\sigma^2\right\}\\ &\quad\leq \text{Prob}\left\{\sum_{i=1}^{t-1}\frac{M_i}{M}\exp\{\|\delta_i\|^2/\sigma^2\} \geq \exp\{4+\beta\}\right\} \leq \exp\{-\beta\},\end{aligned}$$

where the first inequality is by the convexity of $\exp\{x\}$, and the last inequality is by $\mathbb{E}[\exp\{\delta_t^2/\sigma^2\}] \leq \exp\{4\}$ and Markov's inequality. □

4.2 Proof of Theorem 3.1

Proof Let (x^*, y^*) be a solution of (1), and $z^* = (x^*, y^*, \lambda)$, for any $\lambda \in \mathbb{R}^l$, according to the parameter settings in Theorem 3.1, the sequences $\left\{\frac{a_t}{2A_t\eta_t}\right\}$, $\left\{\frac{a_t b_t}{2A_t}\right\}$ are decreasing sequences, and $\left\{\frac{a_t}{2A_t\rho_t}\right\}$ is a constant sequence, then

$$\sum_{t=1}^{N-1}\frac{a_t}{2A_t\eta_t}\left[\|x^* - x_t\|^2 - \|x^* - x_{t+1}\|^2\right] \leq \frac{a_1}{2A_1\eta_1}D_{x^*}^2, \tag{53}$$

$$\sum_{t=1}^{N-1}\frac{a_t}{2A_t}\left[b_t\|By^* - By_t\|^2 - c_t\|By^* - By_{t+1}\|^2\right] \leq \frac{a_1 b_1}{2A_1}\|K\|^2 D_{x^*}^2, \tag{54}$$

where the equalities follow from Lemma 2.4 in [11], the optimal condition $Kx^* + By^* - c = 0$, and the initial condition $Kx_1 + By_1 - c = 0$. Also,

$$\sum_{t=1}^{N-1}\frac{a_t}{2A_t\rho_t}\left[\|\lambda - \lambda_t\|^2 - \|\lambda - \lambda_{t+1}\|^2\right] = \frac{a_1}{2A_1\rho_1}[\|\lambda - \lambda_1\|^2 - \|\lambda - \lambda_N\|^2]. \tag{55}$$

Combining (53)–(55) with (48), by the fact that $c_i = b_i$ and $\lambda_1 = 0$,

$$\begin{aligned} &\mathbb{E}\Big[\frac{1}{A_{N-1}}G(z_N^{\text{ag}}; z^*)\Big] \\ &\le \mathbb{E}\Big[\sum_{t=1}^{N-1}\frac{a_i\sigma^2}{2A_i(\eta_i^{-1}-La_i)} + \frac{a_1 D_{x^*}^2}{2A_1}\Big(\frac{1}{\eta_1}+b_1\|K\|^2\Big) \\ &\quad + \frac{a_1}{2A_1\rho_1}[\|\lambda-\lambda_1\|^2 - \|\lambda-\lambda_N\|^2]\Big]. \end{aligned} \tag{56}$$

In addition, we can see that $\|\lambda-\lambda_1\|^2 - \|\lambda-\lambda_N\|^2 \le 2\langle\lambda_N-\lambda_1, \lambda\rangle$, since $\lambda_1 = 0$. Then, in view of (32), we have

$$\begin{aligned} &\mathbb{E}\Big[d\Big(\frac{a_1 A_{N-1}}{A_1\rho_1}(\lambda_1-\lambda_N), z_N^{\text{ag}}\Big)\Big] \\ &\le A_{N-1}\mathbb{E}\Big[\sum_{t=1}^{N-1}\frac{a_i\sigma^2}{2A_i(\eta_i^{-1}-La_i)} + \frac{a_1 D_{x^*}^2}{2A_1}\Big(\frac{1}{\eta_1}+b_1\|K\|^2\Big)\Big]. \end{aligned} \tag{57}$$

Thus by Proposition 4.1, we have (17), and

$$Kx_N^{\text{ag}} + By_N^{\text{ag}} - c = \frac{a_1 A_{N-1}}{A_1\rho_1}(\lambda_1-\lambda_N) \text{ a.e..} \tag{58}$$

Since $\|\lambda_1-\lambda_N\|^2 \le 2\left(\|\lambda^*-\lambda_1\|^2 + \|\lambda^*-\lambda_N\|^2\right)$, $G(z_N^{\text{ag}}; x^*, y^*, \lambda^*) \ge 0$ (see [11]), and (56), it follows

$$\begin{aligned} &\mathbb{E}\Big[\frac{a_1}{4A_1\rho_1}\|\lambda_1-\lambda_N\|^2\Big] \\ &\le \mathbb{E}\Big[\sum_{t=1}^{N-1}\frac{a_i\sigma^2}{2A_i(\eta_i^{-1}-La_i)} + \frac{a_1 D_{x^*}^2}{2A_1}\Big(\frac{1}{\eta_1}+b_1\|K\|^2\Big) + \frac{a_1\|\lambda^*-\lambda_1\|^2}{A_1\rho_1}\Big]. \end{aligned} \tag{59}$$

Then, by (58) and direct computation, we obtain (18).

Moreover, if X is compact, it can be verified that $\mathbb{E}[\exp\{\delta_t^2/\sigma^2\}] \le \exp\{4\}$, applying Lemma 4.4 for (57), then, for $\beta > 0$,

$$\text{Prob}\left[d\Big(\frac{a_1 A_{N-1}}{A_1\rho_1}(\lambda_1-\lambda_N), z_N^{\text{ag}}\Big) \ge V_1(N)\right] \le \exp\left\{-\frac{\beta^2}{9}\right\} + \exp\{-\beta\}, \tag{60}$$

where $V_1(N)$ is given in (21), then by Proposition 4.1, we get (19).

Consider the feasibility violation, we can see that for any nonnegative $r_1, r_2, \ldots, r_s M_1 = \sum_{i=1}^s r_i$, $M_2 = \sum_{i=1}^s r_i^2$, $s > 1$, we have

$$\text{Prob}\Big[\|Kx_N^{\text{ag}} + By_N^{\text{ag}} - c\| \ge M_1\Big] \le \text{Prob}\left[\left(\frac{a_1 A_{N-1}}{A_1\rho_1}\right)^2\|\lambda_1-\lambda_N\|^2 \ge M_2\right].$$

Then applying Lemma 4.4 for (59), we conclude (20). □

4.3 Proof of Theorem 3.2

Proof Let (x^*, y^*) be a solution of (3), $\lambda \in W = \text{dom } g^*$, and $z^* = (x^*, y^*, \lambda)$, then by Lemma 4.1 and the parameter settings, $B = -I$, $Kx^* - y^* = 0$ and $c_t = \rho_t$, we have

$$\begin{aligned}
&\mathbb{E}^{(w)}\Big[\frac{1}{A_{t-1}}G(z_t^{\text{ag}}; z^*)\Big] \\
\leq\ & \mathbb{E}^{(w)}\Bigg[\sum_{i=1}^{t-1}\frac{a_i\sigma^2}{2A_i\left(\eta_i^{-1} - La_i\right)} + \sum_{i=1}^{t-1}\frac{a_i}{2A_i\eta_i}\left[\|x^* - x_i\|^2 - \|x^* - x_{i+1}\|^2\right] \\
&+\sum_{i=1}^{t-1}\frac{a_i}{2A_i}\left[b_i\|y^* - y_i\|^2 - c_i\|y^* - y_{i+1}\|^2\right] \\
&+\sum_{i=1}^{t-1}\frac{a_i}{2A_i\rho_i}\left[\|\lambda - \lambda_i\|^2 - \|\lambda - \lambda_{i+1}\|^2\right] \\
&+\sum_{i=1}^{t-1}\frac{a_i(c_i - b_i)}{2A_i}\|Kx^* - Kx^{i+1}\|^2\Bigg]. \qquad (61)
\end{aligned}$$

According to the parameter settings in Theorem 3.2, $\left\{\frac{a_t}{2A_t\eta_t}\right\}$ is an increasing sequence, then by Lemma 2.4 in [11],

$$\sum_{i=1}^{t-1}\frac{a_i}{2A_i\eta_i}\left[\|x^* - x_i\|^2 - \|x^* - x_{i+1}\|^2\right] \leq \frac{a_{t-1}}{2A_{t-1}\eta_{t-1}}D_X^2. \qquad (62)$$

Moreover, since $c_t \geq b_t$ and $b_1 = 0$, it is easy to see that

$$\sum_{t=1}^{t-1}\frac{a_i}{2A_i}\left[b_i\|y^* - y_i\|^2 - c_i\|y^* - y_{i+1}\|^2\right] \leq 0, \qquad (63)$$

$$\sum_{i=1}^{t-1}\frac{a_i(c_i - b_i)}{2A_i}\|Kx_{i+1} - Kx^*\|^2 \leq \sum_{i=1}^{t-1}\frac{a_i(c_i - b_i)\|K\|^2}{2A_i}D_X^2. \qquad (64)$$

Also, in view of (11)(12), Moreau decomposition (see, e.g., [29,30]), and (2.36) in [11], we have $\{\lambda_t\} \subseteq W$. And $\left\{\frac{a_t}{2A_t\rho_t}\right\}$ is an increasing sequence, then for any $\lambda \in W$,

$$\sum_{i=1}^{t-1}\frac{a_i}{2A_i\rho_i}\left[\|\lambda - \lambda_i\|^2 - \|\lambda - \lambda_{i+1}\|^2\right] \leq \frac{a_{t-1}}{2A_{t-1}\rho_{t-1}}D_W^2. \qquad (65)$$

Combing (62)–(65) with (61), we have

$$\begin{aligned}&\mathbb{E}^{(w)}\Big[\frac{1}{A_{t-1}}G(z_t^{\mathrm{ag}};z^*)\Big]\\&\le\sum_{i=1}^{t-1}\frac{a_i\sigma^2}{2A_i(\eta_i^{-1}-La_i)}+\frac{a_{t-1}}{2A_{t-1}\rho_{t-1}}D_W^2\\&\quad+\left(\sum_{i=1}^{t-1}\frac{a_i(c_i-b_i)\|K\|^2}{2A_i}+\frac{a_{t-1}}{2A_{t-1}\eta_{t-1}}\right)D_X^2.\end{aligned}\tag{66}$$

Since $g(y)$ is finite-valued and Lipschitz continuous, by Proposition 2.2 in [31], we have $f(x_t^{\mathrm{ag}})+g(Kx_t^{\mathrm{ag}})-\big(f(x^*)+g(y^*)\big)\le\sup_{\lambda\in W}G(z^*;z_t^{\mathrm{ag}})$. Then, it leads to (25). Applying Lemma 4.4 for (25), we can similarly obtain the probability estimation in (26). □

5 Numerical Experiments

In this paper, we conducted two experiments to examine the performance of SAI and compared with OS-ADMM in [2], which outperforms the stochastic ADMMs in [3,4].

5.1 Total Variation Regularized Linear Inversion

This experiment is on the following TV-regularized linear inversion problem:

$$\min_{x\in X}\ \mathbb{E}_i\Big[\frac{m}{2}\big(\langle A_i,x\rangle-b_i\big)^2\Big]+\gamma\|Dx\|_{2,1},\tag{67}$$

where A_i is the ith row of matrix A, i serves as the random variable with uniform distribution in $\{1,2,\ldots,m\}$, x is the image to be reconstructed, b_i represents the ith component of the observed data, and $\|Dx\|_{2,1}$ is the discrete form of TV semi-norm. The feasible set $X=\{x\in\mathbb{R}^{n\times n}:\|x\|\le D_X/2\}$. Problem (67) is a special case of (3), that $F(x,i)=\frac{m}{2}\big(\langle A_i,x\rangle-b_i\big)^2$, and $g(y)=\|y\|_{2,1}$, with the constraint $y=\gamma Dx$. We consider four instances: (1) A is generated by the normal distribution $N(0,1)$ with nonzero element density 0.2; (2) A is generated by $N(0,10)$ with density 0.2; (3) A is generated by $N(0,10)$ with density 0.1; (4) A is generated by the uniform distribution $U(0,1)$ and the norm of A_i is randomly scaled according to $U(0,10)$. Then, b is obtained by $b=Ax_{\mathrm{true}}+\epsilon$, where $x_{\mathrm{true}}\in\mathbb{R}^{n\times n}$ is generated from $N(0,1)$, and the noise $\epsilon\in\mathbb{R}^{m\times n}$ is generated from $N(0,0.001)$. We apply SAI and OS-ADMM to solve this problem. We set $m=5000$, $n=128$, and $\rho=1$. The performance of SAI and OS-ADMM in terms of relative error, primal objective value, and feasibility violation versus iteration number is shown in Figs. 1 and 2. The relative error of an approximate solution x is defined by $\frac{\|x-x_{\mathrm{true}}\|_2}{\|x_{\mathrm{true}}\|_2}$, and the feasibility violation is $\|y-\gamma Dx\|^2$. It can be observed from Figs. 1 and 2 that SAI outperforms OS-ADMM

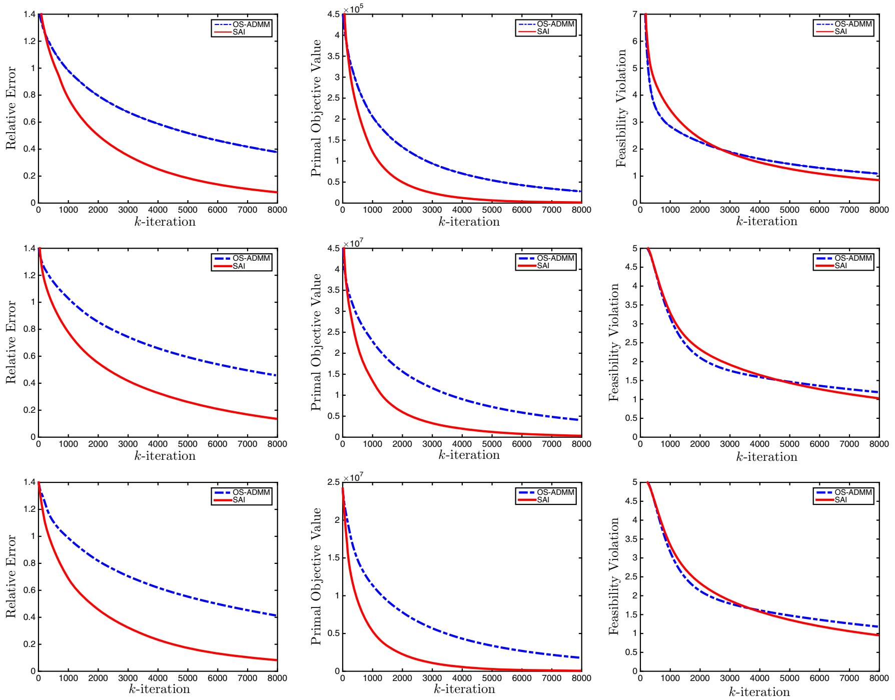

Fig. 1 Comparisons on TV-regularized linear inversion problem with A generating from: (1) N(0, 1) with sparsity density 0.2, (2) N(0, 10) with sparsity density 0.2, (3) N(0, 10) with sparsity density 0.1. Left: the relative errors from SAI and OS-ADMM. Middle: the objective function values from SAI and OS-ADMM. Right: the feasibility violations from SAI and OS-ADMM

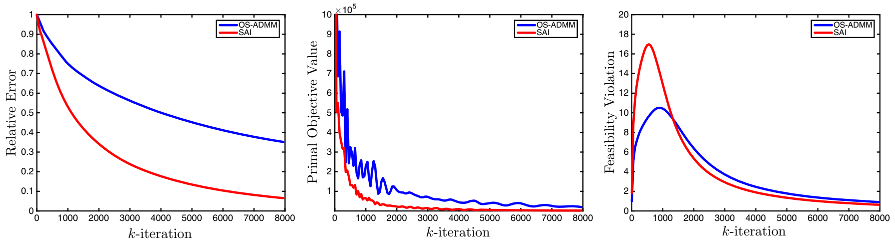

Fig. 2 Comparisons on TV-regularized linear inversion problem with A generating from U(0, 1) and the norm of A_i is subject to U(0, 10). Left: the relative errors from SAI and OS-ADMM. Middle: the objective function values from SAI and OS-ADMM. Right: the feasibility violations from SAI and OS-ADMM

in solving problem (67). This is consistent with our theoretical analysis results. For the case that A is generated from $U(0, 1)$ with the norm of A_i randomly scaled according to $U(0, 10)$, the results indicate that SAI not only accelerates the convergence, but also reduces the oscillations. We observed from the experiments that importance sampling can improve the performance of stochastic ADMM algorithms by increasing the step size and weakening the oscillations.

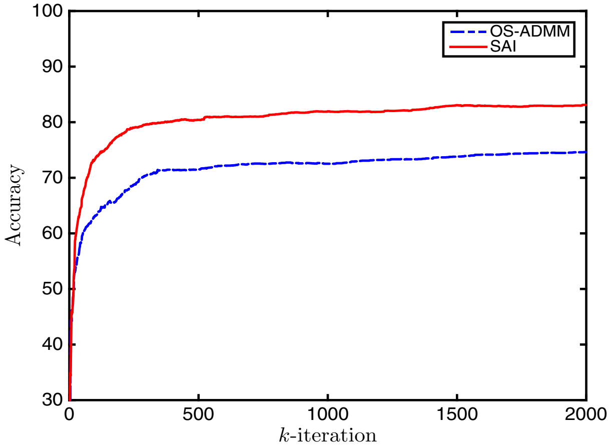

Fig. 3 Comparison of SAI and OS-ADMM on accuracies for multi-class classification

5.2 Graph-Guided Fused Lasso

The second experiment is on the graph-guided fused lasso (GFLasso) followed by the work in [2,3]. As a concrete example of generalized lasso, a graph-fused regularizer is introduced to enforce certain desired structure. Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ be a graph, where $\mathcal{V} = \{x_1, x_2, \ldots, x_n\}$ is the set of the vertices and $\mathcal{E}$ is the set of edges among $\mathcal{V}$. A weight w_{ij} is assigned for each edge $\{x_i, x_j\}$ in $\mathcal{E}$. The difference between x_i and x_j is penalized according to the edge weight w_{ij}, if there is such an edge. Then the GFLasso model for classification can be formulated as

$$\min_x \mathbb{E}_\xi[L(x,\xi)] + \gamma\|x\|_1 + \beta \sum_{\{i,j\}\in\mathcal{E}} w_{ij}|x_i - x_j|, \tag{68}$$

where $L(x,\xi) = \frac{1}{2}(l - x^T s)^2$ for feature-label pair (s, l) in training sample ξ, s is the feature vector and l is the label of sample ξ. We also consider the large-margin modification for (68), as introduced in [2], and reformulate the problem (68) to the form of (3) as follows:

$$\min_{x,y} \quad \mathbb{E}_\xi[L(x,\xi)] + \gamma\|x\|_2^2 + \beta\|y\|_1, \quad s.t. \;\; Fx = y, \tag{69}$$

where the matrix F satisfies that $\|Fx\|_1 = \sum_{\{i,j\}\in\mathcal{E}} w_{ij}|x_i - x_j|$.

We applied SAI and OS-ADMM algorithms to solve this problem and compared their performances. We used a public dataset on 20 newsgroups.[1] This dataset contains binary occurrences of 100 popular words for 16,242 newsgroup postings. All the samples are labeled into four categories: computer, recreation, science and talks. We do multi-class classification by one-vs-rest scheme. For each category, we trained our

[1] http://www.cs.nyu.edu/~roweis/data.html.

algorithm based on 80% of the total samples, and test prediction accuracy on the other 20%. Following the work in [2,3], in order to obtain F, we use the sparse inverse covariance selection method in [32] to get the sparsity pattern of the inverse covariance matrix, thus determine F. Figure 3 shows that SAI owns a faster convergence rate than OS-ADMM, also a better accuracy on test data. It is worth to mention that we can achieve satisfied accuracy by 2000 samples, only using 16% of the whole training dataset.

6 Conclusions

We propose a new accelerated stochastic ADMM with importance sampling that improves the convergence rate in terms of the dependence on the variance of stochastic gradients. The rates of convergence for the primal residual and feasibility violation are established. This algorithm can also solve problems with unbound feasible sets. The numerical experiments show that SAI outperforms the existing accelerated stochastic ADMM algorithm, OS-ADMM in several cases. This is more evident, if the norms of stochastic gradients are more oscillated.

Acknowledgements Funding was provided by National Science Foundation (Grant No. DMS 1719932).

References

1. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
2. Azadi, S., Sra, S.: Towards an optimal stochastic alternating direction method of multipliers. In: Proceedings of the 31st ICML, pp. 620–628 (2014)
3. Ouyang, H., He, N., Tran, L., Gray, A.: Stochastic alternating direction method of multipliers. In: Proceedings of the 30th ICML, pp. 80–88 (2013)
4. Suzuki, T.: Dual averaging and proximal gradient descent for online alternating direction multiplier method. In: Proceedings of the 30th ICML, pp. 392–400 (2013)
5. Wang, H., Banerjee, A.: Online alternating direction method. arXiv preprint arXiv:1306.3721 (2013)
6. Strohmer, T., Vershynin, R.: A randomized Kaczmarz algorithm with exponential convergence. J. Fourier Anal. Appl. **15**(2), 262–278 (2009)
7. Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J. Optim. **22**(2), 341–362 (2012)
8. Needell, D., Srebro, N., Ward, R.: Stochastic gradient descent and the randomized Kaczmarz algorithm. arXiv preprint arXiv:1310.5715 (2013)
9. Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Math. Program. **162**, 83–112 (2013)
10. Zhao, P., Zhang, T.: Stochastic optimization with importance sampling for regularized loss minimization. In: International Conference on Machine Learning, pp. 1–9 (2015)
11. Chen, Y., Lan, G., Ouyang, Y.: Optimal primal-dual methods for a class of saddle point problems. SIAM J. Optim. **24**(4), 1779–1814 (2014)
12. Jacob, L., Obozinski, G., Vert, J.P.: Group lasso with overlap and graph lasso. In: Proceedings of the 26th ICML, pp. 433–440. ACM (2009)
13. Tomioka, R., Hayashi, K., Kashima, H.: Estimation of low-rank tensors via convex optimization. arXiv preprint arXiv:1010.0789 (2010)
14. Goldstein, T., Osher, S.: The split bregman method for l1-regularized problems. SIAM J. Imaging Sci. **2**(2), 323–343 (2009)
15. Goldfarb, D., Ma, S., Scheinberg, K.: Fast alternating linearization methods for minimizing the sum of two convex functions. Math. Program. **141**, 349–382 (2013)

16. Touzi, N.: Stochastic control and application to finance. Pisa. Special Research Semester on Financial Mathematics, Scuola Normale Superiore (2002)
17. Ziemba, W.T., Vickson, R.G.: Stochastic Optimization Models in Finance. World Scientific, Singapore (1975)
18. Nesterov, Y.: A method of solving a convex programming problem with convergence rate o (1/k2). Sov. Math. Doklady **27**(2), 372–376 (1983)
19. Nesterov, Y.: Introductory Lectures on Convex Optimization, vol. 87. Springer, Berlin (2004)
20. Zhao, S.Y., Li, W.J., Zhou, Z.H.: Scalable stochastic alternating direction method of multipliers. arXiv preprint arXiv:1502.03529 (2015)
21. Zhang, C., Shen, Z., Qian, H., Zhou, T.: Accelerated stochastic ADMM with variance reduction. arXiv preprint arXiv:1611.04074 (2016)
22. Zheng, S., Kwok, J.T.: Stochastic variance-reduced ADMM. arXiv preprint arXiv:1604.07070 (2016)
23. Liu, Y., Shang, F., Cheng, J.: Accelerated variance reduced stochastic ADMM. In: AAAI, pp. 2287–2293 (2017)
24. Shapiro, A.: Monte carlo sampling methods. Handb. Oper. Res. Manag. Sci. **10**, 353–425 (2003)
25. Shapiro, A., Nemirovski, A.: On complexity of stochastic programming problems. In: Continuous Optimization, pp. 111–146. Springer (2005)
26. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
27. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
28. Lan, G., Nemirovski, A., Shapiro, A.: Validation analysis of mirror descent stochastic approximation method. Math. Program. **134**(2), 425–458 (2012)
29. Esser, E., Zhang, X., Chan, T.F.: A general framework for a class of first order primal–dual algorithms for convex optimization in imaging science. SIAM J. Imaging Sci. **3**(4), 1015–1046 (2010)
30. Moreau, J.J.: Décomposition orthogonale dun espace hilbertien selon deux cônes mutuellement polaires. CR Acad. Sci. Paris **255**, 238–240 (1962)
31. Ouyang, Y., Chen, Y., Lan, G., Pasiliao Jr, E.: An accelerated linearized alternating direction method of multipliers. arXiv preprint arXiv:1401.6607 (2014)
32. Banerjee, O., Ghaoui, L.E., dAspremont, A.: Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. J. Mach. Learn. Res. **9**(Mar), 485–516 (2008)