



# Adaptive primal-dual stochastic gradient method for expectation-constrained convex stochastic programs

Yonggui Yan<sup>1</sup> · Yangyang Xu<sup>1</sup>

Received: 20 December 2020 / Accepted: 10 December 2021

© Springer-Verlag GmbH Germany, part of Springer Nature and Mathematical Optimization Society 2022

## Abstract

Stochastic gradient methods (SGMs) have been widely used for solving stochastic optimization problems. A majority of existing works assume no constraints or easy-to-project constraints. In this paper, we consider convex stochastic optimization problems with expectation constraints. For these problems, it is often extremely expensive to perform projection onto the feasible set. Several SGMs in the literature can be applied to solve the expectation-constrained stochastic problems. We propose a novel primal-dual type SGM based on the Lagrangian function. Different from existing methods, our method incorporates an adaptiveness technique to speed up convergence. At each iteration, our method inquires an unbiased stochastic subgradient of the Lagrangian function, and then it renews the primal variables by an adaptive-SGM update and the dual variables by a vanilla-SGM update. We show that the proposed method has a convergence rate of  $O(1/\sqrt{k})$  in terms of the objective error and the constraint violation. Although the convergence rate is the same as those of existing SGMs, we observe its significantly faster convergence than an existing non-adaptive primal-dual SGM and a primal SGM on solving the Neyman–Pearson classification and quadratically constrained quadratic programs. Furthermore, we modify the proposed method to solve convex–concave stochastic minimax problems, for which we perform adaptive-SGM updates to both primal and dual variables. A convergence rate of  $O(1/\sqrt{k})$  is also established to the modified method for solving minimax problems in terms of primal-dual gap. Our code has been released at <https://github.com/RPI-OPT/APriD>.

**Keywords** Stochastic gradient method · Adaptive methods · Expectation-constrained stochastic optimization · Saddle-point problem

**Mathematics Subject Classification** 90C15 · 90C52 · 49M37 · 65K05

---

✉ Yangyang Xu  
xuy21@rpi.edu

Yonggui Yan  
yany4@rpi.edu

<sup>1</sup> Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

# 1 Introduction

The stochastic approximation method can trace back to [28], where a root-finding problem is considered. Stochastic gradient methods (SGMs), as first-order stochastic approximation methods, have attracted a lot of research interests from both theory and applications. In the literature, a majority of works on SGMs focus on problems without constraints or with easy-to-project constraints.

In this work, we consider expectation-constrained convex stochastic programs, which are formulated as

$$\begin{aligned} \min_{\mathbf{x} \in X} f_0(\mathbf{x}) &\equiv \mathbb{E}_{\xi_0}[F_0(\mathbf{x}; \xi_0)], \\ \text{s.t. } f_i(\mathbf{x}) &\equiv \mathbb{E}_{\xi_i}[F_i(\mathbf{x}; \xi_i)] \leq 0, \quad i \in [M]. \end{aligned} \quad (1.1)$$

Here,  $[M]$  denotes  $\{1, \dots, M\}$ ,  $X$  is a compact convex set in  $\mathbb{R}^n$  that admits an easy projection,  $\xi_i$  is a random variable and  $f_i$  is a convex function on  $X$  for each  $i = 0, 1, \dots, M$ , and  $\mathbb{E}_{\xi_i}$  takes the expectation about  $\xi_i$ . When  $\xi_i$  is distributed uniformly on a finite set  $\{\xi_{i1}, \dots, \xi_{iN_i}\}$  for each  $i = 0, 1, \dots, M$ , (1.1) reduces to the finite-sum structured function-constrained problem

$$\begin{aligned} \min_{\mathbf{x} \in X} f_0(\mathbf{x}) &\equiv \frac{1}{N_0} \sum_{j=1}^{N_0} F_0(\mathbf{x}; \xi_{0j}) \\ \text{s.t. } f_i(\mathbf{x}) &\equiv \frac{1}{N_i} \sum_{j=1}^{N_i} F_i(\mathbf{x}; \xi_{ij}) \leq 0, \quad i \in [M], \end{aligned} \quad (1.2)$$

where each  $N_i$  can be a big integer number.

Due to the possible nonlinearity of  $\{f_i\}_{i=1}^M$ , the expectation format, and the possibly big numbers  $\{N_i\}_{i=1}^M$  and/or  $M$ , it can be extremely expensive to perform projection onto the feasible set of (1.1) or (1.2). Toward finding a solution of (1.1) or (1.2) with a desired accuracy, we will design a primal-dual SGM that does not require the expensive projection. Our method will only require unbiased stochastic subgradients of the Lagrangian function of (1.1) or (1.2) and the weighted projection onto  $X$ . To have fast convergence, we will incorporate into the designed method an adaptiveness technique, which has been extensively used in solving unconstrained problems, such as training deep learning models.

## 1.1 Motivating applications

Many applications can be formulated into (1.1) or (1.2). We give a few concrete examples below.

*Neyman–Pearson classification* One example of (1.1) is the Neyman–Pearson classification (NPC) [27,32]. For a binary classification problem, a classifier  $h(\mathbf{x}; \mathbf{a})$ , parameterized by  $\mathbf{x}$ , is learned by finding an optimal  $\mathbf{x}$  from a set of training data.

Each training data point  $\mathbf{a}$  has a label  $b \in \{+1, -1\}$ . The classifier categorizes a new sample  $\mathbf{a}$  to the “+1” class if  $h(\mathbf{x}, \mathbf{a}) \geq 0$  and “-1” otherwise. The prediction incurs two types of errors: type-I error and type-II error. The former is also called false-positive error, defined as

$$R^-(\mathbf{x}) = \mathbf{Prob}(h(\mathbf{x}; \mathbf{a}) \geq 0 \mid b = -1) = \mathbb{E}[\iota_{\geq 0}(h(\mathbf{x}; \mathbf{a})) \mid b = -1]$$

and the latter is also called false-negative error, defined as

$$R^+(\mathbf{x}) = \mathbf{Prob}(-h(\mathbf{x}; \mathbf{a}) > 0 \mid b = +1) = \mathbb{E}[\iota_{\geq 0}(-h(\mathbf{x}; \mathbf{a})) \mid b = +1].$$

Here,  $b$  denotes the true label of  $\mathbf{a}$ , and  $\iota_{\geq 0}$  denotes the indicator function of nonnegative numbers, i.e.,  $\iota_{\geq 0}(z) = 1$  if  $z \geq 0$  and  $\iota_{\geq 0}(z) = 0$  if  $z < 0$ . In certain applications, the cost of making these two types of error could be severely different. For example, in medical diagnosis, missing a malignant tumor is much more concerned than diagnosing a benign tumor to a malignant one. Hence, for these applications, it would be beneficial to minimize the more concerned type of error while controlling another type of error at an acceptable level. NPC aims at minimizing the type-II error by controlling the type-I error. It searches for the best parameter  $\mathbf{x}$  in a given space  $X$  by solving the error-constrained problem:

$$\min_{\mathbf{x} \in X} R^+(\mathbf{x}), \quad \text{s.t.} \quad R^-(\mathbf{x}) \leq c, \quad (1.3)$$

where  $c > 0$  is a user-specified level of error. Due to the nonconvexity of the indicator function  $\iota_{\geq 0}(z)$ , the problem in (1.3) is often computationally intractable. Hence, it is common (cf. [27]) to relax the indicator function  $\iota_{\geq 0}(z)$  by a convex non-decreasing surrogate  $\varphi(z) : \mathbb{R} \rightarrow \mathbb{R}^+$ . This way, if  $h(\mathbf{x}; \mathbf{a})$  is affine with respect to  $\mathbf{x}$  in a convex set  $X$ , we can obtain a relaxed convex stochastic program in the form of (1.1) by relaxing  $\iota_{\geq 0}(z)$  to  $\varphi(z)$  with  $z = h(\mathbf{x}; \mathbf{a})$  and  $z = -h(\mathbf{x}; \mathbf{a})$  respectively in the formulations of  $R^-$  and  $R^+$ .

**Fairness-constrained classification** The data involved in a classification problem may contain certain sensitive features, such as gender and race. Fairness-constrained classification [43,44] aims at finding a classifier that is fair to the samples with different sensitive features. For example, consider the binary classification problem. Let  $(\mathbf{a}, z)$  denote a data point, following a certain distribution and having a label  $b \in \{+1, -1\}$ . Here,  $z \in \{0, 1\}$  represents a unidimensional binary sensitive feature. Suppose  $h(\mathbf{x}; \mathbf{a}, z)$  is a classifier, parameterized by  $\mathbf{x}$ , such that a given sample  $(\mathbf{a}, z)$  is predicted to class  $\hat{b} = \text{sign}(h(\mathbf{x}; \mathbf{a}, z))$ . One can learn a fair classifier  $h(\mathbf{x}; \mathbf{a}, z)$  by solving the problem:

$$\min_{\mathbf{x} \in X} L(\mathbf{x}), \quad \text{s.t.} \quad \mathbf{Prob}(\hat{b} \neq b \mid z = 0) = \mathbf{Prob}(\hat{b} \neq b \mid z = 1), \quad (1.4)$$

where  $L(\mathbf{x})$  measures a certain expected loss, e.g.,  $L(\mathbf{x}) = \mathbb{E}[\log(1 + \exp(-b \cdot h(\mathbf{x}; \mathbf{a}, z)))]$ , and  $X$  is a convex set. The constraint in (1.4) is one kind of fairness constraint. More kinds of fairness constraints can be found in [43]. Roughly speaking,

a fairness constraint enforces that the classifier should not be correlated to the sensitive features. In general, the constraint in (1.4) is nonconvex. [43] introduces a relaxation to (1.4) by using an empirical estimation of the covariance between  $z$  and  $h(\mathbf{x}; \mathbf{a}, z)$ , which results in the following problem:

$$\min_{\mathbf{x} \in X} \tilde{L}(\mathbf{x}), \quad \text{s.t.} \quad -c \leq \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})h(\mathbf{x}; \mathbf{a}_i, z_i) \leq c. \quad (1.5)$$

Here,  $\tilde{L}$  is an empirical loss,  $\bar{z} = \mathbb{E}[z]$ ,  $\{(\mathbf{a}_i, z_i)\}_{i=1}^n$  is a set of training samples, and  $c > 0$  is a given threshold that trades off accuracy and unfairness. If  $h(\mathbf{x}; \mathbf{a}, z)$  is affine about  $\mathbf{x}$ , then the constraint in (1.5) will be convex about  $\mathbf{x}$ , and in this case, (1.5) with a convex loss function  $\tilde{L}(\mathbf{x})$  becomes a convex problem in the form of (1.2).

*Chance-constrained problems* Another example is the chance-constrained problem. It can be modeled as (cf. [33])

$$\min_{\mathbf{x} \in X} f(\mathbf{x}), \quad \text{s.t.} \quad \mathbf{Prob}\{F(\mathbf{x}, \xi) \leq 0\} \geq p, \quad (1.6)$$

where  $\xi \in \Xi$  is a random vector,  $X$  is a convex set, and  $p \in (0, 1)$  is a user-specified probability value. Notice  $\mathbf{Prob}\{F(\mathbf{x}, \xi) \leq 0\} = 1 - \mathbf{Prob}\{F(\mathbf{x}, \xi) > 0\} = 1 - \mathbb{E}[\iota_{>0}(F(\mathbf{x}, \xi))]$ , where  $\iota_{>0}$  denotes the indicator function of positive numbers. We can rewrite (1.6) as

$$\min_{\mathbf{x} \in X} f(\mathbf{x}), \quad \text{s.t.} \quad \mathbb{E}[\iota_{>0}(F(\mathbf{x}, \xi))] \leq 1 - p. \quad (1.7)$$

One approach to solve (1.7) is the sample average approximation method [22,26]. It approximates the expectation in the constraint by the empirical mean and forms the approximation problem:

$$\min_{\mathbf{x} \in X} f(\mathbf{x}), \quad \text{s.t.} \quad \frac{1}{N} \sum_{i=1}^N \iota_{>0}(F(\mathbf{x}, \xi_i)) \leq 1 - \hat{p}, \quad (1.8)$$

where  $\hat{p}$  is carefully selected such that the solution of (1.8) can satisfy certain desired property. When  $f$  is convex,  $F$  is linear about  $\mathbf{x}$ , and the indicator function  $\iota_{>0}$  is relaxed by a convex surrogate, (1.7) and (1.8) are convex in the form of (1.1) and (1.2) respectively.

Another approach to solve (1.6) is the scenario approximation [3,4], which takes  $M$  independent samples  $\xi_1, \xi_2, \xi_3, \dots, \xi_M$  of  $\xi$  and forms the problem

$$\min_{\mathbf{x} \in X} f(\mathbf{x}), \quad \text{s.t.} \quad F(\mathbf{x}, \xi_i) \leq 0, i \in [M]. \quad (1.9)$$

When  $f(\mathbf{x})$  and  $F(\mathbf{x}, \xi)$  are convex about  $\mathbf{x} \in X$  for all  $\xi$ , (1.9) becomes one instance of (1.1) with  $f_i(\mathbf{x}) = F(\mathbf{x}, \xi_i)$  for each  $i$ .

*More* More examples of (1.1) include the risk averse optimization (cf. [33, Chapter 6]), asset-allocation problems (cf. [29]), and the monotonicity constrained machine learning [11].

## 1.2 Related work

Several existing methods can be applied to solve (1.1) or (1.2). The method that we will propose is of primal-dual type and adopts an adaptiveness technique to have faster convergence. Below we review the existing works that are closely related to ours.

### 1.2.1 Existing methods for solving (1.1) or (1.2)

Lan and Zhou [18] proposes a cooperative stochastic approximation (CSA) method for stochastic programs that have one function constraint. Let  $g(\mathbf{x}) = \sum_{i=1}^M [f_i(\mathbf{x})]_+$ . The constraint  $\{\mathbf{x} : f_i(\mathbf{x}) \leq 0, \forall i \in [M]\}$  can be reformulated to  $g(\mathbf{x}) \leq 0$ . Hence, CSA can be applied to (1.1) with this reformulation. At the  $k$ -th iterate  $\mathbf{x}^k$  for all  $k \geq 0$ , CSA first calculates an unbiased estimator  $\widehat{G}_k$  of  $g(\mathbf{x}^k)$ . Then depending on whether  $\widehat{G}_k \leq \eta_k$  or not for some tolerance  $\eta_k > 0$ , it moves either along the negative direction of an unbiased stochastic approximation of the subgradient  $\tilde{\nabla} f_0(\mathbf{x}^k)$  or  $\tilde{\nabla} g(\mathbf{x}^k)$ . Given a maximum iteration  $K$ , CSA takes a weighted average of all iterates indexed by  $\mathcal{B}^K = \{s \leq k \leq K \mid \widehat{G}_k \leq \eta_k\}$  for some  $1 \leq s \leq K$ . An ergodic convergence of  $O(1/\sqrt{K})$  can be shown for CSA under the convexity assumption.

When  $M$  is big in (1.1) and  $f_i$  is a deterministic function for each  $i \in [M]$ , one can apply the primal-dual SGM in [38] or its adaptive variant. The method in [38] is derived by using the augmented Lagrangian function. At the  $k$ -th iteration, it inquires an unbiased stochastic subgradient  $(\mathbf{u}^k, \mathbf{w}^k)$  of the augmented Lagrangian function and then performs the updates

$$\mathbf{x}^{k+1} = \text{Proj}_{X, D_k}(\mathbf{x}^k - D_k^{-1} \mathbf{u}^k), \quad \mathbf{z}^{k+1} = \mathbf{z}^k + \rho_k \mathbf{w}^k, \quad (1.10)$$

where  $\rho_k > 0$  is a dual step size, and  $D_k$  is a positive definite matrix. Two settings of  $D_k$  are provided and analyzed in [38]. One is a non-adaptive setting with  $D_k = \frac{I}{\alpha_k}$ , where  $I$  is the identity matrix and  $\alpha_k > 0$ . The other is an adaptive setting with  $D_k = \text{diag}(\mathbf{v}^k) + \frac{I}{\alpha_k}$  and  $\mathbf{v}^k = \eta \sqrt{\sum_{t=1}^k \frac{(\mathbf{u}^t)^2}{\gamma_t^2}}$ , where  $\alpha_k > 0$ ,  $\eta > 0$  and  $\gamma_t = \max\{1, \|\mathbf{u}^t\|\}$ .

Under both settings, the method has an ergodic convergence rate of  $O(1/\sqrt{k})$  for the convex case. Notice that for the expectation or finite-sum constrained problems, the augmented Lagrangian function will have the compositional expectation or finite-sum terms, which prohibit an easy access to an unbiased stochastic subgradient [37,41]. Hence, the method in [38] will not apply to (1.1) with expectation constraints or (1.2) with big numbers  $\{N_i\}_{i=1}^M$ .

If an easy projection can be performed onto  $X_i = \{\mathbf{x} \in \mathbb{R}^n : f_i(\mathbf{x}) \leq 0\}$  for each  $i \in [M]$ , one can also apply the random projection methods [35,36] to (1.1), and in addition, if the exact gradient of  $f_0$  can be obtained, the stochastic proximal-proximal method in [30] can also be applied. However, if  $X_i, i \in [M]$  are not simple sets

such as the logistic loss function induced constraint set in NPC [27], these random projection methods will be inefficient. For the finite-sum constrained problem (1.2), level-set based methods are studied in [19]. For the special case of (1.1) where exact gradients of all the functions  $\{f_i\}_{i=0}^M$  can be accessed, deterministic first-order methods have been studied in several works, e.g., [2, 20, 21, 39, 40, 42]. However, for (1.2) with big numbers  $\{N_i\}_{i=0}^N$ , these deterministic methods would be inefficient, as computing exact gradients at each iteration is very expensive.

Deterministic or stochastic gradient methods (e.g., [13–15, 24, 46]) have also been studied for saddle-point problems, which can include the function-constrained problems as special cases if certain constraint qualifications hold. For example, one can solve the problem in (1.1) by the saddle-point mirror stochastic approximation (MSA), which is proposed by Nemirovski et al. [24] to solve convex–concave stochastic saddle-point problems. Assuming the strong duality, we can reformulate (1.1) as a saddle-point problem by using the Lagrangian function  $\mathcal{L}(\mathbf{x}, \mathbf{z})$ ; see (2.1) below. Applying MSA to the reformulation will yield the following iterative update:

$$\begin{pmatrix} \mathbf{x}^{k+1} \\ \mathbf{z}^{k+1} \end{pmatrix} = \text{Proj}_{X \times Z} \left[ \begin{pmatrix} \mathbf{x}^k \\ \mathbf{z}^k \end{pmatrix} - \gamma_k \begin{pmatrix} \mathbf{u}^k \\ -\mathbf{w}^k \end{pmatrix} \right], \quad (1.11)$$

where  $(\mathbf{u}^k, \mathbf{w}^k)$  is an unbiased stochastic subgradient of  $\mathcal{L}(\mathbf{x}^k, \mathbf{z}^k)$ ,  $Z$  is an estimated compact set on the dual variable, and  $\gamma_k > 0$  is a step size. Given a maximum iteration  $K$ , MSA with  $\gamma_k = \gamma/\sqrt{K}$ ,  $\forall k$  for some  $\gamma > 0$  can achieve an ergodic convergence rate  $O(1/\sqrt{K})$  in terms of the primal-dual gap.

### 1.2.2 Adaptive SGMs

A key difference between the existing methods reviewed above and our method for solving (1.1) or (1.2) is the use of an adaptiveness technique, which has been widely used to improve the empirical convergence speed of SGMs for machine learning, especially deep learning. The vanilla SGM is easy to implement, but its convergence is often slow. Many adaptive variants have been proposed to improve the empirical convergence speed, such as Adagrad [8], RMSprop [34], Adam [17], Adadelta [45], Nadam [6], and Amsgrad [31]. In the vanilla SGM, the same learning rate is used to all coordinates of the variable at each update, and this often results in poor performance for solving a problem with sparse training data [8]. In contrast, adaptive methods adaptively scale the learning rate for different coordinates of the variable, based on a certain average of all the stochastic gradients that have been generated. The arithmetic average is used by Adagrad [8] and the exponential moving average by others [6, 17, 31, 34, 45]. Besides the adaptive learning rate, momentum-based search direction is adopted in adaptive SGMs. These adaptive methods have been used a lot but mainly for problems without constraints or with easy-to-project constraints.

The adaptiveness technique that will be used in our method is closely related to the technique used by Amsgrad [31]. For a stochastic problem  $\min_{\mathbf{x} \in X} f_0(\mathbf{x}) = \mathbb{E}_{\xi_0}[F_0(\mathbf{x}; \xi_0)]$ , Amsgrad iteratively performs the updates:

$$\mathbf{m}^k = \beta_1 \mathbf{m}^{k-1} + (1 - \beta_1) \mathbf{u}^k,$$

$$\begin{aligned}
\mathbf{v}^k &= \beta_2 \mathbf{v}^{k-1} + (1 - \beta_2)(\mathbf{u}^k)^2, \\
\hat{\mathbf{v}}^k &= \max(\hat{\mathbf{v}}^{k-1}, \mathbf{v}^k), \\
\mathbf{x}^{k+1} &= \mathbf{Proj}_{X, (\hat{\mathbf{v}}^k)^{1/2}}(\mathbf{x}^k - \alpha_k \mathbf{m}^k / (\hat{\mathbf{v}}^k)^{1/2}),
\end{aligned} \tag{1.12}$$

where  $\mathbf{u}^k$  is an unbiased stochastic gradient of  $f_0$  at  $\mathbf{x}^k$ . Slightly different from Amsgrad, Adam [17] sets  $\hat{\mathbf{v}}^k = \mathbf{v}^k$  for all  $k$ . It turns out that the setting in (1.12) is important, as otherwise the method is not guaranteed to converge even for convex problems. The update in (1.12) ensures that each component of  $\hat{\mathbf{v}}^k$  is nondecreasing with respect to  $k$ . This addresses the divergence issue of Adam. However, one drawback is that once an unusual  $\mathbf{u}^k$  with “big” components appears in early iterations, the components of  $\hat{\mathbf{v}}^k$  will remain “big”, which results small effective learning rates and can slow the convergence. We address this issue by using the clipping technique in [1] to modify the update of  $\hat{\mathbf{v}}^k$ .

### 1.3 Contributions

The main contributions are listed below.

- We propose a new adaptive primal-dual stochastic gradient method (APriD) for solving expectation-constrained convex stochastic optimization problems. The method is derived based on the ordinary Lagrangian function of the underlying problem. During each iteration, APriD first inquires an unbiased stochastic subgradient of the Lagrangian function about both primal and dual variables, and then it performs a stochastic subgradient descent update to the primal variable and a stochastic gradient ascent update to the dual variable. For the primal update, an adaptive learning rate is applied, in order to have faster convergence.
- We analyze the convergence rate of APriD. Assuming the existence of a Karush–Kuhn–Tucker (KKT) solution, we prove the boundedness of the dual iterate in expectation. Based on that, we then establish the  $O(1/\sqrt{k})$  ergodic convergence rate, where  $k$  is the number of subgradient inquiries. The convergence rate results are in terms of primal and dual objective value and primal constraint violation.
- We also extend APriD to solve a convex–concave minimax problem. The extended method (named as APriAD), during each iteration, first inquires an unbiased estimation of a subgradient of the minimax problem and then updates the primal and dual variables by performing an adaptive update. An ergodic convergence rate of  $O(1/\sqrt{k})$  is also established in terms of the primal-dual gap.
- We test APriD on NPC [27, 32] and the quadratically constrained quadratic program (QCQP) in the expectation form and the scenario approximation form. We compare the proposed method to CSA [18] and MSA [24]. Although all three methods have the same-order convergence rate, the numerical results demonstrate that APriD can decrease the objective error and the constraint violations (sometimes significantly) faster than CSA and MSA, in terms of iteration number and the running time.

## 1.4 Notation

We use bold lower-case letters  $\mathbf{x}, \mathbf{z}$  for vectors and  $x_i, z_i$  for their  $i$ -th components. The bold number  $\mathbf{0}, \mathbf{1}$  denote the all-zero vector and all-one vector, respectively. We use the convention  $0/0 = 0$ . For any positive integer  $M$ ,  $[M]$  is short for the set  $\{1, 2, \dots, M\}$ . With slight abuse of notation, for any two vectors  $\mathbf{x}$  and  $\mathbf{y}$ ,  $|\mathbf{x}|$  takes element-wise absolute value,  $\mathbf{x}^p$  takes the element-wise  $p$ -th power,  $\mathbf{x}/\mathbf{y}$  takes the element-wise division, and  $\max\{\mathbf{x}, \mathbf{y}\}$  takes the element-wise maximum. We define  $[\mathbf{x}]_+ = \max\{\mathbf{x}, \mathbf{0}\}$  and  $[\mathbf{x}]_- = \max\{-\mathbf{x}, \mathbf{0}\}$ .  $\mathbf{x} \leq \mathbf{y}$  and  $\mathbf{x} \geq \mathbf{y}$  mean that the inequalities hold element-wisely.

For a given symmetric positive semidefinite matrix  $D$ , we define the weighted product between  $\mathbf{x}, \mathbf{z}$  as  $\langle \mathbf{x}, \mathbf{z} \rangle_D = \mathbf{x}^\top D \mathbf{z}$ . Specially, when  $D$  is the identity matrix  $I$ , we briefly denote  $\langle \mathbf{x}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle_I = \mathbf{x}^\top \mathbf{z}$ . For any vector  $\mathbf{v}$ ,  $\text{diag}(\mathbf{v})$  denotes a diagonal matrix with  $\mathbf{v}$  on the diagonal. For a nonnegative vector  $\mathbf{v} \geq \mathbf{0}$ ,  $\langle \mathbf{x}, \mathbf{z} \rangle_{\mathbf{v}} = \mathbf{x}^\top \text{diag}(\mathbf{v}) \mathbf{z}$ . For any vector  $\mathbf{x}$ , we denote  $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ ,  $\|\mathbf{x}\|_D = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_D}$ ,  $\|\mathbf{x}\|_{\mathbf{v}} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{v}}}$ ,  $\|\mathbf{x}\|_1 = \sum_i |x_i|$ , and  $\|\mathbf{x}\|_\infty = \max_i |x_i|$ . Similarly, for a closed convex set  $X$ , we denote the (weighted) projection onto  $X$  as  $\text{Proj}_X(\mathbf{y}) = \arg \min_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{y}\|^2$ ,  $\text{Proj}_{X,D}(\mathbf{y}) = \arg \min_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{y}\|_D^2$  and  $\text{Proj}_{X,\mathbf{v}}(\mathbf{y}) = \arg \min_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{y}\|_{\mathbf{v}}^2$ .

For a convex function  $f(\mathbf{x})$ , we denote  $\tilde{\nabla} f(\mathbf{x})$  as one subgradient and  $\partial f(\mathbf{x})$  as the set of all subgradients of  $f$  at  $\mathbf{x}$ . If  $f$  is differentiable, we simply use  $\nabla f(\mathbf{x})$  as the gradient. For a function  $f(\mathbf{x}, \mathbf{y})$ , we denote the partial gradient or partial subgradient about  $\mathbf{x}$  by  $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$  or  $\tilde{\nabla}_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$  and the set of all partial subgradients by  $\partial_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$ . In the analysis of our algorithm, we denote  $\mathcal{H}^k$  as all history information until the  $k$ -th iteration, i.e.,  $\mathcal{H}^k = \{\mathbf{x}^1, \mathbf{z}^1, \mathbf{x}^2, \mathbf{z}^2, \dots, \mathbf{x}^k, \mathbf{z}^k\}$ .

## 2 Adaptive primal-dual stochastic gradient method

In this section, we give the details of our adaptive primal-dual stochastic gradient method. Since the algorithm only requires the stochastic subgradient of the Lagrangian function of the underlying problem, it can be applied to both (1.1) and (1.2) by specifying how to obtain the stochastic subgradient. Hence, we focus on the more general formulation in (1.1).

Let  $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_M(\mathbf{x})]^\top$  be the vector function that concatenates all the constraint functions in (1.1). Then the Lagrangian function of (1.1) can be written as

$$\mathcal{L}(\mathbf{x}, \mathbf{z}) = f_0(\mathbf{x}) + \mathbf{z}^\top \mathbf{f}(\mathbf{x}), \quad (2.1)$$

where  $\mathbf{z} = [z_1, z_2, \dots, z_M]^\top$  is the Lagrange multiplier or dual variable. The Lagrangian dual problem is

$$\max_{\mathbf{z} \geq \mathbf{0}} \{d(\mathbf{z}) \equiv \min_{\mathbf{x} \in X} \mathcal{L}(\mathbf{x}, \mathbf{z})\}. \quad (2.2)$$



By the Lagrangian function in (2.1), we design an **adaptive primal-dual** stochastic gradient method for solving (1.1), named as APriD. The pseudocode is shown in Algorithm 1. We assume an *oracle* that can return an unbiased stochastic subgradient of  $\mathcal{L}$  at any inquiry point  $(\mathbf{x}, \mathbf{z})$ . At each iteration  $k$ , APriD first calls the oracle to obtain a stochastic subgradient  $(\mathbf{u}^k, \mathbf{w}^k)$  of  $\mathcal{L}$  at  $(\mathbf{x}^k, \mathbf{z}^k)$ . Then it updates the primal variable by performing an adaptive-SGM step and the dual variable by a vanilla-SGM step. Our adaptive update is similar to Amsgrad [31], and the difference is that we clip the primal stochastic gradient  $\mathbf{u}^k$  if its norm is greater than a user-specified parameter  $\theta$ . The clipping technique has been used in existing works about adaptive SGMs, e.g., [23], to avoid potentially too-small effective learning rate.

---

**Algorithm 1** Adaptive primal-dual stochastic gradient (APriD) method for (1.1)
 

---

- 1: **Initialization:** choose  $\mathbf{x}^1 \in X$  and  $\mathbf{z}^1 \geq \mathbf{0}$ , set  $\mathbf{m}^0 = \mathbf{0}$ ,  $\mathbf{v}^0 = \mathbf{0}$ ,  $\hat{\mathbf{v}}^0 = \mathbf{0}$ ;
- 2: **Parameter setting:** set the maximum number  $K$  of iterations; choose  $\beta_1, \beta_2 \in (0, 1)$ ,  $\theta > 0$ , non-increasing step sizes  $\{\alpha_k\}_{k=1}^K$  and  $\rho_1 > 0$ ; if  $\alpha_k = \alpha_1$  for  $k \in [K]$ , i.e.,  $\alpha_k$  is a constant, let  $\eta_1 = \sum_{i=1}^{\infty} \alpha_i \beta_1^{i-1} = \frac{\alpha_1}{1-\beta_1}$ ; otherwise, let  $\eta_1 = \sum_{i=1}^K \alpha_i \beta_1^{i-1}$ .
- 3: **for**  $k = 1, 2, \dots, K$  **do**
- 4:   Call the oracle to obtain a stochastic subgradient  $(\mathbf{u}^k, \mathbf{w}^k)$  of  $\mathcal{L}$  at  $(\mathbf{x}^k, \mathbf{z}^k)$ .
- 5:   Update the primal variable  $\mathbf{x}$  by

$$\mathbf{m}^k = \beta_1 \mathbf{m}^{k-1} + (1 - \beta_1) \mathbf{u}^k, \quad (2.3)$$

$$\hat{\mathbf{u}}^k = \frac{\mathbf{u}^k}{\max\{1, \frac{\|\mathbf{u}^k\|}{\theta}\}}, \quad (2.4)$$

$$\mathbf{v}^k = \beta_2 \mathbf{v}^{k-1} + (1 - \beta_2) (\hat{\mathbf{u}}^k)^2, \quad (2.5)$$

$$\hat{\mathbf{v}}^k = \max\{\hat{\mathbf{v}}^{k-1}, \mathbf{v}^k\}, \quad (2.6)$$

$$\mathbf{x}^{k+1} = \text{Proj}_{X, (\hat{\mathbf{v}}^k)^{1/2}}(\mathbf{x}^k - \alpha_k \mathbf{m}^k / (\hat{\mathbf{v}}^k)^{1/2}). \quad (2.7)$$

- 6:   If  $k \geq 2$ , update the step size  $\rho_k$  by

$$\rho_k = \frac{\rho_{k-1}}{\beta_1 + \frac{\alpha_{k-1}}{\eta_k}}, \text{ with } \eta_k = \frac{\eta_{k-1} - \alpha_{k-1}}{\beta_1}. \quad (2.8)$$

- 7:   Update the dual variable  $\mathbf{z}$  by

$$\mathbf{z}^{k+1} = [\mathbf{z}^k + \rho_k \mathbf{w}^k]_+. \quad (2.9)$$

8: **end for**

---

Below, we give a few facts about the step-size parameters of Algorithm 1. About the primal step size, the inequality in the following lemma will be used many times in our analysis.

**Lemma 1** For any  $1 \leq j \leq t \leq K$ , we have

$$\alpha_j \leq \sum_{k=j}^t \alpha_k \beta_1^{k-j} \leq \frac{\alpha_j}{1 - \beta_1}. \quad (2.10)$$

**Proof** The first inequality is trivial from the positivity of  $\{\alpha_j\}_{j=1}^K$  and  $\beta_1$ . The second inequality follows from the non-increasing monotonicity of  $\{\alpha_j\}_{j=1}^K$  and by noting  $\sum_{k=j}^t \alpha_k \beta_1^{k-j} \leq \alpha_j \sum_{k=j}^t \beta_1^{k-j} = \alpha_j \sum_{k=0}^{t-j} \beta_1^k \leq \alpha_j \sum_{k=0}^{\infty} \beta_1^k$ . The proof is finished by  $\beta_1 \in (0, 1)$ .  $\square$

About the dual step size, we show the next lemma, which will be used to analyze the summation of  $\|\mathbf{z}^j - \mathbf{z}\|^2$ . The proof of the lemma is given in Appendix A.

**Lemma 2** *The dual step size sequence  $\{\rho_j\}_{j=1}^K$  is non-increasing and*

$$\frac{\sum_{k=j-1}^t \alpha_k \beta_1^{k-(j-1)}}{\rho_{j-1}} - \frac{\sum_{k=j}^t \alpha_k \beta_1^{k-j}}{\rho_j} \geq 0, \quad \text{for any } 2 \leq j \leq t \leq K, \quad (2.11)$$

$$\rho_j \leq \frac{\rho_1 \alpha_j}{\alpha_1 (1 - \beta_1)}, \quad \text{for any } j \in [K]. \quad (2.12)$$

Furthermore, if a constant primal step size is set, i.e.,  $\alpha_j = \alpha_1$ , for all  $j \in [K]$  for some  $\alpha_1 > 0$ , then  $\eta_j = \frac{\alpha_1}{1 - \beta_1}$  and  $\rho_j = \rho_1$  for all  $j \in [K]$ , i.e. the dual step size is also a constant.

### 3 Convergence analysis

In this section, we analyze the convergence of Algorithm 1. We first give the assumptions in Sect. 3.1 and then show some preparatory lemmas in Sect. 3.2. In Sect. 3.3, we give the main convergence results, including a uniform bound on  $\mathbb{E}[\mathbf{z}^k]$ .

#### 3.1 Technical assumptions

Throughout our analysis in this section, we make the following three assumptions.

**Assumption 1**  $X$  is a compact convex set in  $\mathbb{R}^n$ , i.e. there exists a constant  $B$  such that

$$\|\mathbf{x}_1 - \mathbf{x}_2\|_{\infty} \leq B, \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in X.$$

For any  $\mathbf{x} \in \mathbb{R}^n$ ,  $\|\mathbf{x}\| \leq \sqrt{n} \|\mathbf{x}\|_{\infty}$ , and from Assumption 1, it holds

$$\|\mathbf{x}_1 - \mathbf{x}_2\|^2 \leq nB^2, \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in X. \quad (3.1)$$

**Assumption 2** There are constants  $P, Q, F$  such that for any  $k \in [K]$ ,

$$\begin{aligned} \mathbb{E}[\mathbf{u}^k | \mathcal{H}^k] &\in \partial_{\mathbf{x}} \mathcal{L}(\mathbf{x}^k, \mathbf{z}^k), \quad \mathbb{E}[\|\mathbf{u}^k\|^2] \leq P \cdot \mathbb{E}[\|\mathbf{z}^k\|^2] + Q, \\ \mathbb{E}[\mathbf{w}^k | \mathcal{H}^k] &= \mathbf{f}(\mathbf{x}^k) = \nabla_{\mathbf{z}} \mathcal{L}(\mathbf{x}^k, \mathbf{z}^k), \quad \mathbb{E}[\|\mathbf{w}^k\|^2] \leq F^2. \end{aligned}$$

By Assumption 2, we have for any  $k \in [K]$ ,

$$\max_{j \in [k]} \mathbb{E}[\|\mathbf{u}^j\|^2] \leq P \cdot \left( \max_{j \in [k]} \mathbb{E}[\|\mathbf{z}^j\|^2] \right) + Q \equiv \widehat{P}_{\mathbf{z}}^k. \quad (3.2)$$

**Assumption 3** There exists a primal-dual solution  $(\mathbf{x}^*, \mathbf{z}^*)$  satisfying the KKT conditions of (1.1):

$$\mathbf{0} \in \partial f_0(\mathbf{x}^*) + \mathcal{N}_X(\mathbf{x}^*) + \sum_{i=1}^M z_i^* \partial f_i(\mathbf{x}^*), \quad (3.3)$$

$$\mathbf{x}^* \in X, \quad f_i(\mathbf{x}^*) \leq 0, \quad \forall i \in [M], \quad (3.4)$$

$$z_i^* \geq 0, \quad z_i^* f_i(\mathbf{x}^*) = 0, \quad \forall i \in [M], \quad (3.5)$$

where  $\mathcal{N}_X(\mathbf{x})$  denotes the normal cone of  $X$  at  $\mathbf{x}$ .

In Assumption 2, the unbiasedness condition on the stochastic gradients is standard in the literature of SGMs. As  $X$  is compact, the uniform boundedness on  $\mathbb{E}[\|\mathbf{w}^k\|^2]$  can hold if  $\mathbf{f}$  is continuous. However, the bound on  $\mathbb{E}[\|\mathbf{u}^k\|^2]$  often depends on the  $\mathbf{z}$ -variable. To see this, we suppose that  $\mathbf{u}_i^k$  is an unbiased stochastic subgradient of  $f_i$  at  $\mathbf{x}^k$  for each  $i = 0, 1, \dots, M$  and let

$$\mathbf{u}^k = \mathbf{u}_0^k + \sum_{i=1}^M z_i^k \mathbf{u}_i^k.$$

Then  $\mathbf{u}^k$  will be an unbiased stochastic subgradient of  $\mathcal{L}(\cdot, \mathbf{z}^k)$  at  $\mathbf{x}^k$ . Denote  $\mathbf{U}^k = [\mathbf{u}_1^k, \dots, \mathbf{u}_M^k]$ . We have

$$\begin{aligned} \mathbb{E}[\|\mathbf{u}^k\|^2] &\leq 2\mathbb{E}[\|\mathbf{u}_0^k\|^2] + 2\mathbb{E}[\|\mathbf{z}^k\|^2 \|\mathbf{U}^k\|^2] \\ &= 2\mathbb{E}[\mathbb{E}[\|\mathbf{u}_0^k\|^2 \mid \mathcal{H}^k]] + 2\mathbb{E}[\|\mathbf{z}^k\|^2 \mathbb{E}[\|\mathbf{U}^k\|^2 \mid \mathcal{H}^k]] \end{aligned}$$

As  $X$  is compact,  $\mathbb{E}[\|\mathbf{u}_0^k\|^2 \mid \mathcal{H}^k]$  and  $\mathbb{E}[\|\mathbf{U}^k\|^2 \mid \mathcal{H}^k]$  can be uniformly bounded by some constants  $P$  and  $Q$ , and thus the boundedness condition on  $\mathbb{E}[\|\mathbf{u}^k\|^2]$  holds.

Assumption 3 is satisfied if a certain constraint qualification holds such as the Slater's condition. Under Assumption 3, the strong duality holds, i.e.

$$d(\mathbf{z}^*) = f_0(\mathbf{x}^*). \quad (3.6)$$

where  $d$  is defined in (2.2).

### 3.2 Preparatory lemmas

In this subsection, we establish a few lemmas, whose proofs are given in Appendixes B–G. First, we have the following lemma under Assumption 3.

**Lemma 3** *For any primal-dual solution  $(\mathbf{x}^*, \mathbf{z}^*)$  satisfying the KKT conditions (3.3)–(3.5), we have that for any  $\mathbf{x} \in X$ ,*

$$f_0(\mathbf{x}) - f_0(\mathbf{x}^*) + \langle \mathbf{z}^*, \mathbf{f}(\mathbf{x}) \rangle \geq 0. \quad (3.7)$$

Furthermore, for any  $\mathbf{z} \geq 0$ ,

$$f_0(\mathbf{x}) - f_0(\mathbf{x}^*) - \langle \mathbf{z}, \mathbf{f}(\mathbf{x}^*) \rangle + \langle \mathbf{z}^*, \mathbf{f}(\mathbf{x}) \rangle \geq 0. \quad (3.8)$$

Under Assumption 2 and by the updates in Algorithm 1, we are able to upper bound  $\mathbb{E}[\|(\hat{\mathbf{v}}^k)^{1/2}\|_1]$  and  $\mathbb{E}[\|\mathbf{m}^k\|_{(\hat{\mathbf{v}}^k)^{-1/2}}^2]$ , as summarized below.

**Lemma 4** *For any integer  $k \in [K]$ ,*

$$\mathbb{E}[\|(\hat{\mathbf{v}}^k)^{1/2}\|_1] \leq n\theta, \quad \mathbb{E}[\|\mathbf{m}^k\|_{(\hat{\mathbf{v}}^k)^{-1/2}}^2] \leq \frac{\sqrt{n}(\theta + \frac{\hat{P}_z^k}{\theta})}{(1 - \beta_2)^{1/2}},$$

where  $\hat{P}_z^k$  is defined in (3.2).

From the update in (2.7) about the primal variable, we have the following result.

**Lemma 5** *For any  $\mathbf{x} \in X$  and  $t \in [K]$ ,*

$$(1 - \beta_1) \sum_{j=1}^t \mathbb{E}[\langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j \rangle] \sum_{k=j}^t \alpha_k \beta_1^{k-j} \leq \frac{n\theta B^2}{2} + \frac{\sqrt{n}(\theta + \frac{\hat{P}_z^t}{\theta}) \sum_{j=1}^t \alpha_j^2}{2(1 - \beta_1)^2(1 - \beta_2)^{1/2}}. \quad (3.9)$$

From the update in (2.9) about the dual variable, we have the follows.

**Lemma 6** *For any  $\mathbf{z} \geq \mathbf{0}$ ,  $j \in [K]$ , it holds*

$$\begin{aligned} \langle \mathbf{z}^j, \mathbf{f}(\mathbf{x}^j) \rangle &\geq \langle \mathbf{z}, \mathbf{f}(\mathbf{x}^j) \rangle + \frac{1}{2\rho_j} \left( \|\mathbf{z}^{j+1} - \mathbf{z}\|^2 - \|\mathbf{z}^j - \mathbf{z}\|^2 \right) \\ &\quad - \frac{\rho_j}{2} \|\mathbf{w}^j\|^2 - \langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j - \mathbf{f}(\mathbf{x}^j) \rangle. \end{aligned} \quad (3.10)$$

Lemma 5 gives an upper bound of  $\sum_{j=1}^t \mathbb{E}[\langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j \rangle] \sum_{k=j}^t \alpha_k \beta_1^{k-j}$ , the next lemma gives a lower bound on this summation.

**Lemma 7** *For  $\mathbf{z} \geq \mathbf{0}$  and  $t \in [K]$  we have*

$$\begin{aligned}
 & \sum_{j=1}^t \mathbb{E}[\langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j \rangle] \sum_{k=j}^t \alpha_k \beta_1^{k-j} \\
 & \geq \sum_{j=1}^t \mathbb{E}[f_0(\mathbf{x}^j) - f_0(\mathbf{x}) - \langle \mathbf{z}^j, \mathbf{f}(\mathbf{x}) \rangle + \langle \mathbf{z}, \mathbf{f}(\mathbf{x}^j) \rangle] \sum_{k=j}^t \alpha_k \beta_1^{k-j} \\
 & \quad + \sum_{j=1}^t \mathbb{E}[-\langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j - \mathbf{f}(\mathbf{x}^j) \rangle + \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle] \sum_{k=j}^t \alpha_k \beta_1^{k-j} \\
 & \quad - \frac{\alpha_1 \mathbb{E}[\|\mathbf{z}^1 - \mathbf{z}\|^2]}{2\rho_1(1-\beta_1)} - \frac{\rho_1 F^2 \sum_{j=1}^t \alpha_j^2}{2\alpha_1(1-\beta_1)^2} + \frac{\alpha_t}{2\rho_t} \mathbb{E}[\|\mathbf{z}^{t+1} - \mathbf{z}\|^2], \quad (3.11)
 \end{aligned}$$

where  $\tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) = \mathbb{E}[\mathbf{u}^j \mid \mathcal{H}^j]$ .

The second term in the right side of (3.11) can be bounded as follows.

**Lemma 8** For any deterministic or stochastic vector  $(\mathbf{x}, \mathbf{z})$  with  $\mathbf{x} \in X$  and  $\mathbf{z} \geq \mathbf{0}$ , it holds for any positive number sequence  $\{\gamma_j\}_{j=1}^K$ , and  $t \in [K]$  that

$$\begin{aligned}
 & \sum_{j=1}^t \gamma_j \mathbb{E}[\langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j - \mathbf{f}(\mathbf{x}^j) \rangle - \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle] \\
 & \leq \frac{1}{2} \left( \mathbb{E}[\|\mathbf{z}^1 - \mathbf{z}\|^2] + nB^2 + (F^2 + \hat{P}_{\mathbf{z}}^t) \sum_{j=1}^t \gamma_j^2 \right), \quad (3.12)
 \end{aligned}$$

where  $\tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) = \mathbb{E}[\mathbf{u}^j \mid \mathcal{H}^j]$  and  $\hat{P}_{\mathbf{z}}^t$  is defined in (3.2). Furthermore, if  $\mathbf{x}$  and  $\mathbf{z}$  are deterministic, then

$$\sum_{j=1}^t \gamma_j \mathbb{E}[\langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j - \mathbf{f}(\mathbf{x}^j) \rangle - \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle] = 0. \quad (3.13)$$

### 3.3 Main convergence results

In this subsection, we use the previous established lemmas to show our main convergence results. First, we obtain the following theorem by combining the inequalities in (3.9) and (3.11).

**Theorem 1** Under Assumptions 1–3 and with the choice of  $\mathbf{z}^1 = \mathbf{0}$ , for any  $\mathbf{x} \in X$ ,  $\mathbf{z} \geq \mathbf{0}$ ,  $t \in [K]$ ,

$$\mathbb{E}[f_0(\bar{\mathbf{x}}^t) - f_0(\mathbf{x}) - \langle \bar{\mathbf{z}}^t, \mathbf{f}(\mathbf{x}) \rangle + \langle \mathbf{z}, \mathbf{f}(\bar{\mathbf{x}}^t) \rangle] \leq \epsilon_{\mathbf{z}} + \epsilon_0 \mathbb{E}[\|\mathbf{z}\|^2], \quad (3.14)$$

where  $\bar{\mathbf{x}}^t = \frac{\sum_{j=1}^t \sum_{k=j}^t \alpha_k \beta_1^{k-j} \mathbf{x}^j}{\sum_{j=1}^t \sum_{k=j}^t \alpha_k \beta_1^{k-j}}$ ,  $\bar{\mathbf{z}}^t = \frac{\sum_{j=1}^t \sum_{k=j}^t \alpha_k \beta_1^{k-j} \mathbf{z}^j}{\sum_{j=1}^t \sum_{k=j}^t \alpha_k \beta_1^{k-j}}$ , and

$$\begin{aligned} \epsilon_{\mathbf{z}} &= \frac{1}{2(1-\beta_1) \sum_{k=1}^t \alpha_k} \left( n(\theta+1)B^2 + \frac{\sqrt{n}(\theta + \frac{\hat{P}_{\mathbf{z}}^t}{\theta}) \sum_{j=1}^t \alpha_j^2}{(1-\beta_1)^2(1-\beta_2)^{1/2}} \right. \\ &\quad \left. + \frac{\rho_1 F^2 \sum_{j=1}^t \alpha_j^2}{\alpha_1(1-\beta_1)} + (F^2 + \hat{P}_{\mathbf{z}}^t) \sum_{j=1}^t \alpha_j^2 \right), \\ \epsilon_0 &= \frac{1}{2(1-\beta_1) \sum_{k=1}^t \alpha_k} \left( \frac{\alpha_1}{\rho_1} + 1 \right). \end{aligned}$$

**Proof** From (3.9), (3.11), (3.12), and  $\mathbf{z}^1 = \mathbf{0}$ , we have

$$\begin{aligned} & \frac{n\theta B^2}{2} + \frac{\sqrt{n}(\theta + \frac{\hat{P}_{\mathbf{z}}^t}{\theta}) \sum_{j=1}^t \alpha_j^2}{2(1-\beta_1)^2(1-\beta_2)^{1/2}} \\ & \stackrel{(3.9)}{\geq} (1-\beta_1) \sum_{j=1}^t \mathbb{E}[\langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j \rangle] \sum_{k=j}^t \alpha_k \beta_1^{k-j} \\ & \stackrel{(3.11), (3.12)}{\geq} (1-\beta_1) \sum_{j=1}^t \mathbb{E} \left[ f_0(\mathbf{x}^j) - f_0(\mathbf{x}) - \langle \mathbf{z}^j, \mathbf{f}(\mathbf{x}) \rangle + \langle \mathbf{z}, \mathbf{f}(\mathbf{x}^j) \rangle \right] \sum_{k=j}^t \alpha_k \beta_1^{k-j} \\ & \quad - \frac{\alpha_1 \mathbb{E}[\|\mathbf{z}\|^2]}{2\rho_1} - \frac{\rho_1 F^2 \sum_{j=1}^t \alpha_j^2}{2\alpha_1(1-\beta_1)} + \frac{(1-\beta_1)\alpha_t}{2\rho_t} \mathbb{E}[\|\mathbf{z}^{t+1} - \mathbf{z}\|^2] \\ & \quad - \frac{1}{2} \left( \mathbb{E}[\|\mathbf{z}\|^2] + nB^2 + (F^2 + \hat{P}_{\mathbf{z}}^t) \sum_{j=1}^t \alpha_j^2 \right), \end{aligned} \quad (3.15)$$

where in the second inequality, we have used (3.12) with  $\gamma_j = \sum_{k=j}^t \alpha_k \beta_1^{k-j} (1 - \beta_1) \leq \alpha_j$  by (2.10).

By the convexity of  $\{f_i\}_{i \geq 0}$ , we have from the above inequality that for any  $\mathbf{z} \geq \mathbf{0}$ ,

$$\begin{aligned} & \mathbb{E}[f_0(\bar{\mathbf{x}}^t) - f_0(\mathbf{x}) - \langle \bar{\mathbf{z}}^t, \mathbf{f}(\mathbf{x}) \rangle + \langle \mathbf{z}, \mathbf{f}(\bar{\mathbf{x}}^t) \rangle] \\ & \leq \mathbb{E} \left[ \sum_{j=1}^t \left( f_0(\mathbf{x}^j) - f_0(\mathbf{x}) - \langle \mathbf{z}^j, \mathbf{f}(\mathbf{x}) \rangle + \langle \mathbf{z}, \mathbf{f}(\mathbf{x}^j) \rangle \right) \frac{\sum_{k=j}^t \alpha_k \beta_1^{k-j}}{\sum_{j=1}^t \sum_{k=j}^t \alpha_k \beta_1^{k-j}} \right] \\ & = \frac{1}{\sum_{j=1}^t \sum_{k=j}^t \alpha_k \beta_1^{k-j}} \sum_{j=1}^t \mathbb{E} \left[ f_0(\mathbf{x}^j) - f_0(\mathbf{x}) - \langle \mathbf{z}^j, \mathbf{f}(\mathbf{x}) \rangle + \langle \mathbf{z}, \mathbf{f}(\mathbf{x}^j) \rangle \right] \sum_{k=j}^t \alpha_k \beta_1^{k-j} \end{aligned}$$

$$\leq \frac{1}{2(1-\beta_1) \sum_{k=1}^t \alpha_k} \left( n(\theta+1)B^2 + \frac{\sqrt{n}(\theta + \frac{\hat{P}_z^t}{\theta}) \sum_{j=1}^t \alpha_j^2}{(1-\beta_1)^2(1-\beta_2)^{1/2}} + \frac{\rho_1 F^2 \sum_{j=1}^t \alpha_j^2}{\alpha_1(1-\beta_1)} \right. \\ \left. + (1 + \frac{\alpha_1}{\rho_1}) \mathbb{E}[\|\mathbf{z}\|^2] + (F^2 + \hat{P}_z^t) \sum_{j=1}^t \alpha_j^2 \right),$$

where in the last inequality, we have used (3.15) for the numerator and the fact  $\sum_{j=1}^t \sum_{k=j}^t \alpha_k \beta_1^{k-j} \geq \sum_{j=1}^t \alpha_j$  by (2.10) for the denominator. Rearranging the above inequality gives the result in (3.14).  $\square$

In (3.14),  $\epsilon_0$  is a constant. In order to bound  $\epsilon_z$  that depends on  $\hat{P}_z^t$  defined in (3.2), we need to bound  $\mathbb{E}[\|\mathbf{z}^j\|^2]$ , for all  $j \in [t]$ . This is shown in the following theorem.

**Theorem 2** Suppose Assumptions 1–3 hold and  $\mathbf{z}^1 = \mathbf{0}$ . Assume that for a given positive number  $K$ ,  $\sum_{j=1}^K \alpha_j^2 \leq \hat{\alpha}$ , and  $\rho_1 < \frac{\alpha_1(1-\beta_1)^4(1-\beta_2)^{1/2}\theta}{2\sqrt{n}P\hat{\alpha}}$ . Then  $\mathbb{E}[\|\mathbf{z}^t\|^2]$  for all  $t \in [K+1]$  has a uniform bound, i.e.,

$$\mathbb{E}[\|\mathbf{z}^t\|^2] \leq \frac{C_1}{1-C_2},$$

where  $C_1 = \frac{2\rho_1}{\alpha_1(1-\beta_1)^2} \left( n\theta B^2 + \frac{\sqrt{n}(\theta^2+Q)\hat{\alpha}}{(1-\beta_1)^2(1-\beta_2)^{1/2}\theta} + \frac{\rho_1 F^2 \hat{\alpha}}{\alpha_1(1-\beta_1)} \right) + \left( 2 + \frac{2}{(1-\beta_1)^2} \right) \|\mathbf{z}^*\|^2$  and  $C_2 = \frac{2\rho_1\sqrt{n}P\hat{\alpha}}{\alpha_1(1-\beta_1)^4(1-\beta_2)^{1/2}\theta}$ .

**Proof** As the first step in the proof of Theorem 1, we obtain from (3.9) and (3.11) with  $\mathbf{x} = \mathbf{x}^*$ ,  $\mathbf{z} = \mathbf{z}^*$  that

$$\frac{n\theta B^2}{2} + \frac{\sqrt{n}(\theta + \frac{\hat{P}_z^t}{\theta}) \sum_{j=1}^t \alpha_j^2}{2(1-\beta_1)^2(1-\beta_2)^{1/2}} \stackrel{(3.9)}{\geq} (1-\beta_1) \sum_{j=1}^t \mathbb{E}[\langle \mathbf{x}^j - \mathbf{x}^*, \mathbf{u}^j \rangle] \sum_{k=j}^t \alpha_k \beta_1^{k-j} \\ \geq (1-\beta_1) \sum_{j=1}^t \mathbb{E}[f_0(\mathbf{x}^j) - f_0(\mathbf{x}^*) - \langle \mathbf{z}^j, \mathbf{f}(\mathbf{x}^*) \rangle + \langle \mathbf{z}^*, \mathbf{f}(\mathbf{x}^j) \rangle] \sum_{k=j}^t \alpha_k \beta_1^{k-j} \\ - \frac{\alpha_1 \|\mathbf{z}^*\|^2}{2\rho_1} - \frac{\rho_1 F^2 \sum_{j=1}^t \alpha_j^2}{2\alpha_1(1-\beta_1)} + \frac{\alpha_t(1-\beta_1)}{2\rho_t} \mathbb{E}[\|\mathbf{z}^{t+1} - \mathbf{z}^*\|^2] \\ + (1-\beta_1) \sum_{j=1}^t \mathbb{E}[-\langle \mathbf{z}^j - \mathbf{z}^*, \mathbf{w}^j - \mathbf{f}(\mathbf{x}^j) \rangle + \langle \mathbf{x}^j - \mathbf{x}^*, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle] \sum_{k=j}^t \alpha_k \beta_1^{k-j}.$$

Since  $\mathbf{x}^*$ ,  $\mathbf{z}^*$  are deterministic, the last term on the right side of the above inequality vanishes by Lemma 8. In addition, we use (3.8) and have from the above inequality that

$$\begin{aligned} & \frac{\alpha_t(1-\beta_1)}{2\rho_t} \mathbb{E}[\|\mathbf{z}^{t+1} - \mathbf{z}^*\|^2] \\ & \leq \frac{n\theta B^2}{2} + \frac{\sqrt{n}(\theta + \frac{\hat{P}_t}{\theta}) \sum_{j=1}^t \alpha_j^2}{2(1-\beta_1)^2(1-\beta_2)^{1/2}} + \frac{\alpha_1 \|\mathbf{z}^*\|^2}{2\rho_1} + \frac{\rho_1 F^2 \sum_{j=1}^t \alpha_j^2}{2\alpha_1(1-\beta_1)}. \end{aligned}$$

By (2.12),  $\frac{2\rho_t}{\alpha_t(1-\beta_1)} \leq \frac{2\rho_1\alpha_t}{\alpha_t\alpha_1(1-\beta_1)^2} = \frac{2\rho_1}{\alpha_1(1-\beta_1)^2}$ . Thus we have

$$\begin{aligned} & \mathbb{E}[\|\mathbf{z}^{t+1} - \mathbf{z}^*\|^2] \\ & \leq \frac{\rho_1}{\alpha_1(1-\beta_1)^2} \left( n\theta B^2 + \frac{\sqrt{n}(\theta + \frac{\hat{P}_t}{\theta}) \sum_{j=1}^t \alpha_j^2}{(1-\beta_1)^2(1-\beta_2)^{1/2}} + \frac{\alpha_1 \|\mathbf{z}^*\|^2}{\rho_1} + \frac{\rho_1 F^2 \sum_{j=1}^t \alpha_j^2}{\alpha_1(1-\beta_1)} \right). \end{aligned}$$

By  $\|\mathbf{z}^{t+1}\|^2 \leq 2(\|\mathbf{z}^{t+1} - \mathbf{z}^*\|^2 + \|\mathbf{z}^*\|^2)$ , we obtain

$$\begin{aligned} \mathbb{E}[\|\mathbf{z}^{t+1}\|^2] & \leq \frac{2\rho_1}{\alpha_1(1-\beta_1)^2} \left( n\theta B^2 + \frac{\sqrt{n}(\theta + \frac{\hat{P}_t}{\theta}) \sum_{j=1}^t \alpha_j^2}{(1-\beta_1)^2(1-\beta_2)^{1/2}} + \frac{\rho_1 F^2 \sum_{j=1}^t \alpha_j^2}{\alpha_1(1-\beta_1)} \right) \\ & \quad + \left( 2 + \frac{2}{(1-\beta_1)^2} \right) \|\mathbf{z}^*\|^2 \\ & \leq C_1 + C_2 \max_{k \in [t]} \left( \mathbb{E}[\|\mathbf{z}^k\|^2] \right), \end{aligned}$$

where in the last inequality, we have used (3.2) and  $\sum_{j=1}^t \alpha_j^2 \leq \sum_{j=1}^K \alpha_j^2 \leq \hat{\alpha}$  for all  $t \in [K]$ . Notice that  $C_2 < 1$  from the selection of  $\rho_1$ . Below we show  $\mathbb{E}[\|\mathbf{z}^t\|^2] \leq \frac{C_1}{1-C_2}$  for all  $t \in [K+1]$  by induction.

- When  $t = 1$ ,  $\mathbf{z}^1 = \mathbf{0}$ , thus  $\mathbb{E}[\|\mathbf{z}^1\|^2] \leq \frac{C_1}{1-C_2}$  holds trivially.
- Assume that  $\mathbb{E}[\|\mathbf{z}^k\|^2] \leq \frac{C_1}{1-C_2}$  holds for  $k \leq t$ , then

$$\mathbb{E}[\|\mathbf{z}^{t+1}\|^2] \leq C_1 + C_2 \left( \max_{k \in [t]} \mathbb{E}[\|\mathbf{z}^k\|^2] \right) \leq C_1 + C_2 \frac{C_1}{1-C_2} = \frac{C_1}{1-C_2}.$$

Therefore, we complete the induction and obtain the desired result.  $\square$

When the conditions in Theorem 2 hold, we have that for  $k \in [K+1]$ ,

$$\hat{P}_z^k = P \cdot \left( \max_{j \in [k]} \mathbb{E}[\|\mathbf{z}^j\|^2] \right) + Q \leq P \frac{C_1}{1-C_2} + Q \equiv \hat{P}.$$

Then  $\epsilon_z$  in Theorem 1 can be bounded by a constant as follows:

$$\epsilon_z \leq \frac{1}{2(1-\beta_1) \sum_{k=1}^t \alpha_k} \left( n(\theta+1)B^2 + \frac{\sqrt{n}(\theta + \frac{\hat{P}}{\theta}) \sum_{j=1}^t \alpha_j^2}{(1-\beta_1)^2(1-\beta_2)^{1/2}} \right)$$



$$+ \frac{\rho_1 F^2 \sum_{j=1}^t \alpha_j^2}{\alpha_1(1-\beta_1)} + (F^2 + \widehat{P}) \sum_{j=1}^t \alpha_j^2 \equiv \epsilon_1. \quad (3.16)$$

**Remark 1** In Theorems 1 and 2, we assume  $\mathbf{z}^1 = \mathbf{0}$  in the initialization of Algorithm 1, which simplifies the proof. But notice that we can get the same-order convergence rate with any  $\mathbf{z}^1 \geq \mathbf{0}$  by a similar proof with slightly different  $\epsilon_0, \epsilon_z, C_1$ .

With Theorems 1 and 2, we are ready to show the convergence rate results by the following lemma, whose proof is given in Appendix H.

**Lemma 9** Let  $\bar{\mathbf{x}} \in X$  and  $\bar{\mathbf{z}} \geq \mathbf{0}$  be random vectors. If for any  $\mathbf{x} \in X$  and  $\mathbf{z} \geq \mathbf{0}$  that may depend on  $(\bar{\mathbf{x}}, \bar{\mathbf{z}})$ , there are two constants  $\epsilon_1, \epsilon_0$  satisfying

$$\mathbb{E}[f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}) - \langle \bar{\mathbf{z}}, \mathbf{f}(\mathbf{x}) \rangle + \langle \mathbf{z}, \mathbf{f}(\bar{\mathbf{x}}) \rangle] \leq \epsilon_1 + \epsilon_0 \mathbb{E}[\|\mathbf{z}\|^2], \quad (3.17)$$

then for any  $(\mathbf{x}^*, \mathbf{z}^*)$  satisfying KKT conditions, we have

$$\mathbb{E}[|f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*)|] \leq 2\epsilon_1 + 9\epsilon_0 \|\mathbf{z}^*\|^2, \quad (3.18)$$

$$\mathbb{E}\left[\sum_{i=1}^M [f_i(\bar{\mathbf{x}})]_+\right] \leq \epsilon_1 + \epsilon_0 \|1 + \mathbf{z}^*\|^2, \quad (3.19)$$

$$\mathbb{E}[d(\mathbf{z}^*) - d(\bar{\mathbf{z}})] \leq \frac{3}{2}(\epsilon_1 + 3\epsilon_0 \|\mathbf{z}^*\|^2). \quad (3.20)$$

Below, we specify the parameters and give convergence rate results for two different choices of  $\{\alpha_k\}_{k=1}^K$ . One is a constant sequence and the other a varying sequence.

**Corollary 1** (Convergence rate with constant step size) Given any positive integer  $K$ , set  $\alpha_j = \frac{\alpha}{\sqrt{K}}$  for  $j \in [K]$  and  $\rho_1 = \frac{\rho}{\sqrt{K}}$  with  $\rho < \frac{(1-\beta_1)^4(1-\beta_2)^{1/2}\theta}{2\alpha\sqrt{n}P}$ . Let  $\{(\mathbf{x}^k, \mathbf{z}^k)\}$  be the sequence generated from Algorithm 1 with  $\mathbf{z}^1 = \mathbf{0}$ , and  $\bar{\mathbf{x}}^K = \frac{\sum_{j=1}^K (1-\beta_1^{K-j+1})\mathbf{x}^j}{\sum_{j=1}^K (1-\beta_1^{K-j+1})}$  and  $\bar{\mathbf{z}}^K = \frac{\sum_{j=1}^K (1-\beta_1^{K-j+1})\mathbf{z}^j}{\sum_{j=1}^K (1-\beta_1^{K-j+1})}$ . Define

$$C_1 = \frac{2\rho}{\alpha(1-\beta_1)^2} \left( n\theta B^2 + \frac{\sqrt{n}(\theta^2 + Q)\alpha^2}{(1-\beta_1)^2(1-\beta_2)^{1/2}\theta} + \frac{\rho\alpha F^2}{(1-\beta_1)} \right) + \left( 2 + \frac{2}{(1-\beta_1)^2} \right) \|\mathbf{z}^*\|^2,$$

$$C_2 = \frac{2\rho\alpha\sqrt{n}P}{(1-\beta_1)^4(1-\beta_2)^{1/2}\theta},$$

$$\widehat{P} = P \frac{C_1}{1 - C_2} + Q,$$

$$\phi = n(\theta + 1)B^2 + \frac{\sqrt{n}\alpha^2(\theta^2 + \widehat{P})}{(1-\beta_1)^2(1-\beta_2)^{1/2}\theta} + \frac{\alpha\rho F^2}{(1-\beta_1)} + (F^2 + \widehat{P})\alpha^2.$$

Then we have

$$\mathbb{E}[|f_0(\bar{\mathbf{x}}^K) - f_0(\mathbf{x}^*)|] \leq \frac{1}{2(1-\beta_1)\alpha\sqrt{K}} \left( 2\phi + \frac{9(\alpha+\rho)}{\rho} \|\mathbf{z}^*\|^2 \right), \quad (3.21)$$

$$\mathbb{E}\left[\sum_{i=1}^M [f_i(\bar{\mathbf{x}}^K)]_+\right] \leq \frac{1}{2(1-\beta_1)\alpha\sqrt{K}} \left( \phi + \frac{\alpha+\rho}{\rho} \|1 + \mathbf{z}^*\|^2 \right), \quad (3.22)$$

$$\mathbb{E}[d(\mathbf{z}^*) - d(\bar{\mathbf{z}}^K)] \leq \frac{3}{4(1-\beta_1)\alpha\sqrt{K}} \left( \phi + \frac{3(\alpha+\rho)}{\rho} \|\mathbf{z}^*\|^2 \right). \quad (3.23)$$

**Proof** From  $\alpha_j = \frac{\alpha}{\sqrt{K}}$  for  $j \in [K]$ , we have  $\sum_{j=1}^t \alpha_j^2 \leq \sum_{j=1}^K \alpha_j^2 = \alpha^2$  for all  $t \in [K]$ . Hence, the conditions required by Theorem 2 hold, and thus  $\mathbb{E}[\|\mathbf{z}^{t+1}\|^2] \leq \frac{C_1}{1-C_2}$  for  $t \in [K]$  with the given  $C_1$ ,  $C_2$ , and  $\hat{P}$ .

Note  $\sum_{k=1}^K \alpha_k = \alpha\sqrt{K}$  and  $\epsilon_{\mathbf{z}} \leq \epsilon_1$  in (3.16). By Theorem 1, we have that for any  $\mathbf{x} \in X$  and  $\mathbf{z} \geq \mathbf{0}$ ,

$$\mathbb{E}[f_0(\bar{\mathbf{x}}^K) - f_0(\mathbf{x}) - \langle \bar{\mathbf{z}}^K, \mathbf{f}(\mathbf{x}) \rangle + \langle \mathbf{z}, \mathbf{f}(\bar{\mathbf{x}}^K) \rangle] \leq \epsilon_1 + \epsilon_0 \mathbb{E}[\|\mathbf{z}\|^2],$$

where

$$\begin{aligned} \epsilon_1 &= \frac{1}{2(1-\beta_1)\alpha\sqrt{K}} \left( n(\theta+1)B^2 + \frac{\sqrt{n}(\theta + \frac{\hat{P}}{\theta})\alpha^2}{(1-\beta_1)^2(1-\beta_2)^{1/2}} + \frac{\alpha\rho F^2}{(1-\beta_1)} + (F^2 + \hat{P})\alpha^2 \right), \\ \epsilon_0 &= \frac{1}{2(1-\beta_1)\alpha\sqrt{K}} \left( \frac{\alpha}{\rho} + 1 \right). \end{aligned}$$

Therefore, by Lemma 9, we complete the proof.  $\square$

**Remark 2** The convergence rate results in Corollary 1 indicate that the parameter  $\theta$  can neither be too big or too small. Let  $\alpha$  be fixed and  $\rho$  proportional to  $\theta$  such that  $\rho < \frac{(1-\beta_1)^4(1-\beta_2)^{1/2}\theta}{2\alpha\sqrt{n}P}$ . Then  $C_2$  is a constant independent of  $\theta$ , and  $C_1$  and  $\phi$  are quadratically dependent on  $\theta$ . However, the second terms in the parenthesis of the right sides of (3.21)–(3.23) are inversely proportional to  $\theta$ . Thus the right-hand sides of (3.21)–(3.23) will approach to infinity if  $\theta \rightarrow \infty$  or  $\theta \rightarrow 0$ . Similarly,  $\theta$  can neither be too big or too small in Corollaries 2, 3 and 4. But notice that if  $\mathbf{u}^k$  in (2.4) is uniformly bounded for all  $k$ , then a too big  $\theta$  will not have effect of clipping. Numerically, we observe that the algorithm can still perform well even with a very small  $\theta$ , even if  $\alpha$  and  $\rho$  are both relatively large.

**Remark 3** Corollary 1 requires  $\rho < \frac{(1-\beta_1)^4(1-\beta_2)^{1/2}\theta}{2\alpha\sqrt{n}P}$ , or equivalently  $\alpha\rho \leq \frac{(1-\beta_1)^4(1-\beta_2)^{1/2}\theta}{2\sqrt{n}P}$ . In some applications,  $P$  is unknown, and even it can be estimated, strictly following the setting may give too conservative stepsizes. In the experiments, we do not follow the condition strictly, but instead we tune  $\alpha$  and  $\rho$ . In Sect. 5.5, we test Algorithm 1 with different combinations of  $(\alpha, \theta, \rho)$ , and it turns out that the algorithm can perform well in a wide range of values.

**Corollary 2** (Convergence rate with varying step size) *For all  $j \in [K]$ , set  $\alpha_j = \frac{\alpha}{\sqrt{j+1} \log(j+1)}$ . In addition, choose  $\rho_1 = \frac{\rho}{\sqrt{2} \log(2)}$  with  $\rho < \frac{(1-\beta_1)^4(1-\beta_2)^{1/2}\theta}{5\alpha\sqrt{n}P}$ . Let  $\{(\mathbf{z}^k, \mathbf{z}^k)\}$  be the sequence generated from Algorithm 1 with  $\mathbf{z}^1 = \mathbf{0}$ , and  $\bar{\mathbf{x}}^K = \frac{\sum_{j=1}^K \sum_{k=j}^K \alpha_k \beta_1^{k-j} \mathbf{x}^j}{\sum_{j=1}^K \sum_{k=j}^K \alpha_k \beta_1^{k-j}}$ , and  $\bar{\mathbf{z}}^K = \frac{\sum_{j=1}^K \sum_{k=j}^K \alpha_k \beta_1^{k-j} \mathbf{z}^j}{\sum_{j=1}^K \sum_{k=j}^K \alpha_k \beta_1^{k-j}}$ , Define*

$$C_1 = \frac{\rho}{\alpha(1-\beta_1)^2} \left( 2n\theta B^2 + \frac{5\sqrt{n}\alpha^2(\theta^2 + Q)}{(1-\beta_1)^2(1-\beta_2)^{1/2}\theta} + \frac{5\alpha\rho F^2}{(1-\beta_1)} \right) + \left( 2 + \frac{2}{(1-\beta_1)^2} \right) \|\mathbf{z}^*\|^2,$$

$$C_2 = \frac{5\alpha\rho\sqrt{n}P}{(1-\beta_1)^4(1-\beta_2)^{1/2}\theta},$$

$$\hat{P} = P \frac{C_1}{1 - C_2} + Q,$$

$$\phi = n(\theta + 1)B^2 + \frac{2.5\alpha^2\sqrt{n}(\theta^2 + \hat{P})}{(1-\beta_1)^2(1-\beta_2)^{1/2}\theta} + \frac{2.5\alpha\rho F^2}{(1-\beta_1)} + 2.5\alpha^2(F^2 + \hat{P}).$$

Then we have

$$\mathbb{E}[|f_0(\bar{\mathbf{x}}^K) - f_0(\mathbf{x}^*)|] \leq \frac{\log(1+K)}{4(1-\beta_1)\alpha(\sqrt{K+2}-\sqrt{2})} \left( 2\phi + \frac{9(\alpha+\rho)}{\rho} \|\mathbf{z}^*\|^2 \right),$$

$$\mathbb{E}\left[\sum_{i=1}^M [f_i(\bar{\mathbf{x}}^K)]_+\right] \leq \frac{\log(1+K)}{4(1-\beta_1)\alpha(\sqrt{K+2}-\sqrt{2})} \left( \phi + \frac{\alpha+\rho}{\rho} \|1 + \mathbf{z}^*\|^2 \right),$$

$$\mathbb{E}[d(\mathbf{z}^*) - d(\bar{\mathbf{z}}^K)] \leq \frac{3\log(1+K)}{8(1-\beta_1)\alpha(\sqrt{K+2}-\sqrt{2})} \left( \phi + \frac{3(\alpha+\rho)}{\rho} \|\mathbf{z}^*\|^2 \right).$$

**Proof** Note that for any  $t \in [K]$ ,

$$\begin{aligned} \sum_{j=1}^t \alpha_j^2 &\leq \sum_{j=1}^{\infty} \alpha_j^2 \leq \alpha^2 \sum_{j=1}^{\infty} \frac{1}{(j+1)(\log(j+1))^2} \\ &\leq \alpha^2 \left( \frac{1}{2(\log 2)^2} + \int_1^{\infty} \frac{1}{(x+1)(\log(x+1))^2} dx \right) \\ &= \alpha^2 \left( \frac{1}{2(\log 2)^2} + \frac{1}{\log 2} \right) \leq 2.5\alpha^2. \end{aligned}$$

Hence, by the choice of  $\rho_1$ , the conditions required by Theorem 2 hold, and thus  $\mathbb{E}[\|\mathbf{z}^{t+1}\|^2] \leq \frac{C_1}{1-C_2}$  for all  $t \in [K]$  with the given  $C_1$ ,  $C_2$ , and  $\hat{P}$ . Note  $\epsilon_{\mathbf{z}} \leq \epsilon_1$  in (3.16) and  $\sum_{k=1}^K \alpha_k$  satisfies

$$\sum_{k=1}^K \alpha_k = \sum_{k=1}^K \frac{\alpha}{\sqrt{k+1} \log(k+1)} \geq \frac{\alpha}{\log(1+K)} \int_{k=1}^{K+1} \frac{1}{\sqrt{k+1}} = \frac{2\alpha(\sqrt{K+2}-\sqrt{2})}{\log(1+K)}.$$

By Theorem 1, we have for any  $\mathbf{x} \in X$ ,  $\mathbf{z} \geq \mathbf{0}$ ,

$$\mathbb{E}[f_0(\bar{\mathbf{x}}^K) - f_0(\mathbf{x}) - \langle \bar{\mathbf{z}}^K, \mathbf{f}(\mathbf{x}) \rangle + \langle \mathbf{z}, \mathbf{f}(\bar{\mathbf{x}}^K) \rangle] \leq \epsilon_1 + \epsilon_0 \mathbb{E}[\|\mathbf{z}\|^2],$$

where

$$\begin{aligned} \epsilon_1 &= \frac{\log(1+K)}{4(1-\beta_1)\alpha(\sqrt{K+2}-\sqrt{2})} \left( n(\theta+1)B^2 + \frac{2.5\alpha^2\sqrt{n}(\theta+\frac{\hat{P}}{\theta})}{(1-\beta_1)^2(1-\beta_2)^{1/2}} \right. \\ &\quad \left. + \frac{2.5\alpha\rho F^2}{(1-\beta_1)} + 2.5\alpha^2(F^2+\hat{P}) \right), \\ \epsilon_0 &= \frac{\log(1+K)}{4(1-\beta_1)\alpha(\sqrt{K+2}-\sqrt{2})} \left( \frac{\alpha}{\rho} + 1 \right). \end{aligned}$$

By Lemma 9, we obtain the desired results and complete the proof.  $\square$

## 4 Extension to Stochastic minimax problem

In this section, we modify the APriD method in Algorithm 1 to solve a stochastic convex–concave minimax problem in the form of

$$\min_{\mathbf{x} \in X} \max_{\mathbf{z} \in Z} \{ \mathcal{L}(\mathbf{x}, \mathbf{z}) = \mathbb{E}_{\xi} [L(\mathbf{x}, \mathbf{z}; \xi)] \}. \quad (4.1)$$

Similar to existing works (e.g., [5,16,24,46]), we assume that  $X \subseteq \mathbb{R}^n$  and  $Z \subseteq \mathbb{R}^m$  in (4.1) are compact convex sets,  $\xi$  is a random vector, and  $\mathcal{L}(\mathbf{x}, \mathbf{z})$  is convex in  $\mathbf{x} \in X$  and concave in  $\mathbf{z} \in Z$ . By the minimax theorem [25], the strong duality holds for (4.1), namely,

$$\min_{\mathbf{x} \in X} \max_{\mathbf{z} \in Z} \mathcal{L}(\mathbf{x}, \mathbf{z}) = \max_{\mathbf{z} \in Z} \min_{\mathbf{x} \in X} \mathcal{L}(\mathbf{x}, \mathbf{z}). \quad (4.2)$$

In Algorithm 1, we perform an adaptive SGM update to the primal variable  $\mathbf{x}$  in (2.3)–(2.7), but a vanilla SGM update to the dual variable  $\mathbf{z}$ . This is because the Lagrangian function of (1.1) has a simple linear dependence on the dual variable. As demonstrated in [38], an adaptive primal–dual SGM performs similarly as well as its non-adaptive counterpart for a linearly constrained problem. In (4.1),  $\mathcal{L}$  can have complex dependence on both  $\mathbf{x}$  and  $\mathbf{z}$ . Hence, we modify Algorithm 1 to solve (4.1) by performing adaptive updates to both  $\mathbf{x}$  and  $\mathbf{z}$ . The modified algorithm is named as APriAD, and its pseudocode is given in Algorithm 2. Similar to Algorithm 1, we assume a stochastic first-order oracle, which can return an unbiased stochastic subgradient of  $\mathcal{L}$  at any inquiry point  $(\mathbf{x}, \mathbf{z})$ .

Below we analyze the convergence of Algorithm 2. Throughout our analysis in this section, we make the following two assumptions.

**Algorithm 2** Adaptive primal-dual stochastic gradient (APriAD) method for (4.1)

- 1: **Initialization:** choose  $\mathbf{x}^1 \in X$ ,  $\mathbf{z}^1 \in Z$ ,  $\mathbf{m}^0 = [\mathbf{m}_x^0, \mathbf{m}_z^0] = \mathbf{0}$ ,  $\mathbf{v}^0 = [\mathbf{v}_x^0, \mathbf{v}_z^0] = \mathbf{0}$ ,  $\hat{\mathbf{v}}^0 = [\hat{\mathbf{v}}_x^0, \hat{\mathbf{v}}_z^0] = \mathbf{0}$ ;
- 2: **Parameter setting:** set the maximum number  $K$  of iterations; choose  $\beta_1, \beta_2 \in (0, 1)$ ,  $\theta > 0$ , non-increasing step sizes  $\{\alpha_k\}_{k=1}^K$  and  $\{\rho_k\}_{k=1}^K$ ;
- 3: **for all**  $k = 1, 2, \dots, K$  **do**
- 4:   Call the oracle to return a stochastic subgradient  $\mathbf{g}^k = (\mathbf{u}^k, \mathbf{w}^k)$  of  $\mathcal{L}$  at  $(\mathbf{x}^k, \mathbf{z}^k)$ .
- 5:   Update the primal variables  $\mathbf{x}$  and dual variables  $\mathbf{z}$  by

$$\mathbf{m}^k = \beta_1 \mathbf{m}^{k-1} + (1 - \beta_1) \mathbf{g}^k, \quad (4.3)$$

$$\hat{\mathbf{g}}^k = (\hat{\mathbf{u}}^k, \hat{\mathbf{w}}^k) = \left( \frac{\mathbf{u}^k}{\max\{1, \frac{\|\mathbf{u}^k\|}{\theta}\}}, \frac{\mathbf{w}^k}{\max\{1, \frac{\|\mathbf{w}^k\|}{\theta}\}} \right), \quad (4.4)$$

$$\mathbf{v}^k = \beta_2 \mathbf{v}^{k-1} + (1 - \beta_2) (\hat{\mathbf{g}}^k)^2, \quad (4.5)$$

$$\hat{\mathbf{v}}^k = \max\{\hat{\mathbf{v}}^{k-1}, \mathbf{v}^k\}, \quad (4.6)$$

$$\mathbf{x}^{k+1} = \text{Proj}_{X, (\hat{\mathbf{v}}_x^k)^{1/2}}(\mathbf{x}^k - \alpha_k \mathbf{m}_x^k / (\hat{\mathbf{v}}_x^k)^{1/2}), \quad (4.7)$$

$$\mathbf{z}^{k+1} = \text{Proj}_{Z, (\hat{\mathbf{v}}_z^k)^{1/2}}(\mathbf{z}^k + \rho_k \mathbf{m}_z^k / (\hat{\mathbf{v}}_z^k)^{1/2}). \quad (4.8)$$

6: **end for**

**Assumption 4**  $X$  and  $Z$  are both compact convex sets, i.e., there exist constants  $B_x, B_z$  such that

$$\|\mathbf{x}_1 - \mathbf{x}_2\|_\infty \leq B_x, \forall \mathbf{x}_1, \mathbf{x}_2 \in X; \quad \|\mathbf{z}_1 - \mathbf{z}_2\|_\infty \leq B_z, \forall \mathbf{z}_1, \mathbf{z}_2 \in Z.$$

**Assumption 5** The stochastic subgradient  $(\mathbf{u}^k, \mathbf{w}^k)$  of  $\mathcal{L}(\mathbf{x}^k, \mathbf{z}^k)$  is unbiased and bounded for all  $k \in [K]$ , i.e., there are constants  $M_x$  and  $M_z$  such that for any  $k \in [K]$ ,

$$\begin{aligned} \mathbb{E}[\mathbf{u}^k \mid \mathcal{H}^k] &= \tilde{\nabla}_x \mathcal{L}(\mathbf{x}^k, \mathbf{z}^k) \in \partial_x \mathcal{L}(\mathbf{x}^k, \mathbf{z}^k), \quad \mathbb{E}[\|\mathbf{u}^k\|^2] \leq M_x^2, \\ \mathbb{E}[\mathbf{w}^k \mid \mathcal{H}^k] &= \tilde{\nabla}_z \mathcal{L}(\mathbf{x}^k, \mathbf{z}^k) \in \partial_z \mathcal{L}(\mathbf{x}^k, \mathbf{z}^k), \quad \mathbb{E}[\|\mathbf{w}^k\|^2] \leq M_z^2. \end{aligned}$$

Different from Assumption 2, here we assume uniform bounds on  $\mathbb{E}[\|\mathbf{u}^k\|^2]$  and  $\mathbb{E}[\|\mathbf{w}^k\|^2]$  as  $X$  and  $Z$  are both compact. By the assumption, we have that for any  $k \in [K]$ ,

$$\max_{j \in [k]} \mathbb{E}[\|\mathbf{u}^j\|^2] \leq M_x^2, \quad \max_{j \in [k]} \mathbb{E}[\|\mathbf{w}^j\|^2] \leq M_z^2.$$

Similarly to Lemma 4, we have the bounds

$$\begin{aligned} \mathbb{E}[(\hat{\mathbf{v}}_x^k)^{1/2} \|\cdot\|_1] &\leq n\theta, \quad \mathbb{E}[\|\mathbf{m}_x^k\|_{(\hat{\mathbf{v}}_x^k)^{-1/2}}^2] \leq \frac{\sqrt{n}(\theta^2 + M_x^2)}{(1 - \beta_2)^{1/2}\theta}, \\ \mathbb{E}[(\hat{\mathbf{v}}_z^k)^{1/2} \|\cdot\|_1] &\leq m\theta, \quad \mathbb{E}[\|\mathbf{m}_z^k\|_{(\hat{\mathbf{v}}_z^k)^{-1/2}}^2] \leq \frac{\sqrt{m}(\theta^2 + M_z^2)}{(1 - \beta_2)^{1/2}\theta}. \end{aligned}$$

By the convexity–concavity of  $\mathcal{L}(\mathbf{x}, \mathbf{z})$ , we can upper bound  $\mathcal{L}(\mathbf{x}^j, \mathbf{z}) - \mathcal{L}(\mathbf{x}, \mathbf{z}^j)$  as follows.

**Lemma 10** For any  $\mathbf{x} \in X$  and any  $\mathbf{z} \in Z$  and  $j \in [K]$  we have

$$\begin{aligned} \mathcal{L}(\mathbf{x}^j, \mathbf{z}) - \mathcal{L}(\mathbf{x}, \mathbf{z}^j) &\leq \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j \rangle - \langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j \rangle + \langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j - \tilde{\nabla}_{\mathbf{z}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle \\ &\quad - \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle, \end{aligned} \quad (4.9)$$

where  $\tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) = \mathbb{E}[\mathbf{u}^j \mid \mathcal{H}^j] \in \partial_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j)$  and  $\tilde{\nabla}_{\mathbf{z}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) = \mathbb{E}[\mathbf{w}^j \mid \mathcal{H}^j] \in \partial_{\mathbf{z}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j)$ .

**Proof** For any  $j \in [K]$ , we have

$$\langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j \rangle = \langle \mathbf{x}^j - \mathbf{x}, \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle + \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle, \quad (4.10)$$

$$\langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j \rangle = \langle \mathbf{z}^j - \mathbf{z}, \tilde{\nabla}_{\mathbf{z}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle + \langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j - \tilde{\nabla}_{\mathbf{z}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle. \quad (4.11)$$

Because  $\mathcal{L}(\mathbf{x}, \mathbf{z})$  is convex in  $\mathbf{x}$  and concave in  $\mathbf{z}$ , we have

$$\begin{aligned} \langle \mathbf{x}^j - \mathbf{x}, \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle &\geq \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) - \mathcal{L}(\mathbf{x}, \mathbf{z}^j), \\ \langle \mathbf{z}^j - \mathbf{z}, \tilde{\nabla}_{\mathbf{z}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle &\leq \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) - \mathcal{L}(\mathbf{x}^j, \mathbf{z}). \end{aligned}$$

Negating the second one of the above two inequalities and adding to the first one give

$$\langle \mathbf{x}^j - \mathbf{x}, \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle - \langle \mathbf{z}^j - \mathbf{z}, \tilde{\nabla}_{\mathbf{z}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle \geq \mathcal{L}(\mathbf{x}^j, \mathbf{z}) - \mathcal{L}(\mathbf{x}, \mathbf{z}^j).$$

Replacing the left two terms of the above inequality by (4.10) and (4.11), we obtain the desired result by rearranging terms.  $\square$

For the first two terms in the right side of (4.9), we have the following lemma by the same arguments as those in the proof of Lemma 5.

**Lemma 11** For any  $t \in [K]$ ,  $\mathbf{x} \in X$ ,  $\mathbf{z} \in Z$ ,

$$\begin{aligned} (1 - \beta_1) \sum_{j=1}^t \mathbb{E}[\langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j \rangle] \sum_{k=j}^t \alpha_k \beta_1^{k-j} &\leq \frac{n\theta B_{\mathbf{x}}^2}{2} + \frac{\sqrt{n}(\theta^2 + M_{\mathbf{x}}^2) \sum_{j=1}^t \alpha_j^2}{2(1 - \beta_1)^2(1 - \beta_2)^{1/2}\theta}, \\ (1 - \beta_1) \sum_{j=1}^t \mathbb{E}[\langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j \rangle] \sum_{k=j}^t \rho_k \beta_1^{k-j} &\geq -\frac{m\theta B_{\mathbf{z}}^2}{2} - \frac{\sqrt{m}(\theta^2 + M_{\mathbf{z}}^2) \sum_{j=1}^t \rho_j^2}{2(1 - \beta_1)^2(1 - \beta_2)^{1/2}\theta}. \end{aligned}$$

The second inequality above holds reversely because of the concavity of  $\mathcal{L}$  about  $\mathbf{z}$ .

For the last two terms in the right side of (4.9), by essentially the same arguments as those in the proof of Lemma 8, we have the following lemma.

**Lemma 12** For any deterministic or stochastic vector  $(\mathbf{x}, \mathbf{z})$  with  $\mathbf{x} \in X$  and  $\mathbf{z} \in Z$ , it holds for any positive number sequence  $\{\gamma_j\}_{j=1}^K$  and  $t \in [K]$  that

$$\begin{aligned} & \sum_{j=1}^t \gamma_j \mathbb{E} \left[ \langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j - \tilde{\nabla}_{\mathbf{z}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle - \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle \right] \\ & \leq \frac{1}{2} \left( nB_{\mathbf{x}}^2 + mB_{\mathbf{z}}^2 + (M_{\mathbf{x}}^2 + M_{\mathbf{z}}^2) \sum_{j=1}^t \gamma_j^2 \right). \end{aligned}$$

where  $\tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) = \mathbb{E}[\mathbf{u}^j \mid \mathcal{H}^j]$  and  $\tilde{\nabla}_{\mathbf{z}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) = \mathbb{E}[\mathbf{w}^j \mid \mathcal{H}^j]$ .

Note that the bounds in Lemmas 11 and 12 do not depend on  $\mathbf{x}$  and  $\mathbf{z}$ . Below, we use the established lemmas to show the main convergence rate result of Algorithm 2. Following [24], we adopt the expected primal-dual gap to measure the quality of a solution  $(\bar{\mathbf{x}}, \bar{\mathbf{z}}) \in X \times Z$  for (4.2), namely,

$$\epsilon(\bar{\mathbf{x}}, \bar{\mathbf{z}}) = \mathbb{E} \left[ \max_{\mathbf{z} \in Z} \mathcal{L}(\bar{\mathbf{x}}, \mathbf{z}) - \min_{\mathbf{x} \in X} \mathcal{L}(\mathbf{x}, \bar{\mathbf{z}}) \right].$$

**Theorem 3** Under Assumptions 4 and 5, let  $\{(\mathbf{x}^j, \mathbf{z}^j)\}$  be generated from Algorithm 2. Suppose  $\alpha_k = c\rho_k$ ,  $\forall k$  for some constant  $c > 0$ . Then for any  $t \in [K]$ , it holds

$$\begin{aligned} \epsilon(\bar{\mathbf{x}}^t, \bar{\mathbf{z}}^t) & \leq \frac{1}{2(1-\beta_1) \sum_{j=1}^t \alpha_j} \left( \frac{\sqrt{n}(\theta^2 + M_{\mathbf{x}}^2) \sum_{j=1}^t \alpha_j^2 + c\sqrt{m}(\theta^2 + M_{\mathbf{z}}^2) \sum_{j=1}^t \rho_j^2}{(1-\beta_1)^2(1-\beta_2)^{1/2}\theta} \right. \\ & \quad \left. + \theta(nB_{\mathbf{x}}^2 + cmB_{\mathbf{z}}^2) + (nB_{\mathbf{x}}^2 + mB_{\mathbf{z}}^2) \sum_{j=1}^t \alpha_j^2 \right). \end{aligned}$$

$$\text{where } \bar{\mathbf{x}}^t = \frac{\sum_{j=1}^t \sum_{k=j}^t \alpha_k \beta_1^{k-j} \mathbf{x}^j}{\sum_{j=1}^t \sum_{k=j}^t \alpha_k \beta_1^{k-j}}, \text{ and } \bar{\mathbf{z}}^t = \frac{\sum_{j=1}^t \sum_{k=j}^t \alpha_k \beta_1^{k-j} \mathbf{z}^j}{\sum_{j=1}^t \sum_{k=j}^t \alpha_k \beta_1^{k-j}}.$$

**Proof** By the convexity–concavity of  $\mathcal{L}$  and Lemma 10, we have that for any  $\mathbf{x} \in X$  and  $\mathbf{z} \in Z$ ,

$$\begin{aligned} \mathcal{L}(\bar{\mathbf{x}}^t, \mathbf{z}) - \mathcal{L}(\mathbf{x}, \bar{\mathbf{z}}^t) & \leq \sum_{j=1}^t \left( \left( \mathcal{L}(\mathbf{x}^j, \mathbf{z}) - \mathcal{L}(\mathbf{x}, \mathbf{z}^j) \right) \frac{\sum_{k=j}^t \alpha_k \beta_1^{k-j}}{\sum_{j=1}^t \sum_{k=j}^t \alpha_k \beta_1^{k-j}} \right) \\ & \stackrel{(4.9)}{\leq} \frac{1}{\sum_{j=1}^t \sum_{k=j}^t \alpha_k \beta_1^{k-j}} \left( \sum_{j=1}^t \left( \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j \rangle - \langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j \rangle \right) \sum_{k=j}^t \alpha_k \beta_1^{k-j} \right. \\ & \quad \left. + \sum_{j=1}^t \left( \langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j - \tilde{\nabla}_{\mathbf{z}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle - \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle \right) \sum_{k=j}^t \alpha_k \beta_1^{k-j} \right). \end{aligned}$$

Let  $\mathbf{x} \in \arg \min_{\mathbf{x} \in X} \mathcal{L}(\mathbf{x}, \bar{\mathbf{z}}^t)$  and  $\mathbf{z} \in \arg \max_{\mathbf{z} \in Z} \mathcal{L}(\bar{\mathbf{x}}^t, \mathbf{z})$ , and take the expectation on the above inequality. By Lemmas 11 and 12 with  $\gamma_j = \sum_{k=j}^t \alpha_k \beta_1^{k-j} (1-\beta_1) \leq \alpha_j$  for

the numerator, and  $\sum_{j=1}^t \sum_{k=j}^t \alpha_k \beta_1^{k-j} \geq \sum_{j=1}^t \alpha_j$  from (2.10) for the denominator, we have

$$\begin{aligned} \epsilon(\bar{\mathbf{x}}^t, \bar{\mathbf{z}}^t) &= \mathbb{E} \left[ \max_{\mathbf{z} \in \mathcal{Z}} \mathcal{L}(\bar{\mathbf{x}}^t, \mathbf{z}) - \min_{\mathbf{x} \in \mathcal{X}} \mathcal{L}(\mathbf{x}, \bar{\mathbf{z}}^t) \right] \\ &\leq \frac{1}{(1-\beta_1) \sum_{j=1}^t \alpha_j} \left( \frac{\theta(nB_{\mathbf{x}}^2 + cmB_{\mathbf{z}}^2)}{2} + \frac{1}{2} (nB_{\mathbf{x}}^2 + mB_{\mathbf{z}}^2 + (M_{\mathbf{x}}^2 + M_{\mathbf{z}}^2) \sum_{j=1}^t \alpha_j^2) \right. \\ &\quad \left. + \frac{\sqrt{n}(\theta^2 + M_{\mathbf{x}}^2) \sum_{j=1}^t \alpha_j^2 + c\sqrt{m}(\theta^2 + M_{\mathbf{z}}^2) \sum_{j=1}^t \rho_j^2}{2(1-\beta_1)^2(1-\beta_2)^{1/2}\theta} \right). \end{aligned}$$

Simplifying the above inequality gives the desired result.  $\square$

Below, we specify the choices of  $\{\alpha_k\}_{k=1}^K$  and  $\{\rho_k\}_{k=1}^K$  and obtain sublinear convergence of Algorithm 2. Corollary 3 can be proved by essentially the same arguments as those in the proof of Corollary 1, and the proof of Corollary 4 is given in Appendix I.

**Corollary 3** (Convergence rate with constant step size) *Given any positive integer  $K$ , set  $\alpha_j = \frac{\alpha}{\sqrt{j}}$  and  $\rho_j = \frac{\rho}{\sqrt{j}}$  for all  $j \in [K]$  and some positive constants  $\alpha, \rho$ . Let  $\{(\mathbf{x}^k, \mathbf{z}^k)\}$  be the sequence generated from Algorithm 2. Let  $\bar{\mathbf{x}}^K = \frac{\sum_{j=1}^K (1-\beta_1^{K-j+1}) \mathbf{x}^j}{\sum_{j=1}^K (1-\beta_1^{K-j+1})}$  and  $\bar{\mathbf{z}}^K = \frac{\sum_{j=1}^K (1-\beta_1^{K-j+1}) \mathbf{z}^j}{\sum_{j=1}^K (1-\beta_1^{K-j+1})}$ . Then*

$$\begin{aligned} \epsilon(\bar{\mathbf{x}}^K, \bar{\mathbf{z}}^K) &\leq \frac{1}{2(1-\beta_1)\alpha\sqrt{K}} \left( \left( \theta(nB_{\mathbf{x}}^2 + \frac{m\alpha}{\rho} B_{\mathbf{z}}^2) + (nB_{\mathbf{x}}^2 + mB_{\mathbf{z}}^2) + \alpha^2(M_{\mathbf{x}}^2 + M_{\mathbf{z}}^2) \right) \right. \\ &\quad \left. + \frac{\alpha^2\sqrt{n}(\theta^2 + M_{\mathbf{x}}^2) + \alpha\rho\sqrt{m}(\theta^2 + M_{\mathbf{z}}^2)}{(1-\beta_1)^2(1-\beta_2)^{1/2}\theta} \right). \end{aligned}$$

**Corollary 4** (Convergence rate with varying step size) *Set  $\alpha_j = \frac{\alpha}{\sqrt{j+1}}$  and  $\rho_j = \frac{\rho}{\sqrt{j+1}}$  for all  $j \in [K]$  and some positive constants  $\alpha, \rho$ . Let  $\{(\mathbf{x}^k, \mathbf{z}^k)\}$  be generated from Algorithm 2. Let  $\bar{\mathbf{x}}^K = \frac{\sum_{j=1}^K \sum_{k=j}^K \alpha_k \beta_1^{k-j} \mathbf{x}^j}{\sum_{j=1}^K \sum_{k=j}^K \alpha_k \beta_1^{k-j}}$  and  $\bar{\mathbf{z}}^K = \frac{\sum_{j=1}^K \sum_{k=j}^K \alpha_k \beta_1^{k-j} \mathbf{z}^j}{\sum_{j=1}^K \sum_{k=j}^K \alpha_k \beta_1^{k-j}}$  for  $K \geq 2$ . Then*

$$\begin{aligned} \epsilon(\bar{\mathbf{x}}^K, \bar{\mathbf{z}}^K) &\leq \frac{\log(1+K)}{4(1-\beta_1)\alpha(\sqrt{K+2}-\sqrt{2})} \left( \left( \theta(nB_{\mathbf{x}}^2 + \frac{m\alpha}{\rho} B_{\mathbf{z}}^2) + (nB_{\mathbf{x}}^2 + mB_{\mathbf{z}}^2) \right) \right. \\ &\quad \left. + \frac{\alpha^2\sqrt{n}(\theta^2 + M_{\mathbf{x}}^2) + \alpha\rho\sqrt{m}(\theta^2 + M_{\mathbf{z}}^2)}{(1-\beta_1)^2(1-\beta_2)^{1/2}\theta} \right). \end{aligned}$$

## 5 Numerical experiments

In this section, we compare the numerical performance of our proposed APriD method in Algorithm 1 to CSA [18] and MSA [24]. The latter two methods have been reviewed



in Sect. 1.2. In all our experiments, we take  $\beta_1 = 0.9$ ,  $\beta_2 = 0.99$ ,  $\theta = 10$  in APriD;  $s = 1$  and  $\eta_k = \eta = 0.04$  in CSA. In [18], the output of CSA is the weighted average of  $\mathbf{x}^t$  over  $t \in \mathcal{B}^k = \{t \in [K] \mid \widehat{G}_t \leq \eta_t\}$ . Note that  $\mathcal{B}^k$  may be empty for a small  $k$ . Hence, we also computed the weighted average of  $\mathbf{x}^t$  over all  $t \in [k]$ . We call the result of CSA by the former weighted average as CSA1 and the latter as CSA2.

In [24], the step sizes of MSA for updating  $\mathbf{x}$  and  $\mathbf{z}$  are both equal to  $\gamma_k$ , as in (1.11). In our experiments, we chose different step sizes for the  $\mathbf{x}$  and  $\mathbf{z}$  updates in order to have better performance. More specifically, we did the updates:

$$\mathbf{x}^{k+1} = \text{Proj}_X(\mathbf{x}^k - \alpha_k \mathbf{u}^k) \quad \text{and} \quad \mathbf{z}^{k+1} = \text{Proj}_Z(\mathbf{z}^k + \rho_k \mathbf{w}^k),$$

where  $\alpha_k$  and  $\rho_k$  are the step sizes. We used the same  $\alpha_k$  and  $\rho_k$  for MSA and APriD. Hence, MSA can be viewed as a non-adaptive counterpart of APriD. We tried different pairs of  $(\alpha_k, \rho_k)$  for MSA. It turned out that the best pair for MSA was also the best for APriD. With the step size  $\gamma_k$  in CSA, we denote the step sizes in the three algorithms as  $(\alpha_k, \rho_k, \gamma_k)$ .

Three problems were tested in our experiments. The first one is NPC in a finite-sum form; the second one is QCQP whose objective and constraint are both in an expectation form; the third one is QCQP with scenario approximation (i.e., with many quadratic constraints). For the third problem, we also compared to PDSG-adp in [38], for which the update is given in (1.10). All experiments were run on MATLAB installed on a MacBook Pro with one 2.9 GHz Dual-Core Intel Core i5 processor and 16 GB memory.

## 5.1 Neyman–Pearson classification problem

In this subsection, we compare the algorithms on solving instances of NPC (1.3). We take the linear classifier  $h(\mathbf{x}; \mathbf{a}) = \mathbf{x}^\top \mathbf{a}$  and the convex surrogate  $\varphi(z) = \log(1 + e^z)$ . This way, (1.3) reduces to

$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbb{E}[\log(1 + e^{-\mathbf{x}^\top \mathbf{a}}) \mid b = +1], \\ \text{s.t.} \quad & \mathbb{E}[\log(1 + e^{\mathbf{x}^\top \mathbf{a}}) \mid b = -1] - c \leq 0. \end{aligned} \quad (5.1)$$

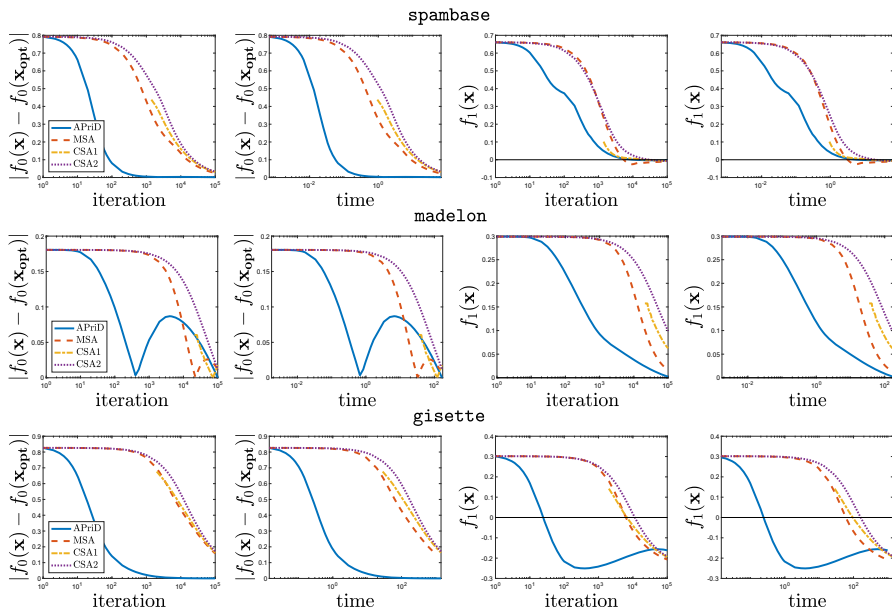
Given a training data set with  $n^+$  positive-class samples  $\{\mathbf{a}_i^+\}_{i=1}^{n^+}$  and  $n^-$  negative-class samples  $\{\mathbf{a}_i^-\}_{i=1}^{n^-}$ , we can obtain and solve a scenario approximation of (5.1), namely,

$$\begin{aligned} \min_{\mathbf{x}} \quad & f_0(\mathbf{x}) \equiv \frac{1}{n^+} \sum_{i=1}^{n^+} \log(1 + e^{-\mathbf{x}^\top \mathbf{a}_i^+}), \\ \text{s.t.} \quad & f_1(\mathbf{x}) \equiv \frac{1}{n^-} \sum_{i=1}^{n^-} \log(1 + e^{\mathbf{x}^\top \mathbf{a}_i^-}) - \widehat{c} \leq 0. \end{aligned} \quad (5.2)$$

**Table 1** Characteristics of data sets and algorithm parameters

Data set	$d$	$(n^+, n^-)$	$\hat{c}$	$(J^0, J^1, J^g)$	$K$	$(\alpha, \rho, \gamma)$
spambase [7]	57	(1813, 2788)	$-\log(0.7)$	(10, 10, 100)	$10^5$	(10, 1, 10)
madelon [12]	500	(1300, 1300)	$-\log(0.6)$	(10, 10, 100)	$10^5$	(10, 1, 10)
gisette [12]	2000	(3500, 3500)	$-\log(0.6)$	(30, 30, 100)	$10^5$	(10, 1, 10)

For the data sets,  $d$  is the number of features of each sample;  $n^+$  and  $n^-$  are respectively the numbers of positive and negative samples.  $\hat{c}$  controls the level of the false-positive error in (5.2).  $J^0$  and  $J^1$  are the batch sizes for evaluating the stochastic gradients of the objective and the constraint functions by the three methods;  $J^g$  is the batch size for obtaining a stochastic estimation of the constraint function value in CSA.  $K$  is the total number of iterations.  $(\alpha_k, \rho_k, \gamma_k) = (\frac{\alpha}{\sqrt{k}}, \frac{\rho}{\sqrt{k}}, \frac{\gamma}{\sqrt{k}})$ ,  $\forall k \leq K$  are constant step sizes



**Fig. 1** The objective error (Left two columns) and constraint value (Right two columns) by three compared methods on solving Example 5.1 with three data sets: Spambase(Top); Madelon(Middle); Gisette(Bottom). The curve with legend “0” represents the zero-error curve

According to [27], we set  $\hat{c} = c - \frac{\kappa}{\sqrt{n}}$ , where  $\kappa$  is a positive constant in order to ensure that the feasible solution of (5.2) is also feasible for (5.1) in a given high probability.

We use three data sets. The information of the data sets and the algorithm parameters are given in Table 1. Before feeding the data sets into the methods, we preprocess the data sets. We first normalize the data sets feature-wisely to have mean 0 and standard deviation 1 and then scale each sample to have unit 2-norm.

We apply a deterministic method iALM [40] to compute the “optimal” solution  $\mathbf{x}_{\text{opt}}$ . The selected  $\hat{c}$  makes sure  $\mathbf{x}_{\text{opt}}$  is feasible. The results, in terms of iteration and time (in seconds), by all compared methods are shown in Fig. 1. The left two columns

of Fig. 1 are the objective error at the output solutions. The right two columns show the value of the constraint at the output solutions.

We can see from Fig. 1 that APriD performs the best on the data sets `spambase` and `gisette`, in terms of either objective error or feasibility. For the data set `madelon`, APriD is the fastest to achieve feasibility. We also note the lack of results for CSA1 at the beginning iterations because  $\mathcal{B}$  is empty. When CSA1 has results, it is better than CSA2 since the former only takes ergodic mean on “good” solutions but the later one takes ergodic mean on “all” solutions.

## 5.2 QCQP in expectation form

In this subsection, we conduct experiments on the QCQP in an expectation form:

$$\begin{aligned} \min_{\mathbf{x} \in X} f_0(\mathbf{x}) &\equiv \mathbb{E}[\frac{1}{2} \|H_\xi \mathbf{x} - \mathbf{c}_\xi\|^2], \\ \text{s.t. } f_1(\mathbf{x}) &\equiv \mathbb{E}[\frac{1}{2} \mathbf{x}^\top Q_\xi \mathbf{x} + \mathbf{a}_\xi^\top \mathbf{x} - b_\xi] \leq 0. \end{aligned} \quad (5.3)$$

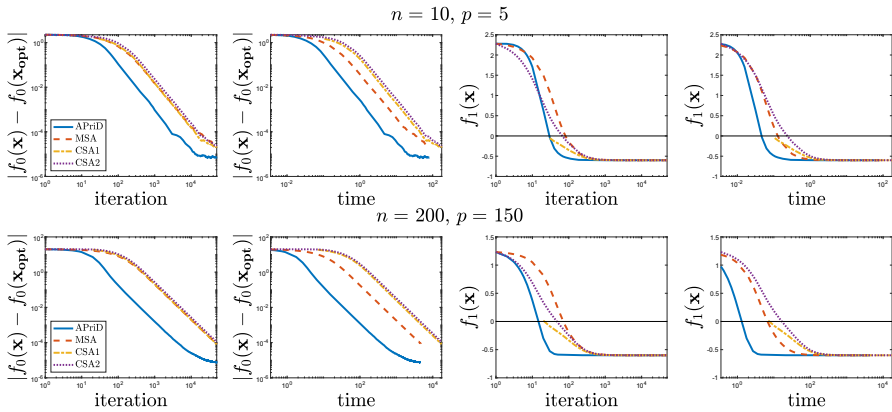
Here, we set  $X = [-10, 10]^n$ ,  $\xi$  is a random variable,  $H_\xi \in \mathbb{R}^{p \times n}$  and  $\mathbf{c}_\xi \in \mathbb{R}^p$  are randomly generated, and their components are generated by standard Gaussian distribution and then normalized. The entries of  $\mathbf{a}_\xi \in \mathbb{R}^n$  are also generated by standard Gaussian distribution and then normalized;  $Q_\xi \in \mathbb{R}^{n \times n}$  is a randomly generated symmetric positive semidefinite matrix with unit 2-norm;  $b_\xi$  follows a uniform distribution on the open interval  $(0, 1, 1.1)$ .

While running the algorithms, we generate  $H_\xi$ ,  $\mathbf{c}_\xi$ ,  $Q_\xi$ ,  $\mathbf{a}_\xi$ ,  $b_\xi$  based on the above distribution once needed for function evaluation or gradient evaluation. Hence, the function value and gradient direction are both unbiased estimations. At a weighted iterate  $\mathbf{x}$ , we generate another  $10^5$  samples to evaluate the objective value and the constraint function value. Also, we obtain the “optimal” solution  $\mathbf{x}_{\text{opt}}$  by using CVX [9,10] to solve a sample approximation problem with the generated  $10^5$  samples.

We test the compared algorithms on QCQP instances of size  $(n, p) = (10, 5)$  and  $(n, p) = (200, 150)$ . In both instances, we run  $K = 5 \times 10^4$  iterations, and we set batch size  $J = 10$  for obtaining stochastic gradients of both objective and constraint functions in all the three methods and  $J^g = 100$  for obtaining a stochastic estimation of the constraint value in CSA. Step sizes are set to  $(\alpha_k, \rho_k, \gamma_k) = (\frac{10}{\sqrt{k}}, \frac{10}{\sqrt{k}}, \frac{\sqrt{10}}{\sqrt{k}})$  for all  $k$ . The results in terms of iteration and time (in seconds) are shown in Fig. 2. From the results, we see again that APriD significantly outperforms over other two compared methods.

## 5.3 Finite-sum structured QCQP with many constraints

In this subsection, we test the algorithms on the QCQP with a finite-sum objective and many constraints:

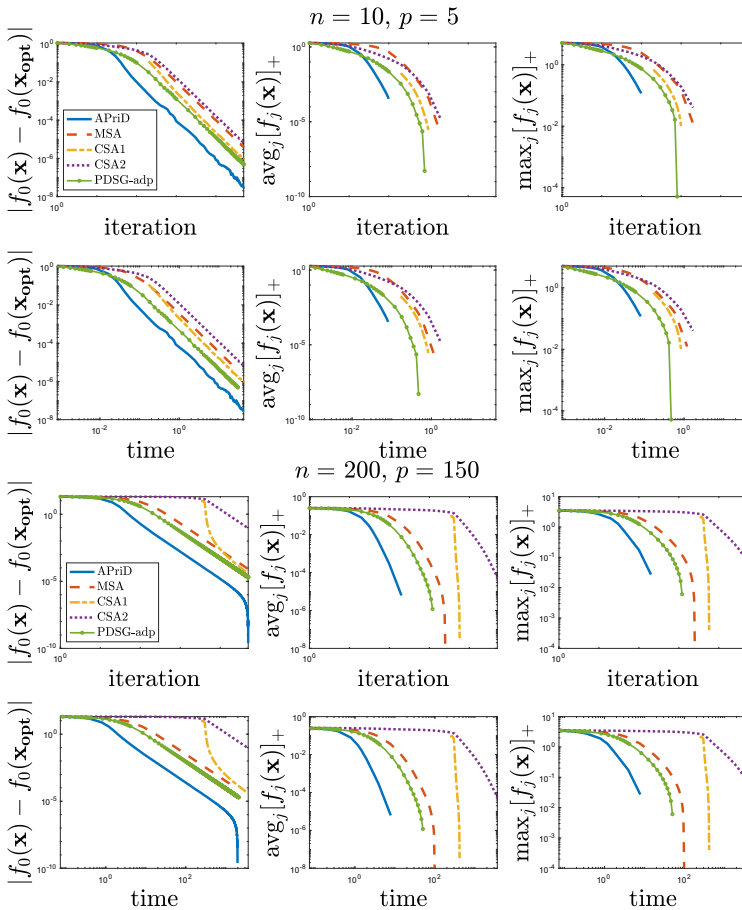


**Fig. 2** The objective error (Left two columns) and constraint value (Right two columns) by three methods on solving QCQP instances of (5.3)

$$\begin{aligned}
 \min_{\mathbf{x} \in X} f_0(\mathbf{x}) &\equiv \frac{1}{2N} \sum_{i=1}^N \|H_i \mathbf{x} - \mathbf{c}_i\|^2, \\
 \text{s.t. } f_j(\mathbf{x}) &\equiv \frac{1}{2} \mathbf{x}^\top Q_j \mathbf{x} + \mathbf{a}_j^\top \mathbf{x} - b_j \leq 0, \quad j \in [M].
 \end{aligned} \tag{5.4}$$

We set  $X = [-10, 10]^n$  in the experiment.  $H_i, \mathbf{c}_i$  for  $i \in [N]$  and  $Q_j, \mathbf{a}_j, b_j$  for  $j \in [M]$  are independently generated from the same distribution as  $H_\xi, \mathbf{c}_\xi, Q_\xi, \mathbf{a}_\xi, b_\xi$  in Sect. 5.2. Two different-size QCQP instances are generated: one with  $(n, p) = (10, 5)$  and the other with  $(n, p) = (200, 150)$ . We set  $N = M = 10^4$  in both instances. Besides CSA and MSA, we also compare APriD with the adaptive method in [38], called PDSG-adp. The update of PDSG-adp is shown in (1.10). For each update of the compared methods, we randomly select 10 component functions of the objective and 10 constraint functions for evaluating stochastic gradients, and for CSA, we randomly pick 100 constraint functions to obtain a stochastic estimation of  $g(\mathbf{x}) \equiv \sum_{j=1}^M [f_j(\mathbf{x})]_+$ . In both instances, we run  $K = 5 \times 10^4$  iterations with step sizes  $(\alpha_k, \rho_k, \gamma_k) = (\frac{10}{\sqrt{K}}, \frac{\sqrt{10}}{\sqrt{K}}, \frac{10}{\sqrt{K}})$  for all  $k$ . We run PDSG-adp to  $K = 5 \times 10^4$  iterations with  $\alpha_k = \frac{20}{\sqrt{K}}, \rho_k = \frac{\sqrt{10}}{\sqrt{K}}, \forall k$  and  $\eta = 0.1$ .

For the smaller-size instance, we obtain the optimal solution  $\mathbf{x}_{\text{opt}}$  by CVX [9, 10]; for the larger one, we use as an estimated optimal solution  $\mathbf{x}_{\text{opt}}$  that is feasible and has the smallest objective value among all iterates from APriD, CSA, MSA and PDSG-adp. We report the objective error, the averaged constraint violation measured by  $\frac{1}{M} \sum_{j=1}^M [f_j(\mathbf{x})]_+$ , and the maximum constraint violation measured by  $\max_{j \in [M]} [f_j(\mathbf{x})]_+$ . The results in terms of iteration and time (in seconds) are shown in Fig. 3. From the results, we see again that APriD outperforms over the other methods in terms of any of the three measures we use.



**Fig. 3** The objective error (Left), averaged constraint violation (Middle), maximum constraint violation (Right) by four methods on solving QCQP instances of (5.4). Rows 1 and 3 are with respect to iteration; rows 2 and 4 are with respect to time (in seconds)

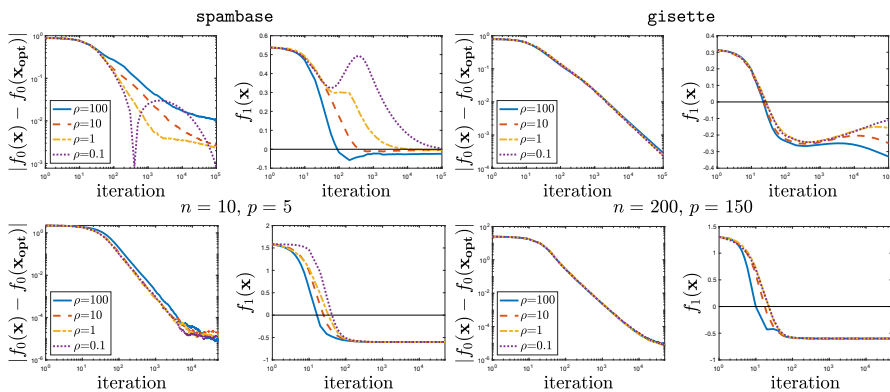
## 5.4 Computing time comparison

In Table 2, we compare the total running time (in seconds) of all the tested methods in Sects. 5.1–5.3. From the table, we see that although APriD needs extra computation in (2.3)–(2.7), it takes similar amount of time (and thus has similar per-iteration cost) as MSA in all examples and PDSG-adp for the example in Sect. 5.3. CSA always takes more time than the other methods because of the extra estimation of  $g(\mathbf{x}^k)$  in each iteration.

**Table 2** Total running time (in seconds) of all the methods that are tested in Sect. 5.1–5.3

Example	Data or size	APriD	MSA	CSA	PDSG-adp
NPC (5.2)	spambase	60.3	58.8	67.1	–
	madelon	175.5	144.6	167.4	–
	gisette	983.8	954.8	1416.6	–
QCQP (5.3)	(10, 5)	78.8	79.6	177.1	–
	(200, 150)	4451.9	4722.6	17592.7	–
QCQP (5.4)	(10, 5)	40.7	30.0	42.2	29.7
	(200, 150)	1964.2	1966.0	3638.1	2087.2

PDSG-adp is only applied to QCQP (5.4)



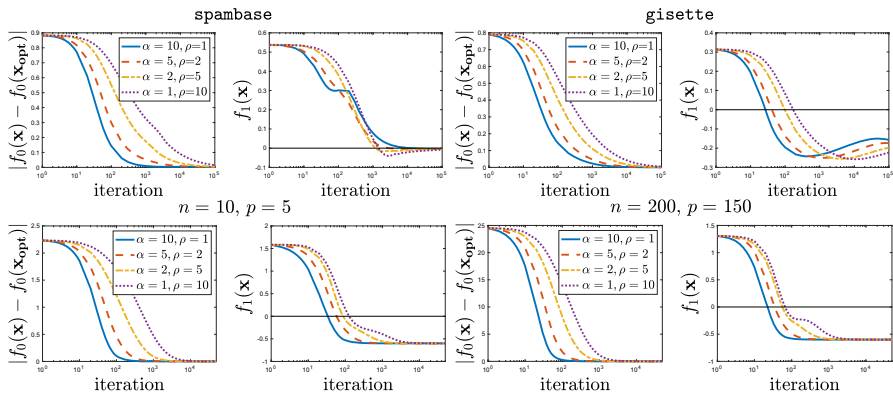
**Fig. 4** The objective error and constraint violation by APriD with  $\theta = 10$ ,  $\alpha = 10$  and different values of  $\rho$  on solving instances of NPC (5.2) with data sets spambase (TopLeft) and gisette (TopRight) and on solving instances of QCQP (5.3) with  $(n, p) = (10, 5)$  (BottomLeft) and  $(n, p) = (200, 150)$  (BottomRight)

## 5.5 Effect of hyper-parameters on algorithm performance

In this subsection, we test how the choices of  $\alpha$ ,  $\rho$  (in the constant step size  $(\alpha_k, \rho_k) = (\frac{\alpha}{K}, \frac{\rho}{K})$  for all  $k \leq K$ ) and  $\theta$  affect the performance of APriD. For all tests in this subsection, we apply APriD to solve instances of NPC (5.2) with the spambase and gisette data sets and QCQP instances (5.3) with  $(n, p) = (10, 5)$  and  $(200, 150)$ . It turns out that APriD can perform reasonably well for a wide range of values of the hyper-parameters.

First, we test the effect of  $\rho$  by fixing  $\theta = 10$  and  $\alpha = 10$ . From the results in Fig. 4, we see that the algorithm performs similarly well with different values of  $\rho$ . The difference is most obvious for the instance of NPC on the spambase data set. Zooming in details of the curves, we can observe that with the biggest  $\rho = 100$ , the objective error decreases slowest but the constraint value decreases fastest, and the convergence behavior with the smallest  $\rho = 0.1$  is exactly the opposite.

Second, we fix  $\theta = 10$  and test the effect of  $(\alpha, \rho)$ . For simplicity, we fix the product  $\alpha\rho = 10$  and test the algorithm with different pairs of  $(\alpha, \rho)$ . The results are shown in Fig. 5. It turns out that the algorithm with a larger  $\alpha$  tends to converge faster in terms



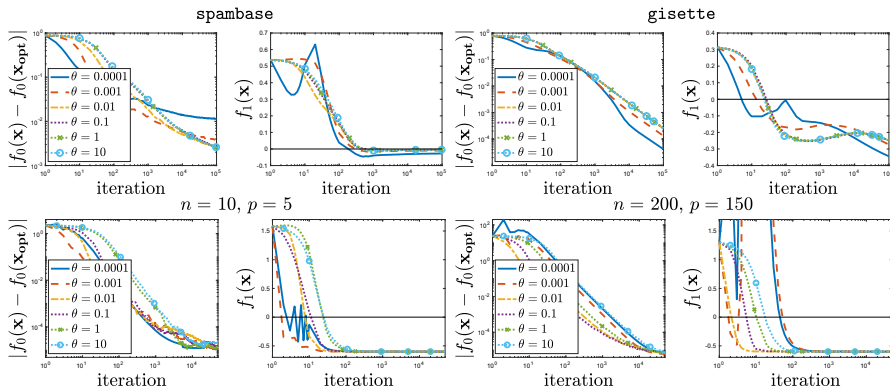
**Fig. 5** The objective error and constraint violation by APriD with different pairs of  $(\alpha, \rho)$  and a fixed  $\theta = 10$ , on solving instances of NPC (5.2) with data sets spambase (TopLeft) and gisette (TopRight) and on solving instances of QCQP (5.3) with  $(n, p) = (10, 5)$  (BottomLeft) and  $(n, p) = (200, 150)$  (BottomRight)

of both the objective error and the constraint violation. Nevertheless, the influence by the choice of  $(\alpha, \rho)$  is not severe, and the algorithm with all four different pairs of  $(\alpha, \rho)$  can perform reasonably well.

Finally, we test the effect of  $\theta$  by fixing  $\alpha = 10$  and  $\rho = 10$ . We vary  $\theta \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ . The results are shown in Fig. 6. For the instances of NPC (5.2), there is almost no difference for  $\theta \in \{0.1, 1, 10\}$ . This is probably because these values of  $\theta$  do not trigger the clipping. For the spambase data set, the best results appear to be given by  $\theta = 10^{-2}$  and  $\theta = 10^{-3}$ . The algorithm can still perform well with  $\theta = 10^{-4}$ , but the corresponding feasibility curve is less smooth. Similar observations are made to the gisette data set. For QCQP instances of (5.3), the performance of the algorithm is more affected by the value of  $\theta$ . The algorithm performs better as  $\theta$  decreases from 10 to  $10^{-2}$ , in terms of both suboptimality and infeasibility. However, for the case of  $n = 200$  and  $p = 150$ , the infeasibility increases rapidly in the beginning iterations. These observations match with our discussion in Remark 2, i.e., the best value of  $\theta$  should not be extremely small.

## 6 Conclusions

We have proposed an adaptive primal-dual stochastic gradient method (SGM) for solving expectation-constrained convex stochastic programming. The method is designed based on the Lagrangian function. At each iteration, it first inquires an unbiased stochastic estimation of the subgradient of the Lagrangian function, and then it performs an adaptive SGM update to the primal variables and a vanilla SGM step to the dual variables. The method has also been extended with a modification to solve stochastic convex-concave minimax problems. For both methods, we have established the convergence rate of  $O(1/\sqrt{k})$ , where  $k$  is the number of inquiries of the stochastic subgradient. Numerical experiments on three examples demonstrate its superior practical performance over two state-of-the-art methods.



**Fig. 6** The objective error and constraint violation by APriD with  $\alpha = 10, \rho = 10$  and  $\theta \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$  on solving instances of NPC (5.2) with data sets spambase (TopLeft) and gisette (TopRight) and on solving instances of QCQP (5.3) with  $(n, p) = (10, 5)$  (BottomLeft) and  $(n, p) = (200, 150)$  (BottomRight)

**Acknowledgements** The authors would like to thank three anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper and also for the careful testing on our codes. The authors are partly supported by the NSF award 2053493 and the RPI-IBM AIRC faculty fund.

**Funding** This work was partly supported by NSF Award 2053493.

**Data Availability Statement** This manuscript has associated data in a data repository at <https://github.com/RPIOPT/APriD>.

**Conflict of interest** The authors declare that they have no conflict of interest.

## A Proof of Lemma 2

**Proof** First, consider the case of non-constant primal step size. By  $\eta_1 = \sum_{i=1}^K \alpha_i \beta_1^{i-1}$  and the  $\eta$ -update in (2.8), we have  $\eta_k = \sum_{i=k}^K \alpha_i \beta_1^{i-k}, k \in [K]$ , and thus the  $\rho$ -update becomes

$$\rho_k = \frac{\rho_{k-1}}{\beta_1 + \frac{\alpha_{k-1}}{\sum_{i=k}^K \alpha_i \beta_1^{i-k}}}.$$

By the above equation, we have for  $2 \leq j \leq K$ ,

$$\rho_j = \frac{\rho_{j-1}}{\beta_1 + \frac{\alpha_{j-1}}{\sum_{k=j}^K \alpha_k \beta_1^{k-j}}} \leq \frac{\rho_{j-1}}{\beta_1 + \frac{\alpha_{j-1}}{\alpha_{j-1} \sum_{k=j}^K \beta_1^{k-j}}} \leq \frac{\rho_{j-1}}{\beta_1 + \frac{1}{\sum_{k=j}^K \beta_1^{k-j}}} = \rho_{j-1},$$



where the inequality follows from the non-increasing monotonicity of  $\{\alpha_j\}_{j=1}^K$  and  $\beta_1 \in (0, 1)$ . Hence,  $\{\rho_j\}_{j=1}^K$  is a non-increasing sequence. Using (2.8) again, we have for  $2 \leq j \leq t \leq K$ ,

$$\rho_j = \frac{\rho_{j-1}}{\beta_1 + \frac{\alpha_{j-1}}{\sum_{k=j}^K \alpha_k \beta_1^{k-j}}} \geq \frac{\rho_{j-1}}{\beta_1 + \frac{\alpha_{j-1}}{\sum_{k=j}^t \alpha_k \beta_1^{k-j}}} = \frac{\sum_{k=j}^t \alpha_k \beta_1^{k-j}}{\sum_{k=j-1}^t \alpha_k \beta_1^{k-(j-1)}} \rho_{j-1},$$

which clearly implies the inequality in (2.11).

For  $j = 1$ , (2.12) holds because  $\beta \in (0, 1)$ . To show it holds for  $2 \leq j \leq K$ , we rewrite  $\rho_j$  and obtain

$$\begin{aligned} \rho_j &= \frac{\rho_{j-1}}{\beta_1 + \frac{\alpha_{j-1}}{\sum_{k=j}^K \alpha_k \beta_1^{k-j}}} = \frac{\sum_{k=j}^K \alpha_k \beta_1^{k-j}}{\sum_{k=j-1}^K \alpha_k \beta_1^{k-(j-1)}} \rho_{j-1} \\ &= \frac{\sum_{k=j}^K \alpha_k \beta_1^{k-j}}{\sum_{k=j-1}^K \alpha_k \beta_1^{k-(j-1)}} \times \frac{\sum_{k=j-1}^K \alpha_k \beta_1^{k-(j-2)}}{\sum_{k=j-2}^K \alpha_k \beta_1^{k-(j-2)}} \times \cdots \times \frac{\sum_{k=2}^K \alpha_k \beta_1^{k-2}}{\sum_{k=1}^K \alpha_k \beta_1^{k-1}} \rho_1 \\ &= \frac{\sum_{k=j}^K \alpha_k \beta_1^{k-j}}{\sum_{k=1}^K \alpha_k \beta_1^{k-1}} \rho_1 \leq \frac{\frac{\alpha_j}{(1-\beta_1)}}{\sum_{k=1}^K \alpha_k \beta_1^{k-1}} \rho_1 \leq \frac{\frac{\alpha_j}{(1-\beta_1)}}{\alpha_1} \rho_1 = \frac{\rho_1 \alpha_j}{\alpha_1 (1-\beta_1)}, \end{aligned}$$

where the third equation recursively applies the second equation, and the inequalities hold by the two inequalities in (2.10).

Now, consider the case of constant primal step size, i.e.,  $\alpha_j = \alpha_1$  for all  $j \in [K]$ . We can prove  $\eta_j = \frac{\alpha_1}{1-\beta_1}$  for all  $j \in [K]$  by the induction, and thus

$$\rho_j = \frac{\rho_{j-1}}{\beta_1 + \frac{\alpha_1}{\alpha_1(1-\beta_1)}} = \frac{\rho_{j-1}}{\beta_1 + (1-\beta_1)} = \rho_{j-1}, \quad \forall j \geq 2,$$

which completes the proof.  $\square$

## B Proof of Lemma 3

**Proof** As  $(\mathbf{x}^*, \mathbf{z}^*)$  satisfies the KKT conditions in Assumption 3, there are  $\tilde{\nabla} f_i(\mathbf{x}^*)$ ,  $\forall i \in [M]$  such that

$$-\sum_{i=1}^M z_i^* \tilde{\nabla} f_i(\mathbf{x}^*) \in \partial f_0(\mathbf{x}^*) + \mathcal{N}_X(\mathbf{x}^*).$$

From the convexity of  $f_0$  and  $X$ , it follows that  $f_0(\mathbf{x}) - f_0(\mathbf{x}^*) \geq -\langle \sum_{i=1}^M z_i^* \tilde{\nabla} f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle$ ,  $\forall \mathbf{x} \in X$ . Since  $f_i$  is convex for each  $i \in [M]$ , we have  $f_i(\mathbf{x}) - f_i(\mathbf{x}^*) \geq \langle \tilde{\nabla} f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle$ . Noticing  $\mathbf{z}^* \geq \mathbf{0}$ , we have for any  $\mathbf{x} \in X$ ,

$$f_0(\mathbf{x}) - f_0(\mathbf{x}^*) \geq - \sum_{i=1}^M z_i^* \langle \tilde{\nabla} f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq - \sum_{i=1}^M z_i^* (f_i(\mathbf{x}) - f_i(\mathbf{x}^*)).$$

Because  $z_i^* f_i(\mathbf{x}^*) = 0$  for all  $i \in [M]$ , we obtain (3.7).

Furthermore, for any  $\mathbf{z} \geq \mathbf{0}$ , we have  $\langle \mathbf{z}, \mathbf{f}(\mathbf{x}^*) \rangle \leq 0$  from  $f_i(\mathbf{x}^*) \leq 0, \forall i \in [M]$ . Hence, combining with (3.7), we have the inequality in (3.8).  $\square$

## C Proof of Lemma 4

**Proof** For  $\hat{\mathbf{u}}^k$  given in (2.4), we have  $\|\hat{\mathbf{u}}^k\| \leq \theta$  and thus each coordinate of  $\hat{\mathbf{u}}^k$  is also less than  $\theta$ , i.e.  $-\theta \mathbf{1} \leq \hat{\mathbf{u}}^k \leq \theta \mathbf{1}$ . Recursively rewriting the updates in (2.3), (2.5) and (2.6) gives

$$\mathbf{m}^k = \beta_1 \mathbf{m}^{k-1} + (1 - \beta_1) \mathbf{u}^k = (1 - \beta_1) \sum_{j=1}^k \beta_1^{k-j} \mathbf{u}^j, \quad (\text{C.1})$$

$$\mathbf{v}^k = \beta_2 \mathbf{v}^{k-1} + (1 - \beta_2) (\hat{\mathbf{u}}^k)^2 = (1 - \beta_2) \sum_{j=1}^k \beta_2^{k-j} (\hat{\mathbf{u}}^j)^2, \quad (\text{C.2})$$

$$\hat{\mathbf{v}}^k = \max\{\hat{\mathbf{v}}^{k-1}, \mathbf{v}^k\} = \max_{j \in [k]} \mathbf{v}^j, \quad (\text{C.3})$$

here,  $\mathbf{m}^0 = \mathbf{v}^0 = \hat{\mathbf{v}}^0 = \mathbf{0}$ . By (C.2) and  $-\theta \mathbf{1} \leq \hat{\mathbf{u}}^k \leq \theta \mathbf{1}$ , we have  $\mathbf{v}^k \leq \theta^2(1 - \beta_2) \sum_{j=1}^k \beta_2^{k-j} \mathbf{1} \leq \theta^2 \mathbf{1}$ . By (C.3), we further have  $\hat{\mathbf{v}}^k \leq \theta^2 \mathbf{1}$ . Thus  $\mathbb{E}[\|\hat{\mathbf{v}}^k\|_1] \leq n\theta$  holds.

Notice  $\mathbb{E}[\|\mathbf{m}^k\|_{(\hat{\mathbf{v}}^k)^{-1/2}}^2] = \mathbb{E}[\|\frac{\mathbf{m}^k}{(\hat{\mathbf{v}}^k)^{1/4}}\|^2]$ . We can lower bound  $\mathbf{v}^k$  by keeping only the last term in (C.2) since  $(\hat{\mathbf{u}}^j)^2 \geq \mathbf{0}$ , i.e.  $\mathbf{v}^k \geq (1 - \beta_2) (\hat{\mathbf{u}}^k)^2$ . By (C.3), we also have  $\hat{\mathbf{v}}^k \geq (1 - \beta_2) \max_{j \in [k]} (\hat{\mathbf{u}}^j)^2$ . Plugging the inequality and (C.1) into  $\mathbb{E}[\|\mathbf{m}^k\|_{(\hat{\mathbf{v}}^k)^{-1/2}}^2]$  gives

$$\begin{aligned} \mathbb{E}[\|\mathbf{m}^k\|_{(\hat{\mathbf{v}}^k)^{-1/2}}^2] &\leq \mathbb{E} \left[ \left\| \frac{\mathbf{m}^k}{((1 - \beta_2) \max_{j \in [k]} (\hat{\mathbf{u}}^j)^2)^{1/4}} \right\|^2 \right] \\ &= \frac{(1 - \beta_1)^2}{(1 - \beta_2)^{1/2}} \mathbb{E} \left[ \left\| \frac{\sum_{j=1}^k \beta_1^{k-j} \mathbf{u}^j}{\max_{j \in [k]} |\hat{\mathbf{u}}^j|^{1/2}} \right\|^2 \right]. \end{aligned} \quad (\text{C.4})$$

Then we bound  $\left\| \frac{\sum_{j=1}^k \beta_1^{k-j} \mathbf{u}^j}{\max_{j \in [k]} |\hat{\mathbf{u}}^j|^{1/2}} \right\|^2$  by the Cauchy-Schwarz inequality.

$$\left\| \frac{\sum_{j=1}^k \beta_1^{k-j} \mathbf{u}^j}{\max_{j \in [k]} |\hat{\mathbf{u}}^j|^{1/2}} \right\|^2 = \left\| \sum_{j=1}^k (\beta_1^{k-j})^{1/2} \frac{(\beta_1^{k-j})^{1/2} \mathbf{u}^j}{\max_{j \in [k]} |\hat{\mathbf{u}}^j|^{1/2}} \right\|^2$$

$$\begin{aligned}
 &\leq \left( \sum_{j=1}^k \left( (\beta_1^{k-j})^{1/2} \right)^2 \right) \left( \sum_{j=1}^k \left\| \frac{(\beta_1^{k-j})^{1/2} \mathbf{u}^j}{\max_{j \in [k]} |\hat{\mathbf{u}}^j|^{1/2}} \right\|^2 \right) \\
 &= \left( \sum_{j=1}^k \beta_1^{k-j} \right) \left( \sum_{j=1}^k \beta_1^{k-j} \left\| \frac{\mathbf{u}^j}{\max_{j \in [k]} |\hat{\mathbf{u}}^j|^{1/2}} \right\|^2 \right) \\
 &\leq \left( \sum_{j=1}^k \beta_1^{k-j} \right) \left( \sum_{j=1}^k \beta_1^{k-j} \left\| \frac{\mathbf{u}^j}{|\hat{\mathbf{u}}^j|^{1/2}} \right\|^2 \right), \tag{C.5}
 \end{aligned}$$

where we use  $\max_{j \in [k]} |\hat{\mathbf{u}}^j|^{1/2} \geq |\hat{\mathbf{u}}^j|^{1/2}$  for  $j \in [k]$  in the second inequality. For  $\left\| \frac{\mathbf{u}^j}{|\hat{\mathbf{u}}^j|^{1/2}} \right\|^2$ , we have  $\hat{\mathbf{u}}^j$  given in (2.4) and notice  $\max \{1, \frac{\|\mathbf{u}^k\|}{\theta}\}$  is a scalar.

$$\begin{aligned}
 \left\| \frac{\mathbf{u}^j}{|\hat{\mathbf{u}}^j|^{1/2}} \right\|^2 &= \left\| \frac{\mathbf{u}^j}{\left| \frac{\mathbf{u}^j}{\max \{1, \frac{\|\mathbf{u}^j\|}{\theta}\}} \right|^{1/2}} \right\|^2 = \max \left\{ 1, \frac{\|\mathbf{u}^j\|}{\theta} \right\} \|\mathbf{u}^j\|_1 \\
 &\leq \max \left\{ 1, \frac{\|\mathbf{u}^j\|}{\theta} \right\} \sqrt{n} \|\mathbf{u}^j\| = \begin{cases} \sqrt{n} \|\mathbf{u}^j\|, & \text{if } \|\mathbf{u}^j\| \leq \theta, \\ \frac{\sqrt{n} \|\mathbf{u}^j\|^2}{\theta}, & \text{if } \|\mathbf{u}^j\| > \theta. \end{cases}
 \end{aligned}$$

So we get  $\left\| \frac{\mathbf{u}^j}{|\hat{\mathbf{u}}^j|^{1/2}} \right\|^2 \leq \sqrt{n} \left( \theta + \frac{\|\mathbf{u}^j\|^2}{\theta} \right)$ . Plug the inequality back to (C.5), and then (C.5) back to (C.4).

$$\begin{aligned}
 &\mathbb{E} \left[ \|\mathbf{m}^k\|_{(\mathbf{v}^k)^{-1/2}}^2 \right] \\
 &\leq \frac{(1-\beta_1)^2}{(1-\beta_2)^{1/2}} \mathbb{E} \left[ \left( \sum_{j=1}^k \beta_1^{k-j} \right) \left( \sum_{j=1}^k \beta_1^{k-j} \sqrt{n} \left( \theta + \frac{\|\mathbf{u}^j\|^2}{\theta} \right) \right) \right] \\
 &= \frac{\sqrt{n}(1-\beta_1)^2}{(1-\beta_2)^{1/2}} \left( \sum_{j=1}^k \beta_1^{k-j} \right) \left( \sum_{j=1}^k \beta_1^{k-j} \left( \theta + \frac{\mathbb{E}[\|\mathbf{u}^j\|^2]}{\theta} \right) \right) \\
 &\leq \frac{\sqrt{n}(1-\beta_1)^2}{(1-\beta_2)^{1/2}} \left( \sum_{j=1}^k \beta_1^{k-j} \right)^2 \max_{j \in [k]} \left( \theta + \frac{\mathbb{E}[\|\mathbf{u}^j\|^2]}{\theta} \right) \\
 &\leq \frac{\sqrt{n} \left( \theta + \frac{\max_{j \in [k]} \mathbb{E}[\|\mathbf{u}^j\|^2]}{\theta} \right)}{(1-\beta_2)^{1/2}},
 \end{aligned}$$

where we have used  $\sum_{j=1}^k \beta_1^{k-j} \leq \sum_{j=1}^{\infty} \beta_1^j \leq \frac{1}{1-\beta_1}$  for  $\beta_1 \in (0, 1)$  in the last inequality. With (3.2), the proof is finished.  $\square$

## D Proof of Lemma 5

**Proof** From the projection (2.7) in the primal variable update, we have for  $k \in [K]$  and  $\forall \mathbf{x} \in X$ ,

$$\begin{aligned} 0 &\geq \left\langle \mathbf{x}^{k+1} - \mathbf{x}, \mathbf{x}^{k+1} - \left( \mathbf{x}^k - \alpha_k \mathbf{m}^k / (\widehat{\mathbf{v}}^k)^{1/2} \right) \right\rangle_{(\widehat{\mathbf{v}}^k)^{1/2}} \\ &= \left\langle \mathbf{x}^{k+1} - \mathbf{x}, (\widehat{\mathbf{v}}^k)^{1/2} (\mathbf{x}^{k+1} - \mathbf{x}^k) + \alpha_k \mathbf{m}^k \right\rangle. \end{aligned} \quad (\text{D.1})$$

The first term of the right side equals to

$$\begin{aligned} &\left\langle \mathbf{x}^{k+1} - \mathbf{x}, (\widehat{\mathbf{v}}^k)^{1/2} (\mathbf{x}^{k+1} - \mathbf{x}^k) \right\rangle \\ &= \frac{1}{2} \left( \|\mathbf{x}^{k+1} - \mathbf{x}\|_{(\widehat{\mathbf{v}}^k)^{1/2}}^2 - \|\mathbf{x}^k - \mathbf{x}\|_{(\widehat{\mathbf{v}}^k)^{1/2}}^2 + \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_{(\widehat{\mathbf{v}}^k)^{1/2}}^2 \right). \end{aligned} \quad (\text{D.2})$$

Recursively rewrite  $\langle \mathbf{x}^{k+1} - \mathbf{x}, \mathbf{m}^k \rangle$  with the update (2.3)

$$\begin{aligned} &\langle \mathbf{x}^{k+1} - \mathbf{x}, \mathbf{m}^k \rangle \quad (\text{D.3}) \\ &= \langle \mathbf{x}^{k+1} - \mathbf{x}^k, \mathbf{m}^k \rangle + (1 - \beta_1) \langle \mathbf{x}^k - \mathbf{x}, \mathbf{u}^k \rangle + \beta_1 \langle \mathbf{x}^k - \mathbf{x}, \mathbf{m}^{k-1} \rangle \\ &= \langle \mathbf{x}^{k+1} - \mathbf{x}^k, \mathbf{m}^k \rangle + (1 - \beta_1) \langle \mathbf{x}^k - \mathbf{x}, \mathbf{u}^k \rangle \\ &\quad + \beta_1 \left( \langle \mathbf{x}^k - \mathbf{x}^{k-1}, \mathbf{m}^{k-1} \rangle + (1 - \beta_1) \langle \mathbf{x}^{k-1} - \mathbf{x}, \mathbf{u}^{k-1} \rangle \right) \\ &\quad \dots \\ &\quad + \beta_1^{k-1} \left( \langle \mathbf{x}^2 - \mathbf{x}^1, \mathbf{m}^1 \rangle + (1 - \beta_1) \langle \mathbf{x}^1 - \mathbf{x}, \mathbf{u}^1 \rangle \right) + \beta_1^k \langle \mathbf{x}^1 - \mathbf{x}, \mathbf{m}^0 \rangle \\ &= \sum_{j=1}^k \beta_1^{k-j} \left( \langle \mathbf{x}^{j+1} - \mathbf{x}^j, \mathbf{m}^j \rangle + (1 - \beta_1) \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j \rangle \right), \end{aligned} \quad (\text{D.4})$$

where the second equation recursively applied the first equation and the last term  $\beta_1^k \langle \mathbf{x}^1 - \mathbf{x}, \mathbf{m}^0 \rangle$  vanishes because  $\mathbf{m}^0 = \mathbf{0}$ . Plugging Eqs. (D.2) and (D.4) into the inequality (D.1) gives

$$\begin{aligned} &\alpha_k \sum_{j=1}^k \beta_1^{k-j} \left( \langle \mathbf{x}^{j+1} - \mathbf{x}^j, \mathbf{m}^j \rangle + (1 - \beta_1) \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j \rangle \right) \\ &\leq \frac{1}{2} \left( -\|\mathbf{x}^{k+1} - \mathbf{x}\|_{(\widehat{\mathbf{v}}^k)^{1/2}}^2 + \|\mathbf{x}^k - \mathbf{x}\|_{(\widehat{\mathbf{v}}^k)^{1/2}}^2 - \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_{(\widehat{\mathbf{v}}^k)^{1/2}}^2 \right). \end{aligned} \quad (\text{D.5})$$

Sum the above inequality (D.5) for  $k = 1$  to  $t$ . About the left side, we have

$$\begin{aligned}
 \sum_{k=1}^t \alpha_k \sum_{j=1}^k \beta_1^{k-j} \left( \langle \mathbf{x}^{j+1} - \mathbf{x}^j, \mathbf{m}^j \rangle \right) &= \sum_{j=1}^t \left( \langle \mathbf{x}^{j+1} - \mathbf{x}^j, \mathbf{m}^j \rangle \right) \sum_{k=j}^t \alpha_k \beta_1^{k-j} \\
 &\geq \sum_{j=1}^t \left( -\frac{\|\mathbf{x}^{j+1} - \mathbf{x}^j\|_{(\hat{\mathbf{v}}^j)^{1/2}}^2}{2 \sum_{k=j}^t \alpha_k \beta_1^{k-j}} - \frac{\sum_{k=j}^t \alpha_k \beta_1^{k-j}}{2} \|\mathbf{m}^j\|_{(\hat{\mathbf{v}}^j)^{-1/2}}^2 \right) \sum_{k=j}^t \alpha_k \beta_1^{k-j} \\
 &\stackrel{(2.10)}{\geq} \sum_{j=1}^t \left( -\frac{\|\mathbf{x}^{j+1} - \mathbf{x}^j\|_{(\hat{\mathbf{v}}^j)^{1/2}}^2}{2} - \frac{\alpha_j^2 \|\mathbf{m}^j\|_{(\hat{\mathbf{v}}^j)^{-1/2}}^2}{2(1-\beta_1)^2} \right). \quad (\text{D.6})
 \end{aligned}$$

About the right side of the sum of the inequality (D.5), by  $(\hat{\mathbf{v}}^k)^{1/2} \geq (\hat{\mathbf{v}}^{k-1})^{1/2} \geq \mathbf{0}$ ,  $k \in [t]$  since the iteration (2.6), and Assumption 1, we have

$$\begin{aligned}
 \sum_{k=1}^t \frac{1}{2} \left( -\|\mathbf{x}^{k+1} - \mathbf{x}\|_{(\hat{\mathbf{v}}^k)^{1/2}}^2 + \|\mathbf{x}^k - \mathbf{x}\|_{(\hat{\mathbf{v}}^k)^{1/2}}^2 \right) \\
 = \frac{1}{2} \left( -\|\mathbf{x}^{t+1} - \mathbf{x}\|_{(\hat{\mathbf{v}}^t)^{1/2}}^2 + \sum_{k=2}^t \|\mathbf{x}^k - \mathbf{x}\|_{(\hat{\mathbf{v}}^k)^{1/2} - (\hat{\mathbf{v}}^{k-1})^{1/2}}^2 + \|\mathbf{x}^1 - \mathbf{x}\|_{(\hat{\mathbf{v}}^1)^{1/2}}^2 \right) \\
 \leq \frac{B^2}{2} \left( \sum_{k=2}^t \|(\hat{\mathbf{v}}^k)^{1/2} - (\hat{\mathbf{v}}^{k-1})^{1/2}\|_1 + \|(\hat{\mathbf{v}}^1)^{1/2}\|_1 \right) = \frac{B^2}{2} \|(\hat{\mathbf{v}}^t)^{1/2}\|_1. \quad (\text{D.7})
 \end{aligned}$$

Thus with inequalities (D.6) and (D.7), the sum of the inequality (D.5) for  $k = 1$  to  $t$  becomes

$$\begin{aligned}
 \sum_{k=1}^t \alpha_k \sum_{j=1}^k \beta_1^{k-j} (1-\beta_1) \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j \rangle + \sum_{j=1}^t \left( -\frac{\|\mathbf{x}^{j+1} - \mathbf{x}^j\|_{(\hat{\mathbf{v}}^j)^{1/2}}^2}{2} - \frac{\alpha_j^2 \|\mathbf{m}^j\|_{(\hat{\mathbf{v}}^j)^{-1/2}}^2}{2(1-\beta_1)^2} \right) \\
 \leq \frac{B^2}{2} \|(\hat{\mathbf{v}}^t)^{1/2}\|_1 - \frac{1}{2} \sum_{k=1}^t \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_{(\hat{\mathbf{v}}^k)^{1/2}}^2.
 \end{aligned}$$

Eliminating the term  $\|\mathbf{x}^{j+1} - \mathbf{x}^j\|_{(\hat{\mathbf{v}}^j)^{1/2}}$  on both sides and exchanging the order of sums in the first term give

$$(1-\beta_1) \sum_{j=1}^t \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j \rangle \sum_{k=j}^t \alpha_k \beta_1^{k-j} \leq \frac{B^2}{2} \|(\hat{\mathbf{v}}^t)^{1/2}\|_1 + \frac{\sum_{j=1}^t \alpha_j^2 \|\mathbf{m}^j\|_{(\hat{\mathbf{v}}^j)^{-1/2}}^2}{2(1-\beta_1)^2}.$$

Then take the expectation on the above inequality. With the bounds given in Lemma 4, we have

$$\begin{aligned}
 (1-\beta_1) \sum_{j=1}^t \mathbb{E} \left[ \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j \rangle \right] \sum_{k=j}^t \alpha_k \beta_1^{k-j} \\
 \leq \frac{n\theta B^2}{2} + \frac{\sum_{j=1}^t \alpha_j^2 \frac{\sqrt{n}(\theta + \frac{\hat{P}_Z^j}{\theta})}{(1-\beta_2)^{1/2}}}{2(1-\beta_1)^2}
 \end{aligned}$$

$$\leq \frac{n\theta B^2}{2} + \frac{\sqrt{n}(\theta + \frac{\hat{P}_z^t}{\theta}) \sum_{j=1}^t \alpha_j^2}{2(1-\beta_1)^2(1-\beta_2)^{1/2}},$$

where the last inequality holds because we notice  $\hat{P}_z^k$  defined in (3.2) is nondecreasing with respect to  $k$ .  $\square$

## E Proof of Lemma 6

**Proof** For the dual variable is projected to the positive region in the update (2.9), it follows that for any  $\mathbf{z} \geq 0$ ,  $j \in [K]$ ,

$$\langle \mathbf{z}^{j+1} - \mathbf{z}, \mathbf{z}^{j+1} - (\mathbf{z}^j + \rho_j \mathbf{w}^j) \rangle \leq 0.$$

It could be rewritten as

$$\begin{aligned} \langle \mathbf{z}^{j+1} - \mathbf{z}, \mathbf{z}^{j+1} - \mathbf{z}^j \rangle &\leq \langle \mathbf{z}^{j+1} - \mathbf{z}, \rho_j \mathbf{w}^j \rangle \\ &= \langle \mathbf{z}^{j+1} - \mathbf{z}^j, \rho_j \mathbf{w}^j \rangle + \langle \mathbf{z}^j - \mathbf{z}, \rho_j \mathbf{w}^j \rangle. \end{aligned} \quad (\text{E.1})$$

For each term of the above inequality (E.1), we have

$$\begin{aligned} \langle \mathbf{z}^{j+1} - \mathbf{z}, \mathbf{z}^{j+1} - \mathbf{z}^j \rangle &= \frac{1}{2} \left( \|\mathbf{z}^{j+1} - \mathbf{z}^j\|^2 + \|\mathbf{z}^{j+1} - \mathbf{z}\|^2 - \|\mathbf{z}^j - \mathbf{z}\|^2 \right), \\ \langle \mathbf{z}^{j+1} - \mathbf{z}^j, \mathbf{w}^j \rangle &\leq \frac{1}{2\rho_j} \|\mathbf{z}^{j+1} - \mathbf{z}^j\|^2 + \frac{\rho_j}{2} \|\mathbf{w}^j\|^2, \\ \langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j \rangle &= \langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j - \mathbf{f}(\mathbf{x}^j) \rangle + \langle \mathbf{z}^j - \mathbf{z}, \mathbf{f}(\mathbf{x}^j) \rangle. \end{aligned}$$

Plugging the above three terms into the inequality (E.1) and eliminating  $\|\mathbf{z}^{j+1} - \mathbf{z}^j\|^2$  give

$$\frac{1}{2\rho_j} \left( \|\mathbf{z}^{j+1} - \mathbf{z}\|^2 - \|\mathbf{z}^j - \mathbf{z}\|^2 \right) \leq \frac{\rho_j}{2} \|\mathbf{w}^j\|^2 + \langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j - \mathbf{f}(\mathbf{x}^j) \rangle + \langle \mathbf{z}^j - \mathbf{z}, \mathbf{f}(\mathbf{x}^j) \rangle.$$

Rearranging the above inequality gives the inequality (3.10).  $\square$

## F Proof of Lemma 7

**Proof** For any  $j \in [K]$ , we have

$$\langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j \rangle = \langle \mathbf{x}^j - \mathbf{x}, \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle + \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle. \quad (\text{F.1})$$

Here  $\tilde{\nabla}_{\mathbf{x}}\mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) = \mathbb{E}[\mathbf{u}^j \mid \mathcal{H}^j] \in \partial_{\mathbf{x}}\mathcal{L}(\mathbf{x}^j, \mathbf{z}^j)$  according to Assumption 2. By the convexity of  $f_i(\mathbf{x})$ ,  $i = 0, 1, \dots, M$ , we know  $\mathcal{L}(\mathbf{x}, \mathbf{z})$  is convex with respect to  $\mathbf{x}$  and

$$\langle \mathbf{x}^j - \mathbf{x}, \tilde{\nabla}_{\mathbf{x}}\mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle \geq \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) - \mathcal{L}(\mathbf{x}, \mathbf{z}^j) = f_0(\mathbf{x}^j) - f_0(\mathbf{x}) + \langle \mathbf{z}^j, \mathbf{f}(\mathbf{x}^j) \rangle - \langle \mathbf{z}^j, \mathbf{f}(\mathbf{x}) \rangle.$$

Plug the lower bound of  $\langle \mathbf{z}^j, \mathbf{f}(\mathbf{x}^j) \rangle$  given in Lemma 6 to the above inequality.

$$\begin{aligned} & \langle \mathbf{x}^j - \mathbf{x}, \tilde{\nabla}_{\mathbf{x}}\mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle \\ & \geq f_0(\mathbf{x}^j) - f_0(\mathbf{x}) - \langle \mathbf{z}^j, \mathbf{f}(\mathbf{x}) \rangle + \langle \mathbf{z}, \mathbf{f}(\mathbf{x}^j) \rangle + \frac{1}{2\rho_j} \left( \|\mathbf{z}^{j+1} - \mathbf{z}\|^2 - \|\mathbf{z}^j - \mathbf{z}\|^2 \right) \\ & \quad - \frac{\rho_j}{2} \|\mathbf{w}^j\|^2 - \langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j - \mathbf{f}(\mathbf{x}^j) \rangle. \end{aligned}$$

Summarizing (F.1) with weights  $\sum_{k=j}^t \alpha_k \beta_1^{k-j}$  for  $j \in [t]$ , and plugging the above inequality give

$$\begin{aligned} & \sum_{j=1}^t \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j \rangle \sum_{k=j}^t \alpha_k \beta_1^{k-j} \\ & \geq \sum_{j=1}^t \left( f_0(\mathbf{x}^j) - f_0(\mathbf{x}) - \langle \mathbf{z}^j, \mathbf{f}(\mathbf{x}) \rangle + \langle \mathbf{z}, \mathbf{f}(\mathbf{x}^j) \rangle + \frac{1}{2\rho_j} \left( \|\mathbf{z}^{j+1} - \mathbf{z}\|^2 - \|\mathbf{z}^j - \mathbf{z}\|^2 \right) \right. \\ & \quad \left. - \frac{\rho_j}{2} \|\mathbf{w}^j\|^2 - \langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j - \mathbf{f}(\mathbf{x}^j) \rangle + \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}}\mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle \right) \sum_{k=j}^t \alpha_k \beta_1^{k-j}. \end{aligned} \quad (\text{F.2})$$

Summation of the term about  $\|\mathbf{z}^{j+1} - \mathbf{z}\|^2 - \|\mathbf{z}^j - \mathbf{z}\|^2$  can be lower bounded:

$$\begin{aligned} & \sum_{j=1}^t \frac{1}{2\rho_j} \left( \|\mathbf{z}^{j+1} - \mathbf{z}\|^2 - \|\mathbf{z}^j - \mathbf{z}\|^2 \right) \sum_{k=j}^t \alpha_k \beta_1^{k-j} \\ & = -\frac{\sum_{k=1}^t \alpha_k \beta_1^{k-1}}{2\rho_1} \|\mathbf{z}^1 - \mathbf{z}\|^2 + \frac{\alpha_t}{2\rho_t} \|\mathbf{z}^{t+1} - \mathbf{z}\|^2 \\ & \quad + \sum_{j=2}^t \left( \frac{\sum_{k=j-1}^t \alpha_k \beta_1^{k-(j-1)}}{2\rho_{j-1}} - \frac{\sum_{k=j}^t \alpha_k \beta_1^{k-j}}{2\rho_j} \right) \|\mathbf{z}^j - \mathbf{z}\|^2 \\ & \stackrel{(2.11)}{\geq} -\frac{\sum_{k=1}^t \alpha_k \beta_1^{k-1}}{2\rho_1} \|\mathbf{z}^1 - \mathbf{z}\|^2 + \frac{\alpha_t}{2\rho_t} \|\mathbf{z}^{t+1} - \mathbf{z}\|^2 \\ & \stackrel{(2.10)}{\geq} -\frac{\alpha_1 \|\mathbf{z}^1 - \mathbf{z}\|^2}{2\rho_1(1 - \beta_1)} + \frac{\alpha_t}{2\rho_t} \|\mathbf{z}^{t+1} - \mathbf{z}\|^2. \end{aligned} \quad (\text{F.3})$$

Summation of  $\|\mathbf{w}^j\|^2$  can also be lower bounded

$$\begin{aligned} -\sum_{j=1}^t \frac{\rho_j}{2} \|\mathbf{w}^j\|^2 \sum_{k=j}^t \alpha_k \beta_1^{k-j} &\stackrel{(2.10)}{\geq} -\sum_{j=1}^t \frac{\rho_j \alpha_j \|\mathbf{w}^j\|^2}{2(1-\beta_1)} \\ &\stackrel{(2.12)}{\geq} -\frac{\rho_1}{2\alpha_1(1-\beta_1)^2} \sum_{j=1}^t \alpha_j^2 \|\mathbf{w}^j\|^2. \end{aligned} \quad (\text{F.4})$$

Plugging the above two inequalities (F.3) and (F.4) into the inequality (F.2) gives

$$\begin{aligned} &\sum_{j=1}^t \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j \rangle \sum_{k=j}^t \alpha_k \beta_1^{k-j} \\ &\geq \sum_{j=1}^t \left( f_0(\mathbf{x}^j) - f_0(\mathbf{x}) - \langle \mathbf{z}^j, \mathbf{f}(\mathbf{x}) \rangle + \langle \mathbf{z}, \mathbf{f}(\mathbf{x}^j) \rangle \right) \sum_{k=j}^t \alpha_k \beta_1^{k-j} \\ &\quad - \frac{\alpha_1 \|\mathbf{z}^1 - \mathbf{z}\|^2}{2\rho_1(1-\beta_1)} + \frac{\alpha_t \|\mathbf{z}^{t+1} - \mathbf{z}\|^2}{2\rho_t} - \frac{\rho_1 \sum_{j=1}^t \alpha_j^2 \|\mathbf{w}^j\|^2}{2\alpha_1(1-\beta_1)^2} \\ &\quad + \sum_{j=1}^t \left( -\langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j - \mathbf{f}(\mathbf{x}^j) \rangle + \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle \right) \sum_{k=j}^t \alpha_k \beta_1^{k-j}. \end{aligned}$$

Then taking the expectation on the above inequality and using Assumption 2 to bound  $\mathbb{E}[\|\mathbf{w}^j\|^2]$  give the result (3.11).  $\square$

## G Proof of Lemma 8

**Proof** If  $(\mathbf{x}, \mathbf{z})$  are deterministic, we can prove (3.13) by the conditional expectation and Assumption 2, i.e.

$$\begin{aligned} \mathbb{E} \left[ \langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j - \mathbf{f}(\mathbf{x}^j) \rangle \right] &= \mathbb{E} \left[ \mathbb{E} \left[ \langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j - \mathbf{f}(\mathbf{x}^j) \rangle \mid \mathcal{H}^k \right] \right] \\ &= \mathbb{E} \left[ \left\langle \mathbf{z}^j - \mathbf{z}, \mathbb{E}[\mathbf{w}^j - \mathbf{f}(\mathbf{x}^j) \mid \mathcal{H}^k] \right\rangle \right] = \mathbb{E} \left[ \langle \mathbf{z}^j - \mathbf{z}, \mathbf{0} \rangle \right] = 0, \\ \mathbb{E} \left[ \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle \right] &= \mathbb{E} \left[ \mathbb{E} \left[ \langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle \mid \mathcal{H}^k \right] \right] \\ &= \mathbb{E} \left[ \left\langle \mathbf{x}^j - \mathbf{x}, \mathbb{E}[\mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \mid \mathcal{H}^k] \right\rangle \right] = \mathbb{E} \left[ \langle \mathbf{x}^j - \mathbf{x}, \mathbf{0} \rangle \right] = 0. \end{aligned}$$

Then we prove the stochastic case through considering the left two terms of (3.12), separately, in a similar way. Let  $\tilde{\mathbf{z}}^1 = \mathbf{z}^1$ ,  $\tilde{\mathbf{z}}^{j+1} = \tilde{\mathbf{z}}^j - \gamma_j(\mathbf{w}^j - \mathbf{f}(\mathbf{x}^j))$ , then  $\mathbf{z}^j - \tilde{\mathbf{z}}^j$  is known given  $\mathcal{H}^k$  and we have  $\mathbb{E}[\langle \mathbf{z}^j - \tilde{\mathbf{z}}^j, \mathbf{w}^j - \mathbf{f}(\mathbf{x}^j) \rangle] = 0$  like the above deterministic case. Thus we have



$$\begin{aligned}
 & \sum_{j=1}^t \gamma_j \mathbb{E}[\langle \mathbf{z}^j - \mathbf{z}, \mathbf{w}^j - \mathbf{f}(\mathbf{x}^j) \rangle] \\
 &= \sum_{j=1}^t \gamma_j \mathbb{E}[\langle \tilde{\mathbf{z}}^j - \mathbf{z}, \mathbf{w}^j - \mathbf{f}(\mathbf{x}^j) \rangle] = \sum_{j=1}^t \mathbb{E}[\langle \tilde{\mathbf{z}}^j - \mathbf{z}, \tilde{\mathbf{z}}^j - \tilde{\mathbf{z}}^{j+1} \rangle] \\
 &= \sum_{j=1}^t \frac{1}{2} \mathbb{E}[\|\tilde{\mathbf{z}}^j - \mathbf{z}\|^2 + \|\tilde{\mathbf{z}}^j - \tilde{\mathbf{z}}^{j+1}\|^2 - \|\mathbf{z} - \tilde{\mathbf{z}}^{j+1}\|^2] \\
 &= \frac{1}{2} \left( \mathbb{E}[\|\tilde{\mathbf{z}}^1 - \mathbf{z}\|^2] + \sum_{j=1}^t \mathbb{E}[\|\tilde{\mathbf{z}}^j - \tilde{\mathbf{z}}^{j+1}\|^2] - \mathbb{E}[\|\mathbf{z} - \tilde{\mathbf{z}}^{t+1}\|^2] \right) \\
 &\leq \frac{1}{2} \left( \mathbb{E}[\|\mathbf{z}^1 - \mathbf{z}\|^2] + \sum_{j=1}^t \gamma_j^2 \mathbb{E}[\|\mathbf{w}^j - \mathbf{f}(\mathbf{x}^j)\|^2] \right) \\
 &\leq \frac{1}{2} \left( \mathbb{E}[\|\mathbf{z}^1 - \mathbf{z}\|^2] + \sum_{j=1}^t \gamma_j^2 \mathbb{E}[\|\mathbf{w}^j\|^2] \right) \leq \frac{1}{2} \left( \mathbb{E}[\|\mathbf{z}^1 - \mathbf{z}\|^2] + F^2 \sum_{j=1}^t \gamma_j^2 \right), \tag{G.1}
 \end{aligned}$$

where the first inequality holds because we drop the nonpositive term and  $\tilde{\mathbf{z}}^1 = \mathbf{z}^1$ ; the second inequality holds because for any random vector  $\mathbf{w}$ ,  $\mathbb{E}[\|\mathbf{w} - \mathbb{E}[\mathbf{w}]\|^2] \leq \mathbb{E}[\|\mathbf{w}\|^2]$ , and here  $\mathbb{E}[\mathbf{w}^j \mid \mathcal{H}^j] = \mathbf{f}(\mathbf{x}^j)$  for  $j \in [t]$ ; and the last inequality holds by Assumption 2.

For the summation of  $\gamma_j \mathbb{E}[\langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle]$ , let  $\tilde{\mathbf{x}}^1 = \mathbf{x}^1$ ,  $\tilde{\mathbf{x}}^{j+1} = \tilde{\mathbf{x}}^j + \gamma_j (\mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j))$ , then  $\mathbb{E}[\langle \mathbf{x}^j - \tilde{\mathbf{x}}^j, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle] = 0$  and

$$\begin{aligned}
 & - \sum_{j=1}^t \gamma_j \mathbb{E}[\langle \mathbf{x}^j - \mathbf{x}, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle] \\
 &= - \sum_{j=1}^t \gamma_j \mathbb{E}[\langle \tilde{\mathbf{x}}^j - \mathbf{x}, \mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j) \rangle] = \sum_{j=1}^t \mathbb{E}[\langle \tilde{\mathbf{x}}^j - \mathbf{x}, \tilde{\mathbf{x}}^j - \tilde{\mathbf{x}}^{j+1} \rangle] \\
 &= \sum_{j=1}^t \frac{1}{2} \mathbb{E}[\|\tilde{\mathbf{x}}^j - \mathbf{x}\|^2 + \|\tilde{\mathbf{x}}^j - \tilde{\mathbf{x}}^{j+1}\|^2 - \|\mathbf{x} - \tilde{\mathbf{x}}^{j+1}\|^2] \\
 &= \frac{1}{2} \left( \mathbb{E}[\|\tilde{\mathbf{x}}^1 - \mathbf{x}\|^2] + \sum_{j=1}^t \mathbb{E}[\|\tilde{\mathbf{x}}^j - \tilde{\mathbf{x}}^{j+1}\|^2] - \mathbb{E}[\|\mathbf{x} - \tilde{\mathbf{x}}^{t+1}\|^2] \right) \\
 &\leq \frac{1}{2} \left( nB^2 + \sum_{j=1}^t \gamma_j^2 \mathbb{E}[\|\mathbf{u}^j - \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j)\|^2] \right) \leq \frac{1}{2} \left( nB^2 + \sum_{j=1}^t \gamma_j^2 \mathbb{E}[\|\mathbf{u}^j\|^2] \right) \\
 &\leq \frac{1}{2} \left( nB^2 + (\max_{j \in [t]} \mathbb{E}[\|\mathbf{u}^j\|^2]) \sum_{j=1}^t \gamma_j^2 \right) \leq \frac{1}{2} \left( nB^2 + \hat{P}_{\mathbf{z}}^t \sum_{j=1}^t \gamma_j^2 \right), \tag{G.2}
 \end{aligned}$$

where the first inequality holds because we drop the nonpositive term and (3.1); the second inequality holds because  $\mathbb{E}[\mathbf{u}^j \mid \mathcal{H}^j] = \tilde{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^j, \mathbf{z}^j)$  for  $j \in [t]$ ; and the last two inequalities holds by Assumption 2 and (3.2).

Adding (G.1) and (G.2) gives the result (3.12).  $\square$

## H Proof of Lemma 9

**Proof** Let  $\mathbf{x} = \mathbf{x}^*$  in (3.17), we have  $\mathbf{f}(\mathbf{x}^*) \leq \mathbf{0}$  and thus

$$\mathbb{E}[f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*) + \langle \mathbf{z}, \mathbf{f}(\bar{\mathbf{x}}) \rangle] \leq \epsilon_1 + \epsilon_0 \mathbb{E}[\|\mathbf{z}\|^2]. \quad (\text{H.1})$$

Since  $f_j(\bar{\mathbf{x}}) \leq [f_j(\bar{\mathbf{x}})]_+$  and  $\mathbf{z}^* \geq 0$ , we have from (3.7) that

$$-\sum_{j=1}^M z_j^* [f_j(\bar{\mathbf{x}})]_+ \leq f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*). \quad (\text{H.2})$$

Substituting (H.2) into (H.1) with  $\mathbf{z}$  given by  $z_j = 1 + z_j^*$  if  $f_j(\bar{\mathbf{x}}) > 0$  and  $z_j = 0$  otherwise for any  $j \in [M]$  gives

$$-\mathbb{E}\left[\sum_{j=1}^M z_j^* [f_j(\bar{\mathbf{x}})]_+\right] + \mathbb{E}\left[\sum_{j=1}^M (1 + z_j^*) [f_j(\bar{\mathbf{x}})]_+\right] \leq \epsilon_1 + \epsilon_0 \|1 + \mathbf{z}^*\|^2.$$

Simplifying the above inequality gives (3.19).

Letting  $z_j = 3z_j^*$  if  $f_j(\bar{\mathbf{x}}) > 0$  and  $z_j = 0$  otherwise for any  $j \in [M]$  in (H.1) and adding (H.2) together gives

$$\mathbb{E}\left[\sum_{j=1}^M z_j^* [f_j(\bar{\mathbf{x}})]_+\right] \leq \frac{\epsilon_1}{2} + \frac{9\epsilon_0}{2} \|\mathbf{z}^*\|^2. \quad (\text{H.3})$$

Hence, by the above inequality and (H.2), we obtain

$$-\mathbb{E}[f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*)] \leq \frac{\epsilon_1}{2} + \frac{9\epsilon_0}{2} \|\mathbf{z}^*\|^2. \quad (\text{H.4})$$

Thus  $\mathbb{E}[f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*)]_- \leq \frac{\epsilon_1}{2} + \frac{9\epsilon_0}{2} \|\mathbf{z}^*\|^2$ . In addition, from (H.1) with  $\mathbf{z} = 0$ , it follows  $\mathbb{E}[f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*)] \leq \epsilon_1$ . Since  $|a|_- = a + 2[a]_-$  for any real number  $a$ , we have

$$\begin{aligned} \mathbb{E}[|f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*)|_-] &= \mathbb{E}[f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*)] + 2\mathbb{E}[f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*)]_- \\ &\leq 2\epsilon_1 + 9\epsilon_0 \|\mathbf{z}^*\|^2, \end{aligned}$$

which gives (3.18).

Furthermore, let  $\mathbf{z} = \mathbf{0}$  in (3.17) and take  $\widehat{\mathbf{x}} \in \operatorname{argmin}_{\mathbf{x} \in X} f_0(\mathbf{x}) + \langle \bar{\mathbf{z}}, \mathbf{f}(\mathbf{x}) \rangle$ . By equation (2.2), we have  $\mathbb{E}f_0(\bar{\mathbf{x}}) \leq \mathbb{E}d(\bar{\mathbf{z}}) + \epsilon_1$ , which together with (3.6) gives

$$\mathbb{E}[d(\mathbf{z}^*) - d(\bar{\mathbf{z}})] \leq \mathbb{E}[f_0(\mathbf{x}^*) - f_0(\bar{\mathbf{x}})] + \epsilon_1. \quad (\text{H.5})$$

Combining the above inequality with (H.4) gives (3.20).  $\square$

## I Proof of Corollary 4

**Proof** For the step sizes  $\{\alpha_j\}_{j=1}^K$ , we have

$$\begin{aligned} \sum_{j=1}^K \alpha_j &= \alpha \sum_{j=1}^K \frac{1}{\sqrt{j+1}} \geq \alpha \int_{s=2}^{K+2} \frac{1}{\sqrt{s}} ds = 2\alpha(\sqrt{K+2} - \sqrt{2}), \\ \sum_{j=1}^K \alpha_j^2 &= \alpha^2 \sum_{j=1}^K \frac{1}{j+1} \leq \alpha^2 \int_{s=1}^{K+1} \frac{1}{s} ds = \alpha^2 \log(K+1). \end{aligned}$$

Similarly, for  $\{\rho_j\}_{j=1}^K$ , it holds  $\sum_{j=1}^K \rho_j^2 \leq \rho^2 \log(K+1)$ . Plug these bounds to the result of Theorem 3 and note  $\log(K+1) \geq 1$  for  $K \geq 2$ . We finish the proof.  $\square$

## References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308–318 (2016)
2. Aybat, N.S., Iyengar, G.: An augmented Lagrangian method for conic convex programming. arXiv preprint [arXiv:1302.6322](https://arxiv.org/abs/1302.6322) (2013)
3. Calafiore, G., Campi, M.C.: Uncertain convex programs: randomized solutions and confidence levels. Math. Progr. **102**(1), 25–46 (2005)
4. Calafiore, G.C., Campi, M.C.: The scenario approach to robust control design. IEEE Trans. Autom. Control **51**(5), 742–753 (2006)
5. 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)
6. Dozat, T.: Incorporating Nesterov momentum into adam. Dostupné z: [http://cs229.stanford.edu/proj2015/054\\_report.pdf](http://cs229.stanford.edu/proj2015/054_report.pdf) (2016)
7. Dua, D., Graff, C.: UCI machine learning repository (2017)
8. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**(Jul), 2121–2159 (2011)
9. Grant, M., Boyd, S.: CVX: Matlab software for disciplined convex programming, version 2.1. <http://cvxr.com/cvx> (2014)
10. Grant, M.C., Boyd, S.P.: Graph implementations for nonsmooth convex programs. In: Recent Advances in Learning and Control, pp. 95–110. Springer (2008)
11. Gupta, M., Cotter, A., Pfeifer, J., Voevodski, K., Canini, K., Mangylov, A., Moczydlowski, W., Van Esbroeck, A.: Monotonic calibrated interpolated look-up tables. J. Mach. Learn. Res. **17**(1), 3790–3836 (2016)
12. Guyon, I., Gunn, S., Ben-Hur, A., Dror, G.: Result analysis of the nips 2003 feature selection challenge. Adv. Neural Inf. Process. Syst. **17**, 545–552 (2004)

13. Hamedani, E.Y., Jalilzadeh, A., Aybat, N.S., Shanbhag, U.V.: Iteration complexity of randomized primal-dual methods for convex-concave saddle point problems. arXiv preprint [arXiv:1806.04118](https://arxiv.org/abs/1806.04118) (2018)
14. Hamedani, E.Y., Aybat, N.S.: A primal-dual algorithm for general convex-concave saddle point problems. arXiv preprint [arXiv:1803.01401](https://arxiv.org/abs/1803.01401) (2018)
15. Hien, L.T.K., Zhao, R., Haskell, W.B.: An inexact primal-dual smoothing framework for large-scale non-bilinear saddle point problems. arXiv preprint [arXiv:1711.03669](https://arxiv.org/abs/1711.03669) (2017)
16. Juditsky, A., Nemirovski, A., Tauvel, C.: Solving variational inequalities with stochastic mirror-prox algorithm. *Stoch. Syst.* **1**(1), 17–58 (2011)
17. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
18. Lan, G., Zhou, Z.: Algorithms for stochastic optimization with function or expectation constraints. *Comput. Optim. Appl.* **76**, 1–38 (2020)
19. Lin, Q., Ma, R., Yang, T.: Level-set methods for finite-sum constrained convex optimization. In: International Conference on Machine Learning, pp. 3112–3121 (2018)
20. Lin, Q., Nadarajah, S., Soheili, N.: A level-set method for convex optimization with a feasible solution path. *SIAM J. Optim.* **28**(4), 3290–3311 (2018)
21. Lu, Z., Zhou, Z.: Iteration-complexity of first-order augmented Lagrangian methods for convex conic programming. arXiv preprint [arXiv:1803.09941](https://arxiv.org/abs/1803.09941) (2018)
22. Luedtke, J., Ahmed, S.: A sample approximation approach for optimization with probabilistic constraints. *SIAM J. Optim.* **19**(2), 674–699 (2008)
23. Luo, L., Xiong, Y., Liu, Y., Sun, X.: Adaptive gradient methods with dynamic bound of learning rate. arXiv preprint [arXiv:1902.09843](https://arxiv.org/abs/1902.09843) (2019)
24. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19**(4), 1574–1609 (2009)
25. Neumann, J.V.: Zur theorie der gesellschaftsspiele. *Math. Ann.* **100**(1), 295–320 (1928)
26. Pagnoncelli, B.K., Ahmed, S., Shapiro, A.: Sample average approximation method for chance constrained programming: theory and applications. *J. Optim. Theory Appl.* **142**(2), 399–416 (2009)
27. Rigollet, P., Tong, X.: Neyman–Pearson classification, convexity and stochastic constraints. *J. Mach. Learn. Res.* **12**(Oct), 2831–2855 (2011)
28. Robbins, H., Monro, S.: A stochastic approximation method. *Ann. Math. Stat.* **22**, 400–407 (1951)
29. Rockafellar, R.T., Uryasev, S., et al.: Optimization of conditional value-at-risk. *J. Risk* **2**, 21–42 (2000)
30. Ryu, E.K., Yin, W.: Proximal-proximal-gradient method. arXiv preprint [arXiv:1708.06908](https://arxiv.org/abs/1708.06908) (2017)
31. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of adam and beyond. In: International Conference on Learning Representations (2018)
32. Scott, C., Nowak, R.: A Neyman–Pearson approach to statistical learning. *IEEE Trans. Inf. Theory* **51**(11), 3806–3819 (2005)
33. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming: Modeling and Theory. SIAM (2014)
34. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. COURSERA Neural Netw. Mach. Learn. **4**(2), 26–31 (2012)
35. Wang, M., Bertsekas, D.P.: Stochastic first-order methods with random constraint projection. *SIAM J. Optim.* **26**(1), 681–717 (2016)
36. Wang, M., Chen, Y., Liu, J., Gu, Y.: Random multi-constraint projection: Stochastic gradient methods for convex optimization with many constraints. arXiv preprint [arXiv:1511.03760](https://arxiv.org/abs/1511.03760) (2015)
37. Wang, M., Fang, E.X., Liu, H.: Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Math. Progr.* **161**(1–2), 419–449 (2017)
38. Xu, Y.: Primal-dual stochastic gradient method for convex programs with many functional constraints. *SIAM J. Optim.* **30**(2), 1664–1692 (2020)
39. Xu, Y.: First-order methods for constrained convex programming based on linearized augmented Lagrangian function. *INFORMS J. Optim.* **3**(1), 89–117 (2021)
40. Xu, Y.: Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. *Math. Progr.* **185**, 199–244 (2021)
41. Xu, Y., Xu, Y.: Katyusha acceleration for convex finite-sum compositional optimization. *INFORMS J. Optim.* **3**(4), 418–443 (2021)

42. Yu, H., Neely, M.J.: A primal-dual type algorithm with the  $O(1/t)$  convergence rate for large scale constrained convex programs. In: 2016 IEEE 55th Conference on Decision and Control (CDC), pp. 1900–1905. IEEE (2016)
43. Zafar, M.B., Valera, I., Gomez-Rodriguez, M., Gummadi, K.P.: Fairness constraints: a flexible approach for fair classification. *J. Mach. Learn. Res.* **20**(75), 1–42 (2019)
44. Zafar, M.B., Valera, I., Rognier, M.G., Gummadi, K.P.: Fairness constraints: mechanisms for fair classification. In: *Artificial Intelligence and Statistics*. PMLR, pp. 962–970 (2017)
45. Zeiler, M.D.: Adadelat: an adaptive learning rate method. arXiv preprint [arXiv:1212.5701](https://arxiv.org/abs/1212.5701) (2012)
46. Zhao, R.: Optimal stochastic algorithms for convex–concave saddle-point problems. arXiv preprint [arXiv:1903.01687](https://arxiv.org/abs/1903.01687) (2019)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.