PRIMAL-DUAL STOCHASTIC GRADIENT METHOD FOR CONVEX PROGRAMS WITH MANY FUNCTIONAL CONSTRAINTS[*]

YANGYANG XU[†]

Abstract. The stochastic gradient method (SGM) has been popularly applied to solve optimization problems with an objective that is stochastic or an average of many functions. Most existing works on SGMs assume that the underlying problem is unconstrained or has an easy-to-project constraint set. In this paper, we consider problems that have a stochastic objective and also many functional constraints. For such problems, it could be extremely expensive to project a point to the feasible set, or even compute subgradient and/or function value of all constraint functions. To find solutions to these problems, we propose a novel (adaptive) SGM based on the classical augmented Lagrangian function. Within every iteration, it inquires a stochastic subgradient of the objective, and a subgradient and the function value of one randomly sampled constraint function. Hence, the per-iteration complexity is low. We establish its convergence rate for convex problems and also problems with strongly convex objective. It can achieve the optimal $O(1/\sqrt{k})$ convergence rate for the convex case and nearly optimal $O\big((\log k)/k\big)$ rate for the strongly convex case. Numerical experiments on a sample approximation problem of the robust portfolio selection and quadratically constrained quadratic programming are conducted to demonstrate its efficiency.

1. Introduction. In this paper, we consider the constrained stochastic program

$$(1.1) \qquad \min_{\mathbf{x} \in X} f_0(\mathbf{x}) \equiv \mathbb{E}_\xi[F_0(\mathbf{x}; \xi)], \text{ s.t. } f_j(\mathbf{x}) \leq 0, j = 1, \ldots, M,$$

where X is a convex set in $\mathbb{R}^n$, ξ is a random variable, and f_j is a convex function for each $j = 0, 1, \ldots, M$. All nonlinear optimization problems in $\mathbb{R}^n$ can be formulated in the form of (1.1). We are particularly interested in the case that M is a large number.

To find a solution of (1.1), we aim at designing a novel primal-dual stochastic gradient method (SGM). We assume an oracle, which can return a stochastic approximation of a subgradient of f_0, and also the function value and a deterministic subgradient of each f_j at any inquired point $\mathbf{x} \in X$. Since M is big, it would be computationally very expensive if at every update, we inquire the objective value and/or subgradient of all f_j's. Based on this observation, our algorithm, at every iteration, will simply call the oracle to return subgradients and function values of a few sampled constraint functions.

The algorithm is derived based on the classical augmented Lagrangian function (cf. [19, 20]) of an equivalent rescaled variant of (1.1), i.e.,

$$\mathcal{L}_\beta(\mathbf{x}, \mathbf{z}) = f_0(\mathbf{x}) + \Psi_\beta(\mathbf{x}, \mathbf{z}).$$

Here, $\beta > 0$ is the penalty parameter, $\mathbf{z}$ is the Lagrangian multiplier or dual variable,

[†]Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180 (xuy21@rpi.edu).

$$(1.2) \qquad \Psi_\beta(\mathbf{x}, \mathbf{z}) = \frac{1}{M} \sum_{j=1}^M \psi_\beta\big(f_j(\mathbf{x}), z_j\big),$$

and

$$(1.3) \qquad \psi_\beta(u, v) = \begin{cases} uv + \frac{\beta}{2}u^2 & \text{if } \beta u + v \geq 0, \\ -\frac{v^2}{2\beta} & \text{if } \beta u + v < 0. \end{cases}$$

Note that Ψ_β is convex in $\mathbf{x}$ and concave in $\mathbf{z}$. Given $\beta > 0$, the augmented dual function is defined as

$$(1.4) \qquad d_\beta(\mathbf{z}) = \min_{\mathbf{x} \in X} \mathcal{L}_\beta(\mathbf{x}, \mathbf{z}).$$

At each iteration k, we first sample one constraint function f_{j_k}. Second, we call the oracle to obtain a stochastic subgradient $\mathbf{g}_0^k$ of f_0, and a subgradient $\tilde{\nabla} f_{j_k}(\mathbf{x}^k)$ and the function value of f_{j_k} at $\mathbf{x}^k$. Let

$$(1.5) \qquad \mathbf{h}^k = \big[\beta f_{j_k}\big(\mathbf{x}^k\big) + z_{j_k}^k\big]_+ \tilde{\nabla} f_{j_k}\big(\mathbf{x}^k\big).$$

Then $\mathbf{g}_0^k + \mathbf{h}^k$ is a stochastic subgradient of $\mathcal{L}_\beta$ with respect to $\mathbf{x}$. Third, we perform a (weighted) projected stochastic subgradient update as in (1.6) to the primal variable $\mathbf{x}$, and finally we update dual variable z_{j_k}.

The pseudocode of the proposed method is shown in Algorithm 1, which iteratively performs a (adaptive) stochastic subgradient update to the primal variable $\mathbf{x}$ and a randomized coordinate update to the dual variable $\mathbf{z}$. In order to have an easy update, $\mathbf{D}_k$ will be set to a diagonal matrix for each k. We will consider two different settings of $\mathbf{D}_k$ in our analysis.

SETTING 1. $\mathbf{D}_k = \frac{\mathbf{I}}{\alpha_k}$, where $\alpha_k > 0$ for all k, and $\mathbf{I}$ is the identity matrix.

Algorithm 1: Primal-dual stochastic gradient method (PDSG) for (1.1)

1 **Initialization:** choose $\mathbf{x}^1 \in X$, $\mathbf{z}^1 = \mathbf{0}$, and $\beta > 0$;

2 **for** $k = 1, 2, \ldots$ **do**

3 $\quad$ Pick $j_k \in [M]$ uniformly at random;

4 $\quad$ Call the oracle to return a stochastic subgradient $\mathbf{g}_0^k$ of f_0 and
$\quad\quad$ subgradient and function value of f_{j_k} at $\mathbf{x}^k$;

5 $\quad$ Obtain $\mathbf{h}^k$ in (1.5), choose $\mathbf{D}_k \succ 0$, and update the primal variable $\mathbf{x}$ by

$$(1.6) \qquad \mathbf{x}^{k+1} = \underset{\mathbf{x} \in X}{\arg\min} \big\|\mathbf{x} - \big(\mathbf{x}^k - \mathbf{D}_k^{-1}(\mathbf{g}_0^k + \mathbf{h}^k)\big)\big\|_{\mathbf{D}_k};$$

$\quad$ Choose $0 < \rho_k \leq \beta$ and update the dual variable $\mathbf{z}$ by

$$(1.7) \qquad z_j^{k+1} = \begin{cases} z_j^k & \text{if } j \neq j_k, \\ z_j^k + \rho_k \cdot \max\left(-\frac{z_j^k}{\beta}, f_j(\mathbf{x}^k)\right) & \text{if } j = j_k \end{cases}$$

SETTING 2. $\mathbf{D}_k = \mathrm{diag}(\mathbf{s}^k) + \frac{\mathbf{I}}{\alpha_k}$, where $\alpha_k > 0$ and $\mathbf{s}^k = \eta\sqrt{\sum_{t=1}^{k}\frac{(\mathbf{g}_0^t + \mathbf{h}^t)^2}{\gamma_t^2}}$ with $\eta > 0$ and $\gamma_t = \max\left(1, \|\mathbf{g}_0^t + \mathbf{h}^t\|\right)$ for all t. Here, $\mathbf{a}^2$ and $\sqrt{\mathbf{a}}$ denote the componentwise square and square-root for a vector $\mathbf{a}$.

Note that in Setting 2, $\mathbf{D}_k$ is adaptive to the primal stochastic subgradient. We scale the subgradient for technical reasons, and it is inspired by [29]. With such a setting, Algorithm 1 is an adaptive primal-dual SGM, and it appears to be the first one under the primal-dual setting. Although the same order of convergence rate will be shown for both settings, the adaptive one can numerically perform significantly better.

We remark that if the potential application has any affine equality constraint $\mathbf{a}^\top\mathbf{x} = b$, we can always write it into two affine inequality constraints $\mathbf{a}^\top\mathbf{x} \leq b$ and $-\mathbf{a}^\top\mathbf{x} \leq -b$ and thus formulate the problem in the form of (1.1), or we can use a technique similar to that in [27] to handle the equality and inequality constraints simultaneously. Furthermore, instead of sampling one constraint function every time, we can sample a small set J_k of constraint functions and let

$$\mathbf{h}^k = \frac{1}{|J_k|}\sum_{j \in J_k}\left[\beta f_j(\mathbf{x}^k) + z_j^k\right]_+\tilde{\nabla}f_j(\mathbf{x}^k)$$

in the update (1.6) and also update z_j $j \in J_k$. All our convergence results can still be obtained.

1.1. Motivating examples. We give a few examples that can be written in the form of (1.1) with a very big M, and our proposed algorithm can be applied.

Stochastic linear programming. A two-stage stochastic linear program (cf. [22, section 2.1]) can be formulated as

$$(1.8) \qquad \min_{\mathbf{x}}\mathbf{c}^\top\mathbf{x} + \mathbb{E}\left[f_\xi(\mathbf{x})\right], \text{ s.t. } \mathbf{A}\mathbf{x} \leq \mathbf{b},$$

where $\xi = (\mathbf{B}, \mathbf{C}, \mathbf{d}, \mathbf{q})$ and $f_\xi(\mathbf{x})$ are respectively the data and the optimal value of the second stage linear program

$$\min_{\mathbf{y}}\mathbf{q}^\top\mathbf{y}, \text{ s.t. } \mathbf{B}\mathbf{x} + \mathbf{C}\mathbf{y} \leq \mathbf{d}.$$

As there are M scenarios in the second stage, i.e., $\xi \in \{\xi_1, \dots, \xi_M\}$ with $\mathrm{Prob}(\xi = \xi_i) = p_i > 0$ and $\sum_{i=1}^{M}p_i = 1$, then

$$\mathbb{E}\left[f_\xi(\mathbf{x})\right] = \sum_{i=1}^{M}p_i f_{\xi_i}(\mathbf{x}) = \sum_{i=1}^{M}p_i\min\left\{\mathbf{q}_i^\top\mathbf{y} : \mathbf{B}_i\mathbf{x} + \mathbf{C}_i\mathbf{y} \leq \mathbf{d}_i\right\}.$$

Hence, (1.8) can be written as a single large-scale linear program:

$$(1.9) \quad \min_{\mathbf{x}, \mathbf{y}_1, \cdots, \mathbf{y}_M}\mathbf{c}^\top\mathbf{x} + \sum_{i=1}^{M}p_i\mathbf{q}_i^\top\mathbf{y}_i, \text{ s.t. } \mathbf{A}\mathbf{x} \leq \mathbf{b}, \ \mathbf{B}_i\mathbf{x} + \mathbf{C}_i\mathbf{y}_i \leq \mathbf{d}_i, \ i = 1, \dots, M.$$

Clearly, (1.9) is in the form of (1.1), and if there are many scenarios, i.e., M is big, it could be extremely expensive to access all the data at every update to the variables.

Chance constrained problems by sampling and discarding. A nonlinear program with chance constraint is formulated as

$$(1.10) \qquad \min_{\mathbf{x} \in X} f_0(\mathbf{x}), \text{ s.t. } \text{Prob}\big(g(\mathbf{x};\xi) \leq 0\big) \geq 1 - \tau,$$

where $X \subseteq \mathbb{R}^n$ is a convex set, ξ is an uncertain parameter on a support set Ξ, and τ is a user-specified risk level of constraint violation. Even though $g(\cdot\,;\xi)$ is convex for any $\xi \in \Xi$, the chance constraint set in (1.10) may not be convex. Hence, exactly solving (1.10) is hard in general. To numerically solve (1.10), the work [4] introduces a sample-based approximation method, called the *sampling and discarding* approach. This method makes N independent samples of ξ, then eliminates p of them, and solves a deterministic problem with the remaining $M = N - p$ constraints, i.e.,

$$(1.11) \qquad \min_{\mathbf{x} \in X} f_0(\mathbf{x}), \text{ s.t. } g(\mathbf{x};\xi_i) \leq 0 \ \forall i = 1, \ldots, M,$$

where $\{\xi_1, \ldots, \xi_M\}$ contains the M samples after discarding. It is shown in [4] that under certain assumptions, for any $\varepsilon \in (0,1)$, if

$$(1.12) \qquad \left(\begin{array}{c} p + n - 1 \\ p \end{array} \right) \sum_{i=0}^{p+n-1} \left(\begin{array}{c} N \\ i \end{array} \right) \tau^i (1-\tau)^{N-i} \leq \varepsilon,$$

the solution of (1.11) is feasible for (1.10) with probability at least $1 - \varepsilon$.

Note that if no *discarding* is performed, the above method is similar to the scenario approximation approaches in [10, 14]. For high-dimensional problems, i.e., n is big, it is required to set a significantly bigger N and also $N - p$ to have (1.12). Therefore, the sample-based approximation problem (1.11) will have many functional constraints and be in the form of (1.1).

Robust optimization by sampling. Different from the chance constrained problem (1.10), robust optimization requires the constraint $g(\mathbf{x};\xi) \leq 0$ to be satisfied for any $\xi \in \Xi$, i.e.,

$$(1.13) \qquad \min_{\mathbf{x} \in X} f_0(\mathbf{x}), \text{ s.t. } g(\mathbf{x};\xi) \leq 0 \quad \forall \xi \in \Xi.$$

Similar to the scenario approximation method for chance constrained problems, the sampling approach (e.g., [3]) has also been proposed to numerically solve (1.13). Let $\{\xi_1, \ldots, \xi_M\}$ be M independently extracted samples. It is shown in [3] that for any $\tau \in (0,1)$ and any $\varepsilon \in (0,1)$, if the number of samples satisfies $M \geq \frac{n}{\tau \varepsilon} - 1$, then the solution to (1.11) will be a τ-level robustly feasible solution with probability at least $1 - \varepsilon$. If n is big, and high feasibility level and high probability are required, then M would be a very big number, and thus (1.11) has an extremely big number of functional constraints.

1.2. Existing methods. In this subsection, we review a few existing methods that could potentially be applied to solve (1.1) and show how our method relates to them. Some of these methods are primal-dual type like our method, and others are purely primal methods.

Stochastic mirror-prox method. The proposed method is closely related to the stochastic mirror-prox method [1, 7] for saddle-point problems or more generally for variational inequality (VI) problems. By the augmented Lagrangian function, one can equivalently formulate (1.1) into the following saddle-point problem (cf. [18]):

$$(1.14) \qquad \min_{\mathbf{x} \in X} \max_{\mathbf{z}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{z}).$$

Assuming $\nabla \mathcal{L}_\beta$ to be Lipschitz continuous and $\mathbf{z}$ in a compact set Z, then we can apply the method in [1] to the above saddle-point problem and have the update[1]

$$(1.15a) \qquad \left(\hat{\mathbf{x}}^k, \hat{\mathbf{z}}^k \right) = \mathrm{Proj}_{X \times Z} \left(\left(\mathbf{x}^k - \alpha_k \mathbf{g}_x^k, \mathbf{z}^k + \alpha_k \mathbf{g}_z^k \right) \right),$$

$$(1.15b) \qquad \left(\mathbf{x}^{k+1}, \mathbf{z}^{k+1} \right) = \mathrm{Proj}_{X \times Z} \left(\left(\mathbf{x}^k - \alpha_k \hat{\mathbf{g}}_x^k, \mathbf{z}^k + \alpha_k \hat{\mathbf{g}}_z^k \right) \right),$$

where $(\mathbf{g}_x^k, \mathbf{g}_z^k)$ and $(\hat{\mathbf{g}}_x^k, \hat{\mathbf{g}}_z^k)$ are stochastic approximation of $\nabla \mathcal{L}_\beta$ at $(\mathbf{x}^k, \mathbf{z}^k)$ and $(\hat{\mathbf{x}}^k, \hat{\mathbf{z}}^k)$. The above update performs two stochastic gradient (SG) projections. To have convergence, it seems to be required for VI problems. However, for saddle-point problems, [13] shows that one SG projection is sufficient for convergence guarantee, namely, simply set $(\hat{\mathbf{x}}^k, \hat{\mathbf{z}}^k) = (\mathbf{x}^k, \mathbf{z}^k)$ and then obtain $(\mathbf{x}^{k+1}, \mathbf{z}^{k+1})$ by (1.15b).

The methods in [1] and [13] both require the dual variable to be in a compact set for convergence guarantee. Generally, it is difficult to estimate a valid bound on the dual variable, especially for a stochastic program. In addition, at each iteration, they use the same step size for both primal and dual variable updates, which seems to be required in their analysis. On the contrary, we will not assume boundedness of $\mathbf{z}$ but instead we can prove the boundedness of the sequence $\{\mathbf{z}^k\}$ in expectation. Furthermore, we allow use of different step sizes, and this is crucial for the convergence analysis of our adaptive method.

The SGM for saddle-point problems is also studied in [15]. However, it requires strong convexity for both primal and dual variables. For bilinear convex-concave saddle-point problems, the authors of [5] give an optimal primal-dual SGM. Without assuming boundedness on either primal or dual variables, they show an $O(1/\sqrt{k})$ convergence rate in terms of a perturbed primal-dual gap; cf. [5, Corollary 3.4]. Applying their method, i.e., [5, Algorithm 3], to an affinely constrained convex problem, one can show that if the primal variable and the output dual iterate are bounded, then the convergence rate is $O(1/\sqrt{k})$ in terms of both primal-dual objective gap and feasibility violation.

Cooperative stochastic approximation. The problem (1.1) can also be equivalently formulated as a stochastic program with a single finite-sum constraint:

$$(1.16) \qquad \min_{\mathbf{x} \in X} f_0(\mathbf{x}), \text{ s.t. } \frac{1}{M} \sum_{j=1}^{M} [f_j(\mathbf{x})]_+ \leq 0,$$

and we can apply the cooperative stochastic approximation (CSA) method in [8] to find an approximate solution. At each iteration k, CSA first samples one constraint function f_{j_k} and checks its value at the iterate $\mathbf{x}^k$. If $f_{j_k}(\mathbf{x}^k) \geq \eta_k$, set $\mathbf{g}^k = \tilde{\nabla} f_{j_k}(\mathbf{x}^k)$, and otherwise, set $\mathbf{g}^k$ to an unbiased estimate of $\tilde{\nabla} f_0(\mathbf{x}^k)$, where $\eta_k > 0$ is a parameter to control constraint violation. Then it updates the iterate by

$$(1.17) \qquad \mathbf{x}^{k+1} = \mathrm{Proj}_X \left(\mathbf{x}^k - \alpha_k \mathbf{g}^k \right),$$

where α_k is a step size.

For convex problems, CSA is shown to enjoy an $O(1/\sqrt{k})$ convergence rate in terms of both objective and feasibility. The order can be improved to $O(1/k)$ if both the objective and constraint functions in (1.16) are strongly convex. We will show that the proposed algorithm can enjoy the same order of convergence rate for convex

---

[1]Here, we use the Euclidean norm square as the proximal term, while [1] actually uses a more general Bregman distance function.

problems. To have an improved rate of $O\big((\log k)/k\big)$, we need strong convexity of the objective function but only convexity on the constraint functions. However, we need an additional assumption on the existence of a primal-dual solution. Hence, our method can have a better convergence rate than CSA for solving a problem with a strongly convex objective but only convex constraint functions, such as finding the projection onto the intersection of many polyhedral sets [16, 23].

Stochastic subgradient with random constraint projection. Let $X_0 = X$ and

$$(1.18) \qquad X_j = \{\mathbf{x} \in \mathbb{R}^n : f_j(\mathbf{x}) \leq 0\},\ j = 1, \ldots, M.$$

Then (1.1) can be written as

$$(1.19) \qquad \min_{\mathbf{x}} f_0(\mathbf{x}),\ \text{s.t.}\ \mathbf{x} \in \mathcal{X} = \cap_{j=0}^M X_j.$$

On solving the above problem, we can apply the method in [24, 25] and iteratively perform the update:

$$(1.20) \qquad \mathbf{x}^{k+1} = \mathrm{Proj}_{X_{j_k}} \left(\mathbf{x}^k - \alpha_k \mathbf{g}_0^k \right),$$

where j_k is randomly chosen from $\{0, 1, \ldots, M\}$, Proj_{X_j} denotes the projection onto X_j, and $\mathbf{g}_0^k$ is a stochastic approximation of a subgradient of f_0 at $\mathbf{x}^k$. Various sampling schemes on j_k are studied in [24]. Under the linear regularity assumption on the set collection $\{X_j\}_{j=0}^M$, a sublinear convergence result is established. If f_0 is convex, the rate is $O(1/\sqrt{k})$ in terms of objective error $|f_0(\mathbf{x}^k) - f_0^*|$ and $O\big((\log k)/k\big)$ in terms of constraint violation $[\mathrm{dist}(\mathbf{x}^k, \mathcal{X})]^2$. In [25], the rate of constraint violation is improved to $O(1/k)$. Furthermore, if f_0 is strongly convex, [25] shows the convergence rate $O\big((\log k)/k\big)$ in terms of objective error and $O(1/k^2)$ of constraint violation. To have efficient computation in the update (1.20), X_j is required to be a simple set for each $j = 0, 1, \ldots, M$. Hence, if Proj_{X_j}'s are difficult to evaluate, such as the logistic loss function induced constraint set in the Neyman–Pearson classification problem [17], the method in [24, 25] will be inefficient. By contrast, our update in (1.6) can be computed efficiently as long as X is simple.

Stochastic proximal-proximal gradient method. Let $r(\mathbf{x}) = \iota_X(\mathbf{x})$ and $g_j(\mathbf{x}) = \iota_{X_j}(\mathbf{x})$, where ι_X denotes the indicator function on X, and X_j's are defined in (1.18). Then (1.1) is equivalent to

$$(1.21) \qquad \min_{\mathbf{x}} r(\mathbf{x}) + \frac{1}{M} \sum_{j=1}^M \left(f_0(\mathbf{x}) + g_j(\mathbf{x}) \right).$$

When f_0 is differentiable, the stochastic proximal-proximal gradient (S-PPG) method [21] can be applied to find a solution of (1.21). It starts from $(\mathbf{x}^0, \mathbf{z}_1^0, \ldots, \mathbf{z}_M^0)$ and iteratively performs the update:

$$(1.22) \qquad \begin{aligned} \mathbf{x}^{k+\frac{1}{2}} &= \mathrm{Proj}_X \left(\frac{1}{M} \sum_{j=1}^M \mathbf{z}_j^k \right), \\ \mathbf{x}^{k+1} &= \mathrm{Proj}_{X_{j_k}} \left(2\mathbf{x}^{k+\frac{1}{2}} - \mathbf{z}_{j_k}^k - \alpha \nabla f_0(\mathbf{x}^{k+\frac{1}{2}}) \right), \\ \mathbf{z}_j^{k+1} &= \begin{cases} \mathbf{z}_j^k + \mathbf{x}^{k+1} - \mathbf{x}^{k+\frac{1}{2}} & \text{if } j = j_k, \\ \mathbf{z}_j^k & \text{if } j \neq j_k, \end{cases} \end{aligned}$$

where j_k is chosen from $\{1, \ldots, M\}$ uniformly at random. Since $\text{Proj}_{X_{j_k}}$ needs be evaluated, S-PPG has the same issue as the update in (1.20). However, it could be more suitable in a distributed system, for which communication cost is a main concern.

Stochastic subgradient with single projection. Let $h(\mathbf{x}) = \max_{1 \le j \le M} f_j(\mathbf{x})$. Then (1.1) is equivalent to

$$(1.23) \qquad \min_{\mathbf{x} \in X} f_0(\mathbf{x}), \text{ s.t. } h(\mathbf{x}) \le 0.$$

For solving the above problem, we can apply the method in [11], which, at every iteration, inquires a stochastic subgradient of f_0 and also a subgradient of h. Although the method in [11] only needs to perform a single projection to the feasible set at the last step, computing the subgradient of h would generally require evaluating the function value of all f_j's, and thus it is inefficient for the big-M case. This issue is partly addressed in [6], which only checks a batch of randomly sampled constraint functions at every iteration. However, depending on the underlying problem and required accuracy, the batch size could be as large as M.

Deterministic primal-dual first-order method. Other related methods are the deterministic primal-dual first-order algorithms in the author's previous works [27, 28]. Although [27, 28] also use the classic augmented Lagrangian function, their algorithm design and targeted applications are fundamentally different from those in this paper. The methods in [27, 28] assume differentiability of f_j's, and they require exact gradient of f_0 and use all $f_j, j = 1, \ldots, M$, to update $\mathbf{x}$ and $\mathbf{z}$. Hence, if the exact gradient of f_0 is not available or is very expensive to compute, or if M is extremely big, the deterministic methods are either inapplicable or inefficient. In addition, the update to $\mathbf{x}$ and $\mathbf{z}$ in Algorithm 1 is Jacobi-type while [27, 28] and all existing works about deterministic augmented Lagrangian method update the primal and dual variables in a Gauss–Seidel manner. Furthermore, due to the stochasticity, the analysis of this paper is fundamentally different and more complicated than that in [27, 28]. Similarly, the deterministic first-order methods in [9, 30, 31] are also very expensive or do not apply for the stochastic program with many constraints.

Besides the above reviewed methods, in the literature there are also other methods that can be applied to (1.1) such as the penalty method with stochastic approximation [8]. Exhausting all the existing methods is impossible. We refer interested readers to the papers above and the references therein.

1.3. Contributions. The main contributions are listed below.
- We propose a novel (adaptive) primal-dual SGM for solving stochastic programs with many functional constraints. The method is derived based on the classical augmented Lagrangian function. Through a stochastic oracle, it alternatingly performs a stochastic subgradient update to the primal variable and a randomized coordinate update to the dual variable. At each iteration, it only needs to sample one out of many constraint functions and thus has low per-iteration complexity.
- We establish convergence rate results of the proposed method for convex problems and also problems with a strongly convex objective. Different from existing analysis of the primal-dual SGM for saddle-point problems, we do not assume the boundedness of the dual variable $\mathbf{z}$, but instead we prove the boundedness of the dual iterate in expectation. For convex problems, we show that the algorithm can achieve the optimal $O(1/\sqrt{k})$ convergence rate, and for problems

with strongly convex objective, we show that it can achieve $O\big((\log k)/k\big)$ convergence rate, where k is the number of subgradient inquiries. All convergence rate results are in terms of primal and/or dual objective value and also primal constraint violation. For the strongly convex case, the $\log k$ factor can be removed if the dual iterate sequence is assumed to be bounded; see Remark 3.4. To the best of our knowledge, no existing work has established an $O(1/k)$ convergence rate result for a primal-dual SGM by assuming strong convexity only on the primal objective function, even if the dual variable is restricted in a bounded set. The CSA method in [8] is a primal SGM, and it has $O(1/k)$ convergence rate if both the objective and constraint functions are strongly convex.

- We show the practical performance of the proposed algorithm by testing it on solving a sample approximation problem of the robust portfolio selection (RPS) and convex quadratically constrained quadratic programs (QCQPs). The numerical results demonstrate that the proposed primal-dual SGM can be significantly better than the stochastic mirror-prox algorithm in [1] and the CSA method in [8].

We emphasize that in comparison to the existing methods mentioned in section 1.2, the novelty of our results lies in both algorithm design and convergence rate analysis. The incorporation of the adaptiveness technique is novel in the primal-dual setting. With adaptiveness, the proposed algorithm can have significantly better numerical performance. In addition, to achieve $O\big((\log k)/k\big)$ convergence rate, we only require strong convexity on the objective function but not on the constraint functions.

1.4. Notation and outline. We use bold lowercase letters $\mathbf{x}, \mathbf{z}, \ldots$ for vectors and $x_i, z_i, \ldots$ for their ith components. The bold numbers $\mathbf{0}$ and $\mathbf{1}$ denote the all-zero and all-one vectors, respectively. $[M]$ is short for the set $\{1, 2, \ldots, M\}$, and $[a]_+ = \max(0, a)$ and $[a]_- = \max(0, -a)$ respectively denote the positive and negative parts of a real number a. Given a symmetric positive semidefinite matrix $\mathbf{D}$, $\|\mathbf{x}\|_{\mathbf{D}}$ is defined as $\sqrt{\mathbf{x}^\top \mathbf{D} \mathbf{x}}$. We use $\|\mathbf{x}\|$ to denote the Euclidean norm of a vector $\mathbf{x}$. For two vectors $\mathbf{x}$ and $\mathbf{y}$ of the same size, $\mathbf{x} \odot \mathbf{y}$ denotes their componentwise product. For a convex function f, we denote by $\tilde{\nabla} f(\mathbf{x})$ a subgradient of f at $\mathbf{x}$, and the set of all subgradients of f at $\mathbf{x}$ is called the subdifferential of f, denoted by $\partial f(\mathbf{x})$. For a closed convex set X, Proj_X denotes the projection operator onto X. We let $\mathcal{H}^k$ contain the history of Algorithm 1 until $(\mathbf{x}^k, \mathbf{z}^k)$, i.e., $\mathcal{H}^k = \{\mathbf{x}^1, \mathbf{z}^1, \mathbf{x}^2, \mathbf{z}^2 \ldots, \mathbf{x}^k, \mathbf{z}^k\}$. $\mathbb{E}[\zeta]$ denotes the expectation of a random variable ζ, and $\mathbb{E}[\zeta \,|\, \xi]$ is for the expectation of ζ conditional on ξ. In addition, we denote

$$(1.24) \qquad \Phi(\bar{\mathbf{x}}; \mathbf{x}, \mathbf{z}) = f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}) + \frac{1}{M} \sum_{j=1}^{M} z_j f_j(\bar{\mathbf{x}}).$$

The rest of the paper is outlined as follows. In section 2, we give the technical assumptions required in our analysis, and in section 3, we analyze the algorithm with nonadaptive setting and show its convergence rate results. The convergence rate result of the algorithm with adaptive setting is given in section 4. Numerical results are provided in section 5, and finally section 6 concludes the paper.

2. Technical assumptions. Throughout our analysis, we make the following assumptions.

Assumption 1. There exists a primal-dual solution $(\mathbf{x}^*, \mathbf{z}^*)$ satisfying the Karush–Kuhn–Tucker (KKT) conditions:

$$0 \in \partial f_0(\mathbf{x}^*) + \mathcal{N}_X(\mathbf{x}^*) + \frac{1}{M}\sum_{j=1}^{M} z_j^* \partial f_j(\mathbf{x}^*), \tag{2.1a}$$

$$\mathbf{x}^* \in X, \ f_j(\mathbf{x}^*) \leq 0, \forall j \in [M], \tag{2.1b}$$

$$z_j^* \geq 0, \ z_j^* f_j(\mathbf{x}^*) = 0, \forall j \in [M], \tag{2.1c}$$

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

Assumption 2. The SG approximation $\mathbf{g}_0^k$ is unbiased and bounded, i.e., there is a constant $\sigma > 0$ such that

$$\mathbb{E}\big[\mathbf{g}_0^k \,\big|\, \mathcal{H}^k\big] \in \partial f_0(\mathbf{x}^k), \quad \mathbb{E}\big[\|\mathbf{g}_0^k\|^2 \,\big|\, \mathcal{H}^k\big] \leq \sigma^2 \,\forall k.$$

In addition, there exist constants F and G such that

$$|f_j(\mathbf{x})| \leq F, \|\tilde{\nabla} f_j(\mathbf{x})\| \leq G \ \forall \tilde{\nabla} f_j(\mathbf{x}) \in \partial f_j(\mathbf{x}), \ \forall j \in [M], \ \forall \mathbf{x} \in X.$$

Assumption 3. For each $j = 0, 1, \ldots, M$, f_j is a closed convex function on X. In addition, f_0 is μ-strongly convex, i.e.,

$$f_0(\mathbf{y}) \geq f_0(\mathbf{x}) + \langle \tilde{\nabla} f_0(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2 \ \forall \mathbf{x}, \mathbf{y} \in X. \tag{2.2}$$

Assumption 1 is satisfied if a certain constraint qualification holds such as the Slater's condition [2]. In Assumption 2, the unbiasedness and boundedness assumption on $\mathbf{g}_0^k$ is standard in the literature of SGM, and the boundedness of each f_j and $\tilde{\nabla} f_j$ is satisfied if X is bounded. In Assumption 3, if $\mu = 0$, then f_0 is simply a convex function.

As the KKT conditions in (2.1) hold, there are $\tilde{\nabla} f_j(\mathbf{x}^*) \forall j \in [M]$ such that

$$-\frac{1}{M}\sum_{j=1}^{M} z_j^* \tilde{\nabla} f_j(\mathbf{x}^*) \in \partial f_0(\mathbf{x}^*) + \mathcal{N}_X(\mathbf{x}^*).$$

Hence, from the convexity of f_0 and X, it follows that

$$f_0(\mathbf{x}) \geq f_0(\mathbf{x}^*) - \left\langle \frac{1}{M}\sum_{j=1}^{M} z_j^* \tilde{\nabla} f_j(\mathbf{x}^*)\mathbf{x} - \mathbf{x}^* \right\rangle \forall \mathbf{x} \in X. \tag{2.3}$$

Since $z_j^* \geq 0$ and f_j is convex for each $j \in [M]$, we have

$$z_j^*\big(f_j(\mathbf{x}) - f_j(\mathbf{x}^*)\big) \geq \langle z_j^* \tilde{\nabla} f_j(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle.$$

The above inequality together with (2.3) and the fact $z_j^* f_j(\mathbf{x}^*) = 0 \,\forall j \in [M]$ implies

$$\Phi(\mathbf{x}; \mathbf{x}^*, \mathbf{z}^*) = f_0(\mathbf{x}) - f_0(\mathbf{x}^*) + \frac{1}{M}\sum_{j=1}^{M} z_j^* f_j(\mathbf{x}) \geq 0 \,\forall \mathbf{x} \in X. \tag{2.4}$$

Furthermore, note that for any $\beta > 0$, it holds that $[\beta f_j(\mathbf{x}^*) + z_j^*]_+ = z_j^* \,\forall j \in [M]$, and thus (2.1a) exactly means $\mathbf{0} \in \partial_{\mathbf{x}} \mathcal{L}_\beta(\mathbf{x}^*, \mathbf{z}^*) + \mathcal{N}_X(\mathbf{x}^*)$. Hence, $\mathbf{x}^*$ is a solution of $\min_{\mathbf{x} \in X} \mathcal{L}_\beta(\mathbf{x}, \mathbf{z}^*)$, which indicates $d_\beta(\mathbf{z}^*) = \mathcal{L}_\beta(\mathbf{x}^*, \mathbf{z}^*)$. From the definitions of Ψ_β and ψ_β in (1.2) and (1.3), and also (2.1b) and (2.1c), it is straightforward to have $\Psi_\beta(\mathbf{x}^*, \mathbf{z}^*) = 0$. Therefore,

$$d_\beta(\mathbf{z}^*) = f_0(\mathbf{x}^*), \tag{2.5}$$

i.e., the strong duality holds, and $\mathbf{x}^*$ and $\mathbf{z}^*$ are primal and dual optimal solutions.

3. Convergence analysis of the nonadaptive method. For ease of understanding, we first analyze the convergence of Algorithm 1 with the nonadaptive Setting 1. Under Assumptions 1 through 3, we show that for convex problems, our method can achieve the optimal convergence rate $O(1/\sqrt{k})$, and for problems with strongly convex objective, it can achieve a near-optimal rate $O((\log k)/k)$, where k is the number of iterations. While existing analysis [1, 12] for saddle-point problems assumes the boundedness of the dual variable, we do not require such an assumption. Instead we can bound all $\mathbf{z}^k$ in expectation by choosing appropriate parameters. In addition, we do not find any existing work that has shown an $O((\log k)/k)$ rate for a primal-dual SGM by assuming strong convexity on the primal objective.

3.1. Preliminary results. We first establish a few preliminary results. The lemma below can be directly verified from the definition of Ψ_β.

LEMMA 3.1. *Let $\beta > 0$. Then for any $\mathbf{x} \in X$ such that $f_j(\mathbf{x}) \leq 0 \, \forall j \in [M]$ and any $\mathbf{z} \geq \mathbf{0}$, it holds that $\Psi_\beta(\mathbf{x}, \mathbf{z}) \leq 0$.*

The next lemma is important to establish the convergence rate of our algorithm. Similar ones in a deterministic form have appeared in [27, 28].

LEMMA 3.2. *Let $\bar{\mathbf{x}} \in X$ and $\bar{\mathbf{z}} \geq \mathbf{0}$ be random vectors, and let $\varepsilon_1 \geq 0$ and $\varepsilon_2 \geq 0$ be scalars. If for any $\mathbf{x} \in X$ and $\mathbf{z} \geq \mathbf{0}$ that may depend on $(\bar{\mathbf{x}}, \bar{\mathbf{z}})$, it holds that*

$$(3.1) \qquad \mathbb{E}\left[f_0(\bar{\mathbf{x}}) + \frac{1}{M}\sum_{j=1}^{M} z_j f_j(\bar{\mathbf{x}})\right] \leq \mathbb{E}\left[f_0(\mathbf{x}) + \Psi_\beta(\mathbf{x}, \bar{\mathbf{z}})\right] + \varepsilon_1 + \varepsilon_2 \mathbb{E}\|\mathbf{z}\|^2,$$

then for any $(\mathbf{x}^*, \mathbf{z}^*)$ satisfying (2.1),

$$(3.2) \qquad \mathbb{E}\big|f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*)\big| \leq 2\varepsilon_1 + 9\varepsilon_2\|\mathbf{z}^*\|^2,$$

$$(3.3) \qquad \mathbb{E}\left[\frac{1}{M}\sum_{j=1}^{M}[f_j(\bar{\mathbf{x}})]_+\right] \leq \varepsilon_1 + \varepsilon_2\|\mathbf{1} + \mathbf{z}^*\|^2,$$

$$(3.4) \qquad \mathbb{E}\big[d_\beta(\mathbf{z}^*) - d_\beta(\bar{\mathbf{z}})\big] \leq \frac{3}{2}\big(\varepsilon_1 + 3\varepsilon_2\|\mathbf{z}^*\|^2\big).$$

Proof. Let $\mathbf{x} = \mathbf{x}^*$ in (3.1) and recall the definition of Φ in (1.24). Then by Lemma 3.1, we have

$$(3.5) \qquad \mathbb{E}\big[\Phi(\bar{\mathbf{x}}; \mathbf{x}^*, \mathbf{z})\big] \leq \varepsilon_1 + \varepsilon_2 \mathbb{E}\|\mathbf{z}\|^2.$$

Since $-z_j^* f_j(\bar{\mathbf{x}}) \geq -z_j^*[f_j(\bar{\mathbf{x}})]_+$, we have from (2.4) that

$$(3.6) \qquad f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*) \geq -\frac{1}{M}\sum_{j=1}^{M} z_j^*[f_j(\bar{\mathbf{x}})]_+.$$

We obtain the inequality in (3.3) by substituting the above inequality into (3.5) with $\mathbf{z}$ given by $z_j = 1 + z_j^*$ if $f_j(\bar{\mathbf{x}}) > 0$ and $z_j = 0$ otherwise for any $j \in [M]$.

Letting $z_j = 3z_j^*$ if $f_j(\bar{\mathbf{x}}) > 0$ and $z_j = 0$ otherwise for each $j \in [M]$ in (3.5) and adding (3.6) together gives

$$(3.7) \qquad \mathbb{E}\left[\frac{1}{M}\sum_{j=1}^{M} z_j^*[f_j(\bar{\mathbf{x}})]_+\right] \leq \frac{\varepsilon_1}{2} + \frac{9\varepsilon_2}{2}\|\mathbf{z}^*\|^2.$$

Hence, by the above inequality and (3.6) we obtain $\mathbb{E}\big[f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*)\big]_- \leq \frac{\varepsilon_1}{2} + \frac{9\varepsilon_2}{2}\|\mathbf{z}^*\|^2$. In addition, from (3.5) with $\mathbf{z} = \mathbf{0}$, it follows that $\mathbb{E}[f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*)] \leq \varepsilon_1$. Since $|a| = a + 2[a]_-$ for any real number a, we have

$$\mathbb{E}\big|f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*)\big| = \mathbb{E}[f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*)] + 2\mathbb{E}\big[f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*)\big]_- \leq 2\varepsilon_1 + 9\varepsilon_2\|\mathbf{z}^*\|^2,$$

which gives (3.2).

Furthermore, in (3.1), let $\mathbf{z} = \mathbf{0}$ and take $\mathbf{x} \in \arg\min_{\mathbf{x} \in X} f_0(\mathbf{x}) + \Psi_\beta(\mathbf{x}, \bar{\mathbf{z}})$. We have $\mathbb{E}f_0(\bar{\mathbf{x}}) \leq \mathbb{E}d_\beta(\bar{\mathbf{z}}) + \varepsilon_1$, which together with (3.6), (3.7), and (2.5) gives the inequality in (3.4). $\qquad\square$

Remark 3.1. From the proof of Lemma 3.2, we see that if (3.5) holds for any $\mathbf{z} \geq \mathbf{0}$, then the inequalities in (3.2) and (3.3) hold.

The following two lemmas will be used to establish an important inequality for running one iteration of Algorithm 1. Their proofs are given in the appendix.

LEMMA 3.3. *For any deterministic or stochastic* $\mathbf{z} \geq \mathbf{0}$, *it holds that*

$$(3.8) \qquad -\Psi_\beta(\mathbf{x}^k, \mathbf{z}^k) + \frac{1}{M}\sum_{j=1}^M z_j f_j(\mathbf{x}^k) + \frac{1}{2\rho_k}\mathbb{E}\big[\|\mathbf{z}^{k+1} - \mathbf{z}\|^2 \,\big|\, \mathcal{H}^k\big]$$

$$\leq \frac{1}{2\rho_k}\|\mathbf{z}^k - \mathbf{z}\|^2 - \frac{1}{2\rho_k}\left(\frac{\beta}{\rho_k} - 1\right)\mathbb{E}\big[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 \,\big|\, \mathcal{H}^k\big]$$
$$+ \mathbb{E}\big[\big\langle \mathbf{z}^k - \mathbf{z}, M\mathbf{e}_{j_k} \odot \nabla_\mathbf{z}\Psi\left(\mathbf{x}^k, \mathbf{z}^k\right) - \nabla_\mathbf{z}\Psi\left(\mathbf{x}^k, \mathbf{z}^k\right)\big\rangle \,\big|\, \mathcal{H}^k\big].$$

LEMMA 3.4. *Under Assumption 2, for any* $\mathbf{x} \in X$ *and any* $\mathbf{z}$, *it holds that*

$$(3.9) \qquad \frac{1}{M}\sum_{j=1}^M \|\tilde{\nabla}_\mathbf{x}\psi_\beta(f_j(\mathbf{x}), z_j)\|^2 \leq 2\beta^2 F^2 G^2 + \frac{2G^2}{M}\|\mathbf{z}\|^2.$$

By the previous three lemmas, we establish an important result for running one iteration of Algorithm 1 and then use it to show the convergence rate results.

THEOREM 3.5 (fundamental result). *Under Assumptions 2 and 3, and assuming* $\mathbf{D}_k \succeq \frac{\mathbf{I}}{\alpha_k} \forall k$ *for a positive number sequence* $\{\alpha_k\}_{k \geq 1}$, *let* $(\mathbf{x}, \mathbf{z})$ *be any deterministic or stochastic vector such that* $\mathbf{x} \in X$ *and* $\mathbf{z} \geq \mathbf{0}$. *Then*

$$\mathbb{E}\left[f_0(\mathbf{x}^k) + \frac{1}{M}\sum_{j=1}^M z_j f_j(\mathbf{x}^k)\right] + \frac{1}{2}\mathbb{E}\|\mathbf{x}^{k+1} - \mathbf{x}\|_{\mathbf{D}_k}^2 + \frac{1}{2\rho_k}\mathbb{E}\|\mathbf{z}^{k+1} - \mathbf{z}\|^2$$

$$(3.10) \qquad \leq \mathbb{E}\big[f_0(\mathbf{x}) + \Psi_\beta(\mathbf{x}, \mathbf{z}^k)\big] + \frac{1}{2}\mathbb{E}\|\mathbf{x}^k - \mathbf{x}\|_{\mathbf{D}_k - \mu\mathbf{I}}^2 + \frac{1}{2\rho_k}\mathbb{E}\|\mathbf{z}^k - \mathbf{z}\|^2$$

$$+ \alpha_k\left(\sigma^2 + 2\beta^2 F^2 G^2 + \frac{2G^2}{M}\mathbb{E}\|\mathbf{z}^k\|^2\right) - \frac{1}{2\rho_k}\left(\frac{\beta}{\rho_k} - 1\right)\mathbb{E}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2$$

$$- \mathbb{E}\left[\big\langle \mathbf{x}^k - \mathbf{x}, \mathbf{g}_0^k - \tilde{\nabla}f_0(\mathbf{x}^k)\big\rangle\right] - \mathbb{E}\left[\big\langle \mathbf{x}^k - \mathbf{x}, \mathbf{h}^k - \tilde{\nabla}_\mathbf{x}\Psi_\beta(\mathbf{x}^k, \mathbf{z}^k)\big\rangle\right]$$

$$+ \mathbb{E}\left[\big\langle \mathbf{z}^k - \mathbf{z}, M\mathbf{e}_{j_k} \odot \nabla_\mathbf{z}\Psi\left(\mathbf{x}^k, \mathbf{z}^k\right) - \nabla_\mathbf{z}\Psi\left(\mathbf{x}^k, \mathbf{z}^k\right)\big\rangle\right],$$

where $\tilde{\nabla}f_0(\mathbf{x}^k) = \mathbb{E}[\mathbf{g}_0^k \,|\, \mathcal{H}^k]$ *and* $\tilde{\nabla}_\mathbf{x}\Psi_\beta(\mathbf{x}^k, \mathbf{z}^k) = \mathbb{E}[\mathbf{h}^k \,|\, \mathcal{H}^k]$.

Proof. From the update (1.6), it follows that for any $\mathbf{x} \in X$,

$$(3.11) \qquad \big\langle \mathbf{x}^{k+1} - \mathbf{x}, \mathbf{g}_0^k + \mathbf{h}^k + \mathbf{D}_k\left(\mathbf{x}^{k+1} - \mathbf{x}^k\right)\big\rangle \leq 0.$$

Next we estimate a lower bound about the left-hand side of the above inequality. First, we write $\langle \mathbf{x}^{k+1} - \mathbf{x}, \mathbf{g}_0^k \rangle = \langle \mathbf{x}^{k+1} - \mathbf{x}^k, \mathbf{g}_0^k \rangle + \langle \mathbf{x}^k - \mathbf{x}, \mathbf{g}_0^k \rangle$. By the Young's inequality, it holds that

$$(3.12) \qquad \langle \mathbf{x}^{k+1} - \mathbf{x}^k, \mathbf{g}_0^k \rangle \geq -\frac{1}{4\alpha_k} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 - \alpha_k \|\mathbf{g}_0^k\|^2.$$

Also, we write

$$\langle \mathbf{x}^k - \mathbf{x}, \mathbf{g}_0^k \rangle = \left\langle \mathbf{x}^k - \mathbf{x}, \tilde{\nabla} f_0(\mathbf{x}^k) \right\rangle + \left\langle \mathbf{x}^k - \mathbf{x}, \mathbf{g}_0^k - \tilde{\nabla} f_0(\mathbf{x}^k) \right\rangle,$$

where $\tilde{\nabla} f_0(\mathbf{x}^k) = \mathbb{E}[\mathbf{g}_0^k \,|\, \mathcal{H}^k] \in \partial f_0(\mathbf{x}^k)$. Hence, from (3.12) and (2.2), it follows that

$$\langle \mathbf{x}^{k+1} - \mathbf{x}, \mathbf{g}_0^k \rangle \geq -\frac{1}{4\alpha_k} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 - \alpha_k \|\mathbf{g}_0^k\|^2 + f_0(\mathbf{x}^k) - f_0(\mathbf{x}) + \frac{\mu}{2} \|\mathbf{x}^k - \mathbf{x}\|^2$$
$$+ \left\langle \mathbf{x}^k - \mathbf{x}, \mathbf{g}_0^k - \tilde{\nabla} f_0(\mathbf{x}^k) \right\rangle.$$

Taking conditional expectation, we have from the above inequality and Assumption 2 that

(3.13)

$$\mathbb{E}\left[\langle \mathbf{x}^{k+1} - \mathbf{x}, \mathbf{g}_0^k \rangle \,|\, \mathcal{H}^k \right]$$
$$\geq -\frac{1}{4\alpha_k} \mathbb{E}\left[\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \,|\, \mathcal{H}^k \right] - \alpha_k \sigma^2 + \mathbb{E}\left[f_0\left(\mathbf{x}^k\right) - f_0(\mathbf{x}) + \frac{\mu}{2} \|\mathbf{x}^k - \mathbf{x}\|^2 \,|\, \mathcal{H}^k \right]$$
$$+ \mathbb{E}\left[\left\langle \mathbf{x}^k - \mathbf{x}, \mathbf{g}_0^k - \tilde{\nabla} f_0(\mathbf{x}^k) \right\rangle \,|\, \mathcal{H}^k \right].$$

Similar to (3.13), we have

(3.14)

$$\mathbb{E}\left[\langle \mathbf{x}^{k+1} - \mathbf{x}, \mathbf{h}^k \rangle \,|\, \mathcal{H}^k \right]$$
$$\geq -\mathbb{E}\left[\frac{1}{4\alpha_k} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \alpha_k \|\mathbf{h}^k\|^2 \,|\, \mathcal{H}^k \right] + \mathbb{E}\left[\Psi_\beta\left(\mathbf{x}^k, \mathbf{z}^k\right) - \Psi_\beta\left(\mathbf{x}, \mathbf{z}^k\right) \,|\, \mathcal{H}^k \right]$$
$$+ \mathbb{E}\left[\left\langle \mathbf{x}^k - \mathbf{x}, \mathbf{h}^k - \tilde{\nabla}_{\mathbf{x}} \Psi_\beta\left(\mathbf{x}^k, \mathbf{z}^k\right) \right\rangle \,|\, \mathcal{H}^k \right],$$

where $\tilde{\nabla}_{\mathbf{x}} \Psi_\beta(\mathbf{x}^k, \mathbf{z}^k) = \mathbb{E}[\mathbf{h}^k \,|\, \mathcal{H}^k]$. Since j_k is chosen from $[M]$ uniformly at random, by (1.5), (3.9), and the Young's inequality, we have

$$-\alpha_k \mathbb{E}\left[\|\mathbf{h}^k\|^2 \,|\, \mathcal{H}^k \right] = -\frac{\alpha_k}{M} \sum_{j=1}^{M} \left\| \tilde{\nabla}_{\mathbf{x}} \psi_\beta\left(f_j(\mathbf{x}^k), z_j^k \right) \right\|^2$$
$$(3.15) \qquad\qquad\qquad \geq -\alpha_k \left(2\beta^2 F^2 G^2 + \frac{2G^2}{M} \|\mathbf{z}^k\|^2 \right).$$

In addition,

(3.16)

$$\langle \mathbf{x}^{k+1} - \mathbf{x}, \mathbf{D}_k(\mathbf{x}^{k+1} - \mathbf{x}^k) \rangle = \frac{1}{2} \left[\|\mathbf{x}^{k+1} - \mathbf{x}\|_{\mathbf{D}_k}^2 - \|\mathbf{x}^k - \mathbf{x}\|_{\mathbf{D}_k}^2 + \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_{\mathbf{D}_k}^2 \right].$$

Taking expectation on both sides of (3.13) through (3.16), summing them up, substituting into (3.11), and noting $\mathbf{D}_k \succeq \frac{\mathbf{I}}{\alpha_k}$ gives

$$\mathbb{E}\left[f_0(\mathbf{x}^k) - f_0(\mathbf{x}) + \Psi_\beta(\mathbf{x}^k, \mathbf{z}^k) - \Psi_\beta(\mathbf{x}, \mathbf{z}^k)\right] + \frac{1}{2}\mathbb{E}\|\mathbf{x}^{k+1} - \mathbf{x}\|_{\mathbf{D}_k}^2$$

$$(3.17) \qquad \leq \frac{1}{2}\mathbb{E}\|\mathbf{x}^k - \mathbf{x}\|_{\mathbf{D}_k - \mu\mathbf{I}}^2 + \alpha_k\left(\sigma^2 + 2\beta^2 F^2 G^2 + \frac{2G^2}{M}\mathbb{E}\|\mathbf{z}^k\|^2\right)$$

$$- \mathbb{E}\left[\left\langle \mathbf{x}^k - \mathbf{x}, \mathbf{g}_0^k - \tilde{\nabla}f_0(\mathbf{x}^k)\right\rangle\right] - \mathbb{E}\left[\left\langle \mathbf{x}^k - \mathbf{x}, \mathbf{h}^k - \tilde{\nabla}_{\mathbf{x}}\Psi_\beta(\mathbf{x}^k, \mathbf{z}^k)\right\rangle\right].$$

Taking expectation on both sides of (3.8), adding it to (3.17), and rearranging terms yield the desired result.   □

By Theorem 3.5, we can bound the growth of $\mathbb{E}\|\mathbf{z}^k\|^2$ as below. Its proof is given in the appendix.

PROPOSITION 3.6. *Under Assumptions* 1 *through* 3, *and assuming* $\mathbf{D}_k \succeq \frac{\mathbf{I}}{\alpha_k} \forall k$ *for a positive number sequence* $\{\alpha_k\}_{k\geq 1}$, *let* $\left\{(\mathbf{x}^k, \mathbf{z}^k)\right\}$ *be the sequence generated from Algorithm* 1 *with parameters satisfying*

$$(3.18) \qquad\qquad \rho_k\mathbf{D}_k \succeq \rho_{k+1}(\mathbf{D}_{k+1} - \mu\mathbf{I})\,\forall k \geq 1;$$

then for any $t \geq 1$, *it holds that*

(3.19)

$$\mathbb{E}\|\mathbf{z}^{t+1}\|^2$$

$$\leq 2\rho_1\|\mathbf{x}^1 - \mathbf{x}^*\|_{\mathbf{D}_1 - \mu\mathbf{I}}^2 + 4\|\mathbf{z}^*\|^2 + \sum_{k=1}^{t} 4\alpha_k\rho_k\left(\sigma^2 + 2\beta^2 F^2 G^2 + \frac{2G^2}{M}\mathbb{E}\|\mathbf{z}^k\|^2\right),$$

where $(\mathbf{x}^*, \mathbf{z}^*)$ *is any point satisfying the KKT conditions in* (2.1).

3.2. Convergence rate for convex problems. In this subsection, we establish the convergence rate of Algorithm 1 for convex problems, i.e., $\mu = 0$. Different from existing analysis for saddle-point problems, we do not assume the boundedness of the dual variable $\mathbf{z}$ but instead we can bound $\mathbf{z}^k$ in expectation.

Using Proposition 3.6, we specify the parameters and bound $\mathbb{E}\|\mathbf{z}^k\|^2$. The proofs of both propositions below are given in the appendix.

PROPOSITION 3.7 (predetermined maximum number of iterations).   *Under Assumptions* 1 *through* 3, *given a positive integer* K, *set*

$$(3.20) \qquad\qquad \mathbf{D}_k = \frac{\sqrt{K}}{\alpha}\mathbf{I},\ \rho_k = \frac{\rho}{\sqrt{K}},\ \beta \geq \rho,\ \forall 1 \leq k \leq K,$$

where α, ρ, *and* β *are positive scalars satisfying* $\alpha\rho < \frac{M}{8G^2}$. *Then for any* $1 \leq k \leq K + 1$, *it holds that*

$$(3.21) \qquad\qquad \mathbb{E}\|\mathbf{z}^k\|^2 \leq \frac{C_1}{1 - \frac{8\alpha\rho G^2}{M}},$$

where

$$(3.22) \qquad C_1 = \frac{2\rho}{\alpha}\|\mathbf{x}^1 - \mathbf{x}^*\|^2 + 4\|\mathbf{z}^*\|^2 + 4\alpha\rho(\sigma^2 + 2\beta^2 F^2 G^2).$$

If the maximum number of iterations is not predetermined, we set parameters adaptive to iteration numbers and can still bound $\mathbb{E}\|\mathbf{z}^k\|^2$.

PROPOSITION 3.8 (varying maximum number of iterations). *Under Assumptions 1 through 3, let $\{(\mathbf{x}^k, \mathbf{z}^k)\}$ be the sequence generated from Algorithm 1 with parameters set to*

$$(3.23) \qquad \mathbf{D}_k = \frac{\sqrt{k+1}\log(k+1)}{\alpha}\mathbf{I}, \ \rho_k = \frac{\rho}{\sqrt{k+1}\log(k+1)}, \ \beta \geq \rho, \ \forall k \geq 1,$$

where α, ρ, and β are positive scalars satisfying $\alpha\rho < \frac{M}{20G^2}$. Then for any $k \geq 1$, it holds that

$$(3.24) \qquad \mathbb{E}\|\mathbf{z}^k\|^2 \leq \frac{C_2}{1 - \frac{20\alpha\rho G^2}{M}},$$

where

$$(3.25) \qquad C_2 = \frac{2\rho}{\alpha}\|\mathbf{x}^1 - \mathbf{x}^*\|^2 + 4\|\mathbf{z}^*\|^2 + 10\alpha\rho(\sigma^2 + 2\beta^2 F^2 G^2).$$

To show the convergence rate results, we need the following lemma to handle the last three expectation terms in (3.10). Its proof is given in the appendix and follows the proof of [13, Lemma 3.1].

LEMMA 3.9. *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 $\{\alpha_k\}$ that*

$$(3.26) \qquad -\sum_{k=1}^{K}\alpha_k\mathbb{E}\left\langle\mathbf{x}^k - \mathbf{x}, \mathbf{g}_0^k - \tilde{\nabla}f_0(\mathbf{x}^k)\right\rangle \leq \frac{1}{2}\mathbb{E}\|\mathbf{x}^1 - \mathbf{x}\|^2 + \frac{\sigma^2}{2}\sum_{k=1}^{K}\alpha_k^2,$$

$$(3.27) \qquad \begin{aligned} -\sum_{k=1}^{K}\alpha_k\mathbb{E}\left\langle\mathbf{x}^k - \mathbf{x}, \mathbf{h}^k - \nabla_{\mathbf{x}}\Psi_\beta(\mathbf{x}^k, \mathbf{z}^k)\right\rangle &\leq \frac{1}{2}\mathbb{E}\|\mathbf{x}^1 - \mathbf{x}\|^2 \\ &+ \sum_{k=1}^{K}\alpha_k^2\left(\beta^2 F^2 G^2 + \frac{G^2}{M}\mathbb{E}\|\mathbf{z}^k\|^2\right), \end{aligned}$$

$$(3.28) \qquad \begin{aligned} \sum_{k=1}^{K}\alpha_k\mathbb{E}\left\langle\mathbf{z}^k - \mathbf{z}, M\mathbf{e}_{j_k} \odot \nabla_{\mathbf{z}}\Psi(\mathbf{x}^k, \mathbf{z}^k) - \nabla_{\mathbf{z}}\Psi(\mathbf{x}^k, \mathbf{z}^k)\right\rangle \\ \leq \frac{1}{2}\mathbb{E}\|\mathbf{z}^1 - \mathbf{z}\|^2 + \frac{F^2}{2}\sum_{k=1}^{K}\alpha_k^2. \end{aligned}$$

Using Theorem 3.5 and also the boundedness of $\mathbb{E}\|\mathbf{z}^k\|^2$, we are now ready to show the convergence rate results for the case $\mu = 0$. First, we establish a result with constant step sizes, and the order is $O(1/\sqrt{k})$, where k is the iteration number.

THEOREM 3.10 (convergence rate for convex case with constant step sizes). *Under Assumptions 1 through 3, let $\{(\mathbf{x}^k, \mathbf{z}^k)\}$ be the sequence generated from Algorithm 1. Given any positive integer K, set the parameters according to (3.20), let $\bar{\mathbf{x}}^K = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}^k$ and $\bar{\mathbf{z}}^K = \frac{1}{K}\sum_{k=1}^{K}\mathbf{z}^k$, and define*

$$(3.29) \qquad \phi_1(\mathbf{x}) = \frac{3}{2\alpha}\|\mathbf{x}^1 - \mathbf{x}\|^2 + \alpha\left(\frac{3}{2}\sigma^2 + 3\beta^2 F^2 G^2 + \frac{3G^2}{M}\frac{C_1}{1 - \frac{8\alpha\rho G^2}{M}} + \frac{F^2}{2}\right),$$

where C_1 is defined in (3.22). Then

$$(3.30a) \qquad \mathbb{E}\big|f_0(\bar{\mathbf{x}}^K) - f_0(\mathbf{x}^*)\big| \leq \frac{1}{\sqrt{K}}\left(2\phi_1(\mathbf{x}^*) + \frac{9(\alpha+\rho)}{2\alpha\rho}\|\mathbf{z}^*\|^2\right),$$

$$(3.30b) \qquad \mathbb{E}\left[\frac{1}{M}\sum_{j=1}^{M}[f_j(\bar{\mathbf{x}}^K)]_+\right] \leq \frac{1}{\sqrt{K}}\left(\phi_1(\mathbf{x}^*) + \frac{\alpha+\rho}{2\alpha\rho}\|\mathbf{1} + \mathbf{z}^*\|^2\right).$$

In addition, if X is bounded, then

$$(3.30c) \qquad \mathbb{E}\big[d_\beta(\mathbf{z}^*) - d_\beta(\bar{\mathbf{z}}^K)\big] \leq \frac{3}{2\sqrt{K}} \left(\max_{\mathbf{x}\in X} \phi_1(\mathbf{x}) + \frac{3(\alpha+\rho)}{2\alpha\rho} \|\mathbf{z}^*\|^2 \right).$$

Proof. When the parameters are set according to (3.20), we have (3.21). Hence, multiplying $\alpha_k = \frac{\alpha}{\sqrt{K}}$ to (3.10), summing it up from $k=1$ through K, using (3.26) through (3.28), and noting $\mathbf{z}^1 = \mathbf{0}$ give

$$\frac{\alpha}{\sqrt{K}} \sum_{k=1}^{K} \mathbb{E}\left[f_0(\mathbf{x}^k) + \frac{1}{M} \sum_{j=1}^{M} z_j f_j(\mathbf{x}^k) \right] \leq \frac{\alpha}{\sqrt{K}} \sum_{k=1}^{K} \mathbb{E}\left[f_0(\mathbf{x}) + \Psi_\beta(\mathbf{x}, \mathbf{z}^k) \right]$$

$$+ \frac{3}{2}\mathbb{E}\|\mathbf{x}^1 - \mathbf{x}\|^2 + \left(\frac{\alpha}{2\rho} + \frac{1}{2} \right) \mathbb{E}\|\mathbf{z}\|^2 + \alpha^2 \left(\frac{3}{2}\sigma^2 + 3\beta^2 F^2 G^2 + \frac{3G^2}{M} \frac{C_1}{1 - \frac{8\alpha\rho G^2}{M}} + \frac{F^2}{2} \right).$$

Since $\mathbf{z} \geq \mathbf{0}$, by the convexity of f_j's and also concavity of Ψ_β about $\mathbf{z}$, we have from the above inequality and the definition of ϕ_1 in (3.29) that

$$(3.31)$$
$$\mathbb{E}\left[f_0(\bar{\mathbf{x}}^K) + \frac{1}{M} \sum_{j=1}^{M} z_j f_j(\bar{\mathbf{x}}^K) \right] \leq \mathbb{E}\left[f_0(\mathbf{x}) + \Psi_\beta(\mathbf{x}, \bar{\mathbf{z}}^K) \right] + \frac{1}{\sqrt{K}} \mathbb{E}\left[\phi_1(\mathbf{x}) + \frac{\alpha+\rho}{2\alpha\rho} \|\mathbf{z}\|^2 \right].$$

Let $\mathbf{x} = \mathbf{x}^*$ in the above inequality. Then by Lemma 3.1 and the definition of Φ in (1.24), we have

$$\mathbb{E}\big[\Phi(\bar{\mathbf{x}}^K; \mathbf{x}^*, \mathbf{z})\big] \leq \frac{\phi_1(\mathbf{x}^*)}{\sqrt{K}} + \frac{1}{\sqrt{K}} \frac{\alpha+\rho}{2\alpha\rho} \mathbb{E}\|\mathbf{z}\|^2 \quad \forall \mathbf{z} \geq \mathbf{0}.$$

Hence, (3.30a) and (3.30b) follow from the proof of Lemma 3.2 and Remark 3.1.

Furthermore, as X is bounded, the inequality (3.31) implies

$$\mathbb{E}\left[f_0(\bar{\mathbf{x}}^K) + \frac{1}{M} \sum_{j=1}^{M} z_j f_j(\bar{\mathbf{x}}^K) \right]$$

$$\leq \mathbb{E}\left[f_0(\mathbf{x}) + \Psi_\beta(\mathbf{x}, \bar{\mathbf{z}}^K) \right] + \frac{1}{\sqrt{K}} \left[\max_{\mathbf{x}\in X} \phi_1(\mathbf{x}) + \frac{\alpha+\rho}{2\alpha\rho} \mathbb{E}\|\mathbf{z}\|^2 \right].$$

Therefore, we obtain (3.30c) from Lemma 3.2 and complete the proof. $\quad\square$

Below we make a few remarks about the results in Theorem 3.10. Similar remarks also apply to Theorems 3.11 and 3.14 established later.

Remark 3.2. From the proof of Theorem 3.10, we see that the setting of ρ_k is for bounding $\mathbb{E}\|\mathbf{z}^k\|^2$. If the dual variable $\mathbf{z}$ is bounded, then ρ_k can be taken as large as the augmented penalty parameter β.

Remark 3.3. By the Markov's inequality $\text{Prob}(\xi \geq \varepsilon) \leq \frac{\mathbb{E}[\xi]}{\varepsilon}$ for a nonnegative random variable ξ, one can easily have a high-probability result from Theorem 3.10. One drawback of the result is that in (3.30b), the bound is on the average of all inequality constraint violations. Let $\gamma = \frac{\mathbb{E}[\max_{j\in[M]}[f_j(\bar{\mathbf{x}}^K)]_+]}{\mathbb{E}[\frac{1}{M}\sum_{j=1}^{M}[f_j(\bar{\mathbf{x}}^K)]_+]}$. Then (3.30b) implies $\mathbb{E}[\max_{j\in[M]}[f_j(\bar{\mathbf{x}}^K)]_+] \leq \frac{\gamma}{\sqrt{K}}(\phi_1(\mathbf{x}^*) + \frac{\|\mathbf{1}+\mathbf{z}^*\|^2}{2\rho})$. If $\gamma = O(1)$, then the maximum violation of the inequality constraint is similar to the avarage violation. However, in the worse case, γ could be as large as M.

One may argue that since the averaged constraint violation is used as a measure in the convergence rate result, it could be more natural to work on the equivalent problem (1.16), for which only one dual variable is needed instead of the many more M dual variables required in Algorithm 1. We point out two potential issues to pursue in this direction. First, the augmented Lagrangian function of (1.16) has a term that is a composition of ψ_β given in (1.3) with the finite-sum $\frac{1}{M}\sum_{j=1}^{M}[f_j(\mathbf{x})]_+$. For a stochastic program with such a nested structure, the convergence rate of SGM is much worse [26] due to the difficulty of obtaining an unbiased SG. Second, the Slater's condition can never hold for (1.16). Hence, although one dual variable is needed, the existence of a KKT point is not guaranteed even if the Slater's condition holds for the original problem (1.1), and this would affect the convergence analysis. Also, we point out that the use of M dual variables does not cause an issue of memory or computational cost. Compared to the data involved in the M constraint functions, the size of M dual variables is smaller.

With varying step sizes, we can also show a sublinear convergence rate result of Algorithm 1 as follows. The order is worse with an additional logarithmic term.

THEOREM 3.11 (convergence rate for convex case with varying step sizes). *Under Assumptions* 1 *through* 3, *let* $\{(\mathbf{x}^k, \mathbf{z}^k)\}$ *be the sequence generated from Algorithm* 1. *Set parameters according to* (3.23). *For any integer* $K \geq 1$, *let* $\alpha_k = \frac{\alpha}{\sqrt{k+1}\log(k+1)}$ *for* $1 \leq k \leq K$, $\bar{\mathbf{x}}^K = \frac{1}{\sum_{k=1}^{K}\alpha_k}\sum_{k=1}^{K}\alpha_k\mathbf{x}^k$ *and* $\bar{\mathbf{z}}^K = \frac{1}{\sum_{k=1}^{K}\alpha_k}\sum_{k=1}^{K}\alpha_k\mathbf{z}^k$, *and define*

$$(3.32) \quad \phi_2(\mathbf{x}) = \frac{3}{2\alpha}\|\mathbf{x}^1 - \mathbf{x}\|^2 + 2.5\alpha\left(\frac{3}{2}\sigma^2 + 3\beta^2F^2G^2 + \frac{3G^2}{M}\frac{C_2}{1-\frac{20\alpha\rho G^2}{M}} + \frac{F^2}{2}\right),$$

with C_2 *defined in* (3.25). *Then*

$$(3.33a) \quad \mathbb{E}\big|f_0(\bar{\mathbf{x}}^{K+1}) - f_0(\mathbf{x}^*)\big| \leq \frac{\log(K+1)}{2(\sqrt{K+2}-\sqrt{2})}\left(2\phi_2(\mathbf{x}^*) + \frac{9(\alpha+\rho)}{2\alpha\rho}\|\mathbf{z}^*\|^2\right),$$

$$(3.33b) \quad \mathbb{E}\left[\frac{1}{M}\sum_{j=1}^{M}[f_j(\bar{\mathbf{x}}^{K+1})]_+\right] \leq \frac{\log(K+1)}{2(\sqrt{K+2}-\sqrt{2})}\left(\phi_2(\mathbf{x}^*) + \frac{\alpha+\rho}{2\alpha\rho}\|\mathbf{1}+\mathbf{z}^*\|^2\right).$$

In addition, if X *is bounded, then*

$$(3.33c) \quad \mathbb{E}\big[d_\beta(\mathbf{z}^*) - d_\beta(\bar{\mathbf{z}}^K)\big] \leq \frac{3\log(K+1)}{4(\sqrt{K+2}-\sqrt{2})}\left(\max_{\mathbf{x}\in X}\phi_2(\mathbf{x}) + \frac{3(\alpha+\rho)}{2\alpha\rho}\|\mathbf{z}^*\|^2\right).$$

Proof. When the parameters are set according to (3.23), we have (3.24). Hence, multiplying α_k to both sides of (3.10), summing it over k, and using (3.26) through (3.28), we have

$$\sum_{k=1}^{K}\alpha_k\mathbb{E}\left[f_0(\mathbf{x}^k) + \frac{1}{M}\sum_{j=1}^{M}z_j f_j(\mathbf{x}^k)\right] \leq \sum_{k=1}^{K}\alpha_k\mathbb{E}\left[f_0(\mathbf{x}) + \Psi_\beta(\mathbf{x},\mathbf{z}^k)\right] + \frac{3}{2}\mathbb{E}\|\mathbf{x}^1-\mathbf{x}\|^2$$

$$(3.34)$$

$$+ \left(\frac{\alpha}{2\rho} + \frac{1}{2}\right)\mathbb{E}\|\mathbf{z}\|^2 + \sum_{k=1}^{K}\alpha_k^2\left(\frac{3}{2}\sigma^2 + 3\beta^2F^2G^2 + \frac{3G^2}{M}\frac{C_2}{1-\frac{20\alpha\rho G^2}{M}} + \frac{F^2}{2}\right).$$

Note

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

Hence, dividing both sides of (3.34) by $\sum_{k=1}^{K} \alpha_k$, we have from the convexity of f_j's and the concavity of Ψ_β about $\mathbf{z}$, and also using $\sum_{k=1}^{K} \alpha_k^2 \leq 2.5\alpha^2$ from (A.3) and the definition of ϕ_2 in (3.32), that

$$\mathbb{E}\left[f_0(\bar{\mathbf{x}}^K) + \frac{1}{M}\sum_{j=1}^{M} z_j f_j(\bar{\mathbf{x}}^K)\right]$$

$$\leq \mathbb{E}\left[f_0(\mathbf{x}) + \Psi_\beta(\mathbf{x}, \bar{\mathbf{z}}^K)\right] + \frac{\log(K+1)}{2(\sqrt{K+2}-\sqrt{2})}\mathbb{E}\left[\phi_2(\mathbf{x}) + \frac{\alpha+\rho}{2\alpha\rho}\|\mathbf{z}\|^2\right].$$

Now following the same arguments as those below (3.31) in the proof of Theorem 3.10, we obtain the desired results and complete the proof. □

3.3. Convergence rate for strongly convex problems. In this subsection, we analyze the convergence rate of Algorithm 1 for strongly convex problems, i.e., $\mu > 0$ in (2.2). Similar to the convex case, we first bound $\mathbb{E}\|\mathbf{z}^k\|^2$ by choosing appropriate parameters. The proof is shown in the appendix.

PROPOSITION 3.12. *Under Assumptions* 1 *through* 3 *with* $\mu > 0$, *for any given positive integer* K, *let* $\{(\mathbf{x}^k, \mathbf{z}^k)\}$ *be the sequence generated from Algorithm* 1 *with parameters set to*

$$(3.35) \qquad \mathbf{D}_k = \frac{k+1}{\alpha}\mathbf{I}, \ \rho_k = \frac{\rho}{\log(K+1)}, \ \beta \geq \frac{2\rho}{\log 2} \ \forall 1 \leq k \leq K,$$

where $\alpha \geq \frac{1}{\mu}$ *and* $\alpha\rho < \frac{M}{8G^2}$. *Then for any* $1 \leq k \leq K+1$,

$$(3.36) \qquad \mathbb{E}\|\mathbf{z}^k\|^2 \leq \frac{C_3}{1 - \frac{8\alpha\rho G^2}{M}},$$

where

$$(3.37) \qquad C_3 = \frac{2\rho}{\log(K+1)}\left(\frac{2}{\alpha} - \mu\right)\|\mathbf{x}^1 - \mathbf{x}^*\|^2 + 4\|\mathbf{z}^*\|^2 + 4\alpha\rho(\sigma^2 + 2\beta^2 F^2 G^2).$$

Similar to Lemma 3.9, we have the following result bounding the expectation terms in (3.10). The proof is also given in the appendix.

LEMMA 3.13. *Under the assumptions of Proposition* 3.12, *for any deterministic or stochastic vector* $\mathbf{z} \geq \mathbf{0}$, *we have*

$$\sum_{k=1}^{K} \mathbb{E}\left[\langle \mathbf{z}^k - \mathbf{z}, M\mathbf{e}_{j_k} \odot \nabla_{\mathbf{z}}\Psi(\mathbf{x}^k, \mathbf{z}^k) - \nabla_{\mathbf{z}}\Psi(\mathbf{x}^k, \mathbf{z}^k)\rangle\right]$$

$$(3.38) \qquad \leq \frac{\log(K+1)}{2\rho}\mathbb{E}\left[\|\mathbf{z}^1 - \mathbf{z}\|^2 + \sum_{k=1}^{K}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2\right].$$

Using (3.10) and (3.36), we establish the convergence rate result of Algorithm 1 for the case of $\mu > 0$ as follows.

THEOREM 3.14 (convergence rate for strongly convex case).   *Under the assumptions of Proposition* 3.12, *we have*

$$(3.39) \qquad \mathbb{E}\|\mathbf{x}^{K+1} - \mathbf{x}^*\|^2 \leq \frac{2\alpha}{K+1}\left(\phi_3(\mathbf{x}^*) + \frac{\log(K+1)}{\rho}\|\mathbf{z}^*\|^2\right),$$

where

$$(3.40)$$

$$\phi_3(\mathbf{x}) = \left(\frac{1}{\alpha} - \frac{\mu}{2}\right)\|\mathbf{x}^1 - \mathbf{x}\|^2 + \alpha\log(K+1)\left(\sigma^2 + 2\beta^2 F^2 G^2 + \frac{2G^2}{M}\frac{C_3}{1 - \frac{8\alpha\rho G^2}{M}}\right),$$

with C_3 defined in (3.37). *In addition, let $\bar{\mathbf{x}}^K = \frac{\sum_{k=1}^K \mathbf{x}^k}{K}$ and $\bar{\mathbf{z}}^K = \frac{\sum_{k=1}^K \mathbf{z}^k}{K}$. Then*

$$(3.41\text{a}) \qquad \mathbb{E}\big|f_0(\bar{\mathbf{x}}^K) - f_0(\mathbf{x}^*)\big| \leq \frac{1}{K}\left(2\phi_3(\mathbf{x}^*) + \frac{9\log(K+1)}{\rho}\|\mathbf{z}^*\|^2\right),$$

$$(3.41\text{b}) \qquad \mathbb{E}\left[\frac{1}{M}\sum_{j=1}^M [f_j(\bar{\mathbf{x}}^K)]_+\right] \leq \frac{1}{K}\left(\phi_3(\mathbf{x}^*) + \frac{\log(K+1)}{\rho}\|\mathbf{1} + \mathbf{z}^*\|^2\right).$$

Proof. Let $\alpha_k = \frac{\alpha}{k+1} \,\forall\, k \geq 1$. Since $\alpha \geq \frac{1}{\mu}$, it holds that $\frac{k+1}{\alpha} \geq \frac{k+2}{\alpha} - \mu$, i.e., $\frac{1}{\alpha_k} \geq \frac{1}{\alpha_{k+1}} - \mu$. Hence, summing up (3.10) with $\mathbf{x} = \mathbf{x}^*$ from $k = 1$ through K, using Lemma 3.13, and noting $\mathbf{z}^1 = \mathbf{0}$ and the choice of ρ_k yield

$$(3.42)$$

$$\sum_{k=1}^K \mathbb{E}\left[f_0(\mathbf{x}^k) + \frac{1}{M}\sum_{j=1}^M z_j f_j(\mathbf{x}^k)\right] + \frac{1}{2\alpha_K}\mathbb{E}\|\mathbf{x}^{K+1} - \mathbf{x}^*\|^2 + \frac{1}{2\rho_K}\mathbb{E}\|\mathbf{z}^{K+1} - \mathbf{z}\|^2$$

$$\leq \sum_{k=1}^K \mathbb{E}\left[f_0(\mathbf{x}^*) + \Psi_\beta(\mathbf{x}^*, \mathbf{z}^k)\right] + \left(\frac{1}{2\alpha_1} - \frac{\mu}{2}\right)\|\mathbf{x}^1 - \mathbf{x}^*\|^2 + \frac{1}{\rho_1}\mathbb{E}\|\mathbf{z}\|^2$$

$$+ \sum_{k=1}^K \alpha_k\left(\sigma^2 + 2\beta^2 F^2 G^2 + \frac{2G^2}{M}\mathbb{E}\|\mathbf{z}^k\|^2\right) - \sum_{k=1}^K \frac{1}{2\rho_k}\left(\frac{\beta}{\rho_k} - 2\right)\mathbb{E}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2$$

$$\leq Kf_0(\mathbf{x}^*) + \mathbb{E}\left[\phi_3(\mathbf{x}^*) + \frac{1}{\rho_1}\|\mathbf{z}\|^2\right],$$

where in the first inequality, we have used the fact $\mathbb{E}\langle\mathbf{x}^k - \mathbf{x}^*, \mathbf{g}_0^k - \tilde{\nabla} f_0(\mathbf{x}^k)\rangle = 0$ and $\mathbb{E}\langle\mathbf{x}^k - \mathbf{x}^*, \mathbf{h}^k - \nabla_\mathbf{x}\Psi_\beta(\mathbf{x}^k, \mathbf{z}^k)\rangle = 0$, and in the second inequality, we have used (3.36) and (A.6), Lemma 3.1, the setting $\beta \geq 2\rho_k \,\forall\, k$, and also the definition of ϕ_3 in (3.40). Let $\mathbf{z} = \mathbf{z}^*$ in the above inequality. Then by (2.4), we have that

$$\frac{1}{2\alpha_K}\mathbb{E}\|\mathbf{x}^{K+1} - \mathbf{x}^*\|^2 \leq \phi_3(\mathbf{x}^*) + \frac{1}{\rho_1}\|\mathbf{z}^*\|^2,$$

which clearly implies (3.39) by the parameters given in (3.35) and also $\alpha_K = \frac{\alpha}{K+1}$.

Furthermore, dropping the terms about $\|\mathbf{x}^{K+1} - \mathbf{x}^*\|^2$ and $\|\mathbf{z}^{K+1} - \mathbf{z}\|^2$ on the left-hand side of (3.42), and using the convexity of f_j's, we have for any $\mathbf{z} \geq \mathbf{0}$ that

$$\mathbb{E}\left[f_0(\bar{\mathbf{x}}^K) - f_0(\mathbf{x}^*) + \frac{1}{M}\sum_{j=1}^M z_j f_j(\bar{\mathbf{x}}^K)\right] \leq \frac{1}{K}\mathbb{E}\left[\phi_3(\mathbf{x}^*) + \frac{1}{\rho_1}\|\mathbf{z}\|^2\right].$$

Now using Lemma 3.2 and Remark 3.1, we obtain the desired results. $\qquad\square$

Remark 3.4. The order of the established rate is worse than the optimal one obtained for a primal SGM by a $\log(K+1)$ factor. That term appears essentially because of the setting of ρ_k to bound the dual iterate. If we assume $\{\mathbf{z}^k\}$ to be bounded, then we can set $\rho_k = \frac{\beta}{2}$ and remove the logarithmic term. Furthermore, if the maximum number K of iterations is not given, we can set

$$\mathbf{D}_k = \frac{k+1}{\alpha}\mathbf{I},\ \rho_k = \frac{\rho}{\log(k+1)},\ \beta \geq \frac{2\rho}{\log 2}\ \forall k,$$

with $\alpha \geq \frac{1}{\mu}$. These parameters satisfy the conditions in Proposition 3.6, and thus we can still have a sublinear convergence result through first bounding $\mathbb{E}\|\mathbf{z}^k\|^2$. However, there will be an additional $\log(K+1)$ term in the obtained result, i.e., $O\big([\log(K+1)]^2/(K+1)\big)$ for any positive integer K. The result can be shown by following the proofs of Proposition 3.12 and Theorem 3.14. We leave it to the interested readers.

4. Convergence analysis of the adaptive method. In this section, we analyze Algorithm 1 with the adaptive Setting 2 for $\mathbf{D}_k$'s. For simplicity and also due to the page limitation, we only consider the convex case with a predetermined maximum number of iterations. For the convex case with a varying maximum number of iterations and the strongly convex case, we can have results similar to those in section 3.

Similar to the analysis in the previous section, we first bound $\mathbb{E}\|\mathbf{z}^k\|^2$ as follows. Its proof is given in the appendix.

PROPOSITION 4.1. *Assume that X is bounded and also Assumptions* 1 *through* 3 *hold. Given a positive integer K, let $\alpha > 0$ and $\rho > 0$ such that $\alpha\rho < \frac{M}{8G^2}$, and let*

$$(4.1) \qquad \alpha_k = \frac{\alpha}{\sqrt{K}},\ \rho_k = \frac{\rho}{\sqrt{K}},\ \beta \geq \rho\ \forall 1 \leq k \leq K.$$

Suppose that $\{(\mathbf{x}^k, \mathbf{z}^k)\}$ is generated from Algorithm 1 with $\mathbf{D}_k$ set according to Setting 2 and all other parameters specified in (4.1). Then for any $\mathbf{x} \in X$, we have

$$(4.2) \qquad \frac{1}{2}\sum_{k=1}^{K}\big(\|\mathbf{x}^k - \mathbf{x}\|_{\mathbf{D}_k}^2 - \|\mathbf{x}^{k+1} - \mathbf{x}\|_{\mathbf{D}_k}^2\big) \leq \frac{\sqrt{K}}{2\alpha}\|\mathbf{x}^1 - \mathbf{x}\|^2 + \frac{\eta B^2\sqrt{nK}}{2}.$$

In addition, for any $1 \leq k \leq K+1$, it holds that

$$(4.3) \qquad \mathbb{E}\|\mathbf{z}^k\|^2 \leq \frac{C_4}{1 - \frac{8\alpha\rho G^2}{M}}.$$

Here $B = \max_{\mathbf{x}_1,\mathbf{x}_2 \in X}\|\mathbf{x}_1 - \mathbf{x}_2\|_\infty$, and

$$(4.4) \qquad C_4 = \frac{2\rho}{\alpha}\|\mathbf{x}^1 - \mathbf{x}^*\|^2 + 2\rho\eta B^2\sqrt{n} + 4\|\mathbf{z}^*\|^2 + 4\alpha\rho\left(\sigma^2 + 2\beta^2 F^2 G^2\right).$$

By the above proposition, we have the convergence rate estimate of Algorithm 1 with the adaptive Setting 2 about $\mathbf{D}_k$'s.

THEOREM 4.2. *Under Assumptions* 1 *through* 3, *let $\{(\mathbf{x}^k, \mathbf{z}^k)\}$ be generated from Algorithm 1 with $\mathbf{D}_k$ set according to Setting* 2. *Given any positive integer K, set the parameters according to (4.1), let $\bar{\mathbf{x}}^K = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}^k$ and $\bar{\mathbf{z}}^K = \frac{1}{K}\sum_{k=1}^{K}\mathbf{z}^k$, and define*

(4.5)

$$\phi_4(\mathbf{x}) = \frac{3}{2\alpha}\|\mathbf{x}^1 - \mathbf{x}\|^2 + \frac{\eta B^2\sqrt{n}}{2} + \alpha\left(\frac{3}{2}\sigma^2 + 3\beta^2 F^2 G^2 + \frac{3G^2}{M}\frac{C_4}{1 - \frac{8\alpha\rho G^2}{M}} + \frac{F^2}{2}\right),$$

where C_4 is defined in (4.4). If X is bounded, then

$$(4.6a) \qquad \mathbb{E}\big|f_0(\bar{\mathbf{x}}^K) - f_0(\mathbf{x}^*)\big| \leq \frac{1}{\sqrt{K}}\left(2\phi_4(\mathbf{x}^*) + \frac{9(\alpha + \rho)}{2\alpha\rho}\|\mathbf{z}^*\|^2\right),$$

$$(4.6b) \qquad \mathbb{E}\left[\frac{1}{M}\sum_{j=1}^M [f_j(\bar{\mathbf{x}}^K)]_+\right] \leq \frac{1}{\sqrt{K}}\left(\phi_4(\mathbf{x}^*) + \frac{\alpha + \rho}{2\alpha\rho}\|\mathbf{1} + \mathbf{z}^*\|^2\right),$$

$$(4.6c) \qquad \mathbb{E}\big[d_\beta(\mathbf{z}^*) - d_\beta(\bar{\mathbf{z}}^K)\big] \leq \frac{3}{2\sqrt{K}}\left(\max_{\mathbf{x}\in X}\phi_4(\mathbf{x}) + \frac{3(\alpha + \rho)}{2\alpha\rho}\|\mathbf{z}^*\|^2\right).$$

Proof. Multiply $\alpha_k = \frac{\alpha}{\sqrt{K}}$ to (3.10), sum it up from $k = 1$ through K, use (3.26) through (3.28) and also (4.2), and note $\mathbf{z}^1 = \mathbf{0}$. Then we have from (4.3) that

$$\frac{\alpha}{\sqrt{K}}\sum_{k=1}^K \mathbb{E}\left[f_0(\mathbf{x}^k) + \frac{1}{M}\sum_{j=1}^M z_j f_j(\mathbf{x}^k)\right] \leq \frac{\alpha}{\sqrt{K}}\sum_{k=1}^K \mathbb{E}\left[f_0(\mathbf{x}^*) + \Psi_\beta(\mathbf{x}, \mathbf{z}^k)\right] + \frac{\alpha\eta B^2\sqrt{n}}{2}$$

$$+ \frac{3}{2}\mathbb{E}\|\mathbf{x}^1 - \mathbf{x}\|^2 + \left(\frac{\alpha}{2\rho} + \frac{1}{2}\right)\mathbb{E}\|\mathbf{z}\|^2 + \alpha^2\left(\frac{3}{2}\sigma^2 + 3\beta^2 F^2 G^2 + \frac{3G^2}{M}\frac{C_4}{1 - \frac{8\alpha\rho G^2}{M}} + \frac{F^2}{2}\right).$$

Now the desired results can be obtained by following the same arguments as those in the proof of Theorem 3.10. □

Remark 4.1. From the proofs of Proposition 4.1 and Theorem 4.2, we see that the inequality (4.2) is important to bound $\mathbb{E}\|\mathbf{z}^k\|^2$ and to have the convergence rate results. In addition, while proving (4.2), we use the bound $\|\mathbf{s}^k\| = O(\sqrt{k})$. Since we scale the SGs in Setting 2, we automatically have such a bound. Without the scaling process, we may not have it unless we assume the dual variable to be bounded.

5. Numerical experiments. In this section, we test the proposed method (called PDSG) on solving a sample approximation problem of RPS and also three QCQPs. We compare it to the stochastic mirror-prox method in [7] and the CSA method in [8]. The RPS test is performed in MATLAB 2016a installed on a MacBook Pro with 8 GB memory, while the QCQP test is in MATLAB 2018a installed on a Dell workstation with 32 GB memory.

5.1. Sample approximation of robust portfolio selection. Suppose that one investor has a unit of capital to invest on n assets. Assume the return rate of the ith asset follows a uniform distribution on $[\mu_i - \sigma_i, \mu_i + \sigma_i]$ for each $i \in [n]$. The RPS aims to maximize the expected return subject to a minimum return c for all possible return rates, i.e.,

$$(5.1) \qquad \max_{\mathbf{x}\in X}\boldsymbol{\mu}^\top\mathbf{x}, \text{ s.t. } \sum_{i=1}^n \xi_i x_i \geq c, \forall \xi_i \in [\mu_i - \sigma_i, \mu_i + \sigma_i], \forall i \in [n],$$

where $X = \{\mathbf{x} : \mathbf{x} \geq \mathbf{0}, \sum_{i=1}^n x_i = 1\}$. It is easy to see that the above robust constraint is equivalent to $\sum_{i=1}^n (\mu_i - \sigma_i)x_i \geq c$, and thus (5.1) can be equivalently formulated as a linear program with only two linear constraints and also the nonnegativity constraint.

Now suppose that the distribution of the return rate $\boldsymbol{\xi}$ is unknown but its samples are available. Let $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M\}$ be M samples of $\boldsymbol{\xi}$ and $\bar{\boldsymbol{\mu}}$ be the empirical mean. Then we can solve a sample approximation of (5.1), i.e.,
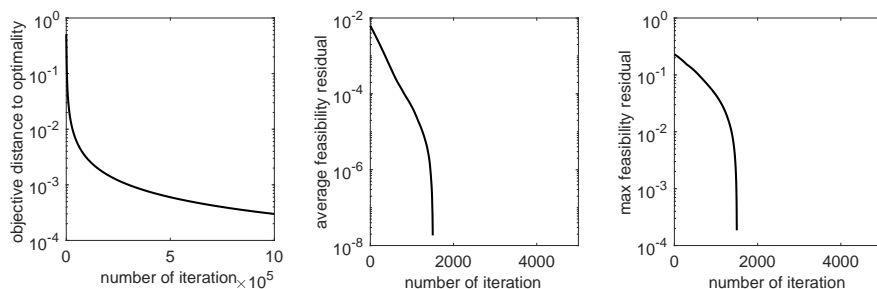
FIG. 1. *Results given by Algorithm* 1 *with nonadaptive setting for solving an instance of the sample approximation* (5.2) *of the RPS. Left: the distance of objective value at averaged point to optimal value* $|f_0(\bar{\mathbf{x}}^k) - f_0(\mathbf{x}^*)|$. *Middle: the average constraint violation at averaged point* $\frac{1}{M}\sum_{j=1}^M [f_j(\bar{\mathbf{x}}^k)]_+$. *Right: the maximum constraint violation at averaged point* $\max_{j\in[M]}[f_j(\bar{\mathbf{x}}^k)]_+$.

$$(5.2) \qquad \max_{\mathbf{x}\in X} \bar{\boldsymbol{\mu}}^\top \mathbf{x}, \text{ s.t. } \boldsymbol{\xi}_j^\top \mathbf{x} \geq c, \forall j \in [M].$$

The sample approximation problem is still a linear program, and one can apply any linear program solver. We use the proposed method in this test simply to see if it can numerically perform well. We set $n = 10$ and $M = 10^4$. All entries of $\bar{\boldsymbol{\mu}}$ are generated independently following the uniform distribution on $[1, 2]$. For each $j \in [M]$, we set $\boldsymbol{\xi}_j = \bar{\boldsymbol{\mu}} + \boldsymbol{\zeta}_j$ with $\boldsymbol{\zeta}_j$ generated by uniform distribution on $[-0.5, 0.5]^n$. Then we let $c = 0.9 \min_{j\in[M], \mathbf{x}\in X} \boldsymbol{\xi}_j^\top \mathbf{x}$ to ensure that (5.2) has a strict feasible solution. The parameters of our algorithm are set according to (3.20) with $K = 10^6$, and $\alpha = \rho = \beta = 1$. The initial point is randomly generated. Figure 1 shows the distance of objective value to the optimal value, the averaged constraint violation, and also the maximum constraint violation, where the optimal objective value is obtained by the MATLAB built-in function `linprog`. The feasibility curves only show the first 5,000 iterations, after which the points remain feasible. We also test the mirror-prox method [7] and the CSA method [8] and find that they perform almost the same as our method on this simple example.

5.2. Quadratically constrained quadratic program. In this subsection, we test the proposed method on a finite-sum structured quadratic program with many quadratic constraints, i.e.,

$$(5.3) \qquad \min_{\mathbf{x}\in X} \frac{1}{2N} \sum_{i=1}^N \|\mathbf{H}_i\mathbf{x} - \mathbf{c}_i\|^2, \text{ s.t. } \frac{1}{2}\mathbf{x}^\top \mathbf{Q}_j \mathbf{x} + \mathbf{a}_j^\top \mathbf{x} \leq b_j, j = 1, \ldots, M.$$

Here $X = [-10, 10]^n$; for each $i \in [N]$, $\mathbf{H}_i \in \mathbb{R}^{p\times n}$ and $\mathbf{c}_i$ are randomly generated with components independently following standard Gaussian distribution; the entries of every $\mathbf{a}_j$ also follow standard Gaussian distribution; $\mathbf{Q}_j$'s are randomly generated symmetric positive semidefinite matrices; each b_j is generated according to uniform distribution on $[0.1, 1.1]$. Note that for the generated data, the Slater's condition holds, and thus there must exist a KKT point for (5.3). Let ξ be a random variable with uniform distribution on $[N]$. Then the objective of (5.3) can be written to $\mathbb{E}_\xi \frac{1}{2}\|\mathbf{H}_\xi \mathbf{x} - \mathbf{c}_\xi\|^2$, and thus (5.3) is in the form of (1.1).

In the experiment, we test three QCQP instances of different sizes. For all of them, we set $N = M = 10^4$ in (5.3), and the dimension (n, p) is set to $(10, 5)$, $(200, 150)$, and $(400, 350)$ respectively for the three instances. We test the proposed algorithm with both nonadaptive and adaptive settings. For the nonadaptive one, we set algorithm parameters according to (3.20) with $K = 50,000$, $\alpha = \rho = \sqrt{10}$, and $\beta = 1$, and it is

called `PDSG-nonadp`. For the adaptive method, i.e., $\mathbf{D}_k$ given according to Setting 2, we set $\eta = \frac{1}{\sqrt{10}}$ and the other parameters according to (4.1) with $K = 50{,}000$, $\alpha = 10$, $\rho = \sqrt{10}$, and $\beta = 1$, and we call it `PDSG-adp`. The stochastic mirror-prox method [7] with update given in (1.15) is applied on the equivalent saddle-point problem (1.14). Although the mirror-prox method requires a compact Z, we simply set $Z = \mathbb{R}^M$, and the method still works well in this test. We use the same penalty parameter $\beta = 1$ and the same step size α_k as for our nonadaptive method. Also we apply the CSA method [8] with update given in (1.17). The same step size α_k is used, and η_k is set to $1/\sqrt{K}$ *for all k*. For all the tested methods, at each iteration, we sample 10 component functions in the objective and also 10 constraint functions to obtain an unbiased SG, i.e., a minibatch of size 10 is applied. Projecting onto the set $\{\mathbf{x} : \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} + \mathbf{a}^\top \mathbf{x} \leq b\}$ does not generally admit an analytic solution and requires an iterative method. Hence, the methods in [21, 24] with updates (1.20) and (1.22) could be inefficient for solving the QCQP problem and are not compared.

Figure 2 shows the results for each method on the three QCQP instances, the objective error, average constraint violation, and maximum constraint violation with
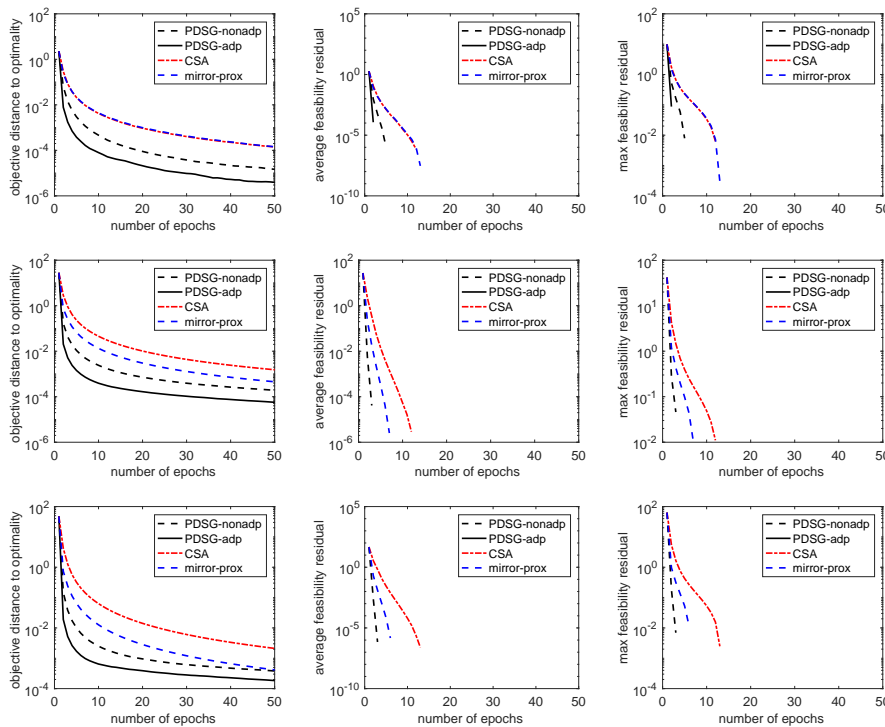

FIG. 2. *Results given by Algorithm* 1 *with both nonadaptive and adaptive settings (called* `PDSG-nonadp` *and* `PDSG-adp`*), the stochastic mirror-prox method in* [7]*, and the CSA method in* [8] *for solving three instances of the* (5.3)*, each instance with $N = M = 10{,}000$. Left: the distance of objective value at averaged point to optimal value $|f_0(\bar{\mathbf{x}}^k) - f_0(\mathbf{x}^*)|$. Middle: the average constraint violation at averaged point $\frac{1}{M}\sum_{j=1}^{M}[f_j(\bar{\mathbf{x}}^k)]_+$. Right: the maximum constraint violation at averaged point $\max_{j\in[M]}[f_j(\bar{\mathbf{x}}^k)]_+$. First row: dimension $n = 10, p = 5$; second row: dimension $n = 200, p = 150$; last row: dimension $n = 400, p = 350$. Missing parts of the feasibility curves indicate that feasible solutions are obtained. In the second and third rows, the solutions by* `PDSG-adp` *remain feasible after the first epoch.*

Running time (in seconds) for each compared method on three instances tested in Figure 2.
Note that the very long time for the third row is because the data size almost reaches the limit of the machine memory.

Method / Dimension	PDSG-nonadp	PDSG-adp	CSA	mirror-prox
$n = 10, p = 5$	20.20	20.69	20.56	20.32
$n = 200, p = 150$	248.94	239.56	250.01	244.82
$n = 400, p = 350$	20129.59	20044.41	20118.03	20161.85

respect to epoch, where the "optimal" solution is computed by running `PDSG-adp` to 1,000 epochs for the smallest instance and 500 epochs for another two. Table 1 shows the running time (in seconds) of each method. Since all the tested methods have almost the same per-iteration complexity, their total running times are almost the same. The very long time for the largest instance is because the data size in this instance almost reaches the limit of machine memory. From the results, we see that the proposed algorithm performs significantly better than the stochastic mirror-prox and CSA methods. In addition, the adaptive PDSG is significantly better than the nonadaptive one. Note that we scale the SGs in the adaptive PDSG. Hence, with the parameters we set, the two PDSGs use roughly the same step size in this experiment. Therefore, the better performance of the adaptive method is mainly attributed to its different setting of $\mathbf{D}_k$.

6. Conclusions. We have proposed a primal-dual (adaptive) stochastic gradient method for stochastic programming with many functional constraints. In every iteration, the method only needs a stochastic subgradient of the objective and a subgradient and the function value of one randomly sampled constraint function. Under standard assumptions, we have established its convergence rate for both convex and strongly convex problems. The order of rate is optimal for the convex case and nearly optimal for the strongly convex case. Numerical experiments on a sample approximation problem of the robust portfolio selection and quadratically constrained quadratic programming demonstrate its nice practical performance.

Appendix A. Proofs of Propositions.

A.1. Proof of Proposition 3.6. Let $(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^*, \mathbf{z}^*)$ in (3.10). Then the last three expectation terms vanish. Since $\rho_k \leq \beta$, we have by the definition of Φ in (1.24) and Lemma 3.1 that

$$\mathbb{E}\left[\Phi(\mathbf{x}^k; \mathbf{x}^*, \mathbf{z}^*)\right] + \frac{1}{2}\mathbb{E}\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{\mathbf{D}_k}^2 + \frac{1}{2\rho_k}\mathbb{E}\|\mathbf{z}^{k+1} - \mathbf{z}^*\|^2$$
$$\leq \frac{1}{2}\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^*\|_{\mathbf{D}_k - \mu\mathbf{I}}^2 + \frac{1}{2\rho_k}\mathbb{E}\|\mathbf{z}^k - \mathbf{z}^*\|^2 + \alpha_k\left(\sigma^2 + 2\beta^2 F^2 G^2 + \frac{2G^2}{M}\mathbb{E}\|\mathbf{z}^k\|^2\right).$$

Multiplying $2\rho_k$ to both sides of the above inequality gives

$$2\rho_k\mathbb{E}\left[\Phi(\mathbf{x}^k; \mathbf{x}^*, \mathbf{z}^*)\right] + \rho_k\mathbb{E}\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{\mathbf{D}_k}^2 + \mathbb{E}\|\mathbf{z}^{k+1} - \mathbf{z}^*\|^2$$
$$\leq \rho_k\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^*\|_{\mathbf{D}_k - \mu\mathbf{I}}^2 + \mathbb{E}\|\mathbf{z}^k - \mathbf{z}^*\|^2 + 2\alpha_k\rho_k\left(\sigma^2 + 2\beta^2 F^2 G^2 + \frac{2G^2}{M}\mathbb{E}\|\mathbf{z}^k\|^2\right).$$

Summing the above inequality from $k = 1$ through t, we have by $\mathbf{z}^1 = \mathbf{0}$, noting $\Phi(\mathbf{x}^k; \mathbf{x}^*, \mathbf{z}^*) \geq 0 \,\forall k$ from (2.4), and using the condition in (3.18) that

$$\mathbb{E}\|\mathbf{z}^{t+1} - \mathbf{z}^*\|^2 \leq \rho_1\|\mathbf{x}^1 - \mathbf{x}^*\|_{\mathbf{D}_1 - \mu\mathbf{I}}^2 + \|\mathbf{z}^*\|^2 + \sum_{k=1}^t 2\alpha_k\rho_k\left(\sigma^2 + 2\beta^2 F^2 G^2 + \frac{2G^2}{M}\mathbb{E}\|\mathbf{z}^k\|^2\right).$$

From the Young's inequality, it follows that $\|\mathbf{z}^{t+1}\|^2 \leq 2\|\mathbf{z}^{t+1} - \mathbf{z}^*\|^2 + 2\|\mathbf{z}^*\|^2$, which together with the above inequality gives the desired result.

A.2. Proof of Proposition 3.7. Let $\alpha_k = \frac{\alpha}{\sqrt{K}}$ $\forall 1 \leq k \leq K$. It is easy to see that the parameters given in (3.20) satisfy the conditions in Proposition 3.6. Hence, for any $t \leq K$, it follows from (3.19) that

(A.1)

$$\mathbb{E}\|\mathbf{z}^{t+1}\|^2 \leq \frac{2\rho}{\alpha}\|\mathbf{x}^1 - \mathbf{x}^*\|^2 + 4\|\mathbf{z}^*\|^2 + 4\alpha\rho\big(\sigma^2 + 2\beta^2 F^2 G^2\big) + \frac{8\alpha\rho G^2}{MK}\sum_{k=1}^t \mathbb{E}\|\mathbf{z}^k\|^2.$$

Now we show the result in (3.21) by induction. Since $\mathbf{z}^1 = \mathbf{0}$, (3.21) holds trivially for $k = 1$. Assume it holds for $k \leq t$. Then from (A.1), it follows that

$$\mathbb{E}\|\mathbf{z}^{t+1}\|^2 \leq C_1 + \frac{8\alpha\rho G^2}{MK}\sum_{k=1}^t \frac{C_1}{1 - \frac{8\alpha\rho G^2}{M}} \leq \frac{C_1}{1 - \frac{8\alpha\rho G^2}{M}},$$

which completes the proof.

A.3. Proof of Proposition 3.8. Let $\alpha_k = \frac{\alpha}{\sqrt{k+1}\log(k+1)}\forall k \geq 1$. It is easy to see that the parameters given in (3.23) satisfy the conditions in Proposition 3.6. Hence, plugging the specified parameters into (3.19) gives

$$\mathbb{E}\|\mathbf{z}^{t+1}\|^2 \leq \frac{2\rho}{\alpha}\|\mathbf{x}^1 - \mathbf{x}^*\|^2 + 4\|\mathbf{z}^*\|^2$$

(A.2)

$$+ \sum_{k=1}^t \frac{4\alpha\rho}{(k+1)(\log(k+1))^2}\left(\sigma^2 + 2\beta^2 F^2 G^2 + \frac{2G^2}{M}\mathbb{E}\|\mathbf{z}^k\|^2\right).$$

By

$$\sum_{k=1}^\infty \frac{1}{(k+1)(\log(k+1))^2} \leq \frac{1}{2(\log 2)^2} + \int_1^\infty \frac{1}{(x+1)(\log(x+1))^2}dx$$

(A.3)

$$= \frac{1}{2(\log 2)^2} + \frac{1}{\log 2} \leq 2.5,$$

we have from (A.2) that

$$\mathbb{E}\|\mathbf{z}^{t+1}\|^2 \leq \frac{2\rho}{\alpha}\|\mathbf{x}^1 - \mathbf{x}^*\|^2 + 4\|\mathbf{z}^*\|^2 + 10\alpha\rho\big(\sigma^2 + 2\beta^2 F^2 G^2\big)$$

(A.4)

$$+ \sum_{k=1}^t \frac{8\alpha\rho}{(k+1)(\log(k+1))^2}\frac{G^2}{M}\mathbb{E}\|\mathbf{z}^k\|^2.$$

Now we show the result in (3.24) by induction. When $k = 1$, it obviously holds. Assume the result holds for $k \leq t$. Then from (A.4), it follows that

$$\mathbb{E}\|\mathbf{z}^{t+1}\|^2 \leq C_2 + \sum_{k=1}^t \frac{8\alpha\rho}{(k+1)(\log(k+1))^2}\frac{G^2}{M}\frac{C_2}{1 - \frac{20\alpha\rho G^2}{M}}$$

$$\leq C_2 + \frac{20\alpha\rho G^2}{M}\frac{C_2}{1 - \frac{20\alpha\rho G^2}{M}} = \frac{C_2}{1 - \frac{20\alpha\rho G^2}{M}},$$

where the second inequality uses (A.3). This completes the proof.

A.4. Proof of Proposition 3.12. Let $\alpha_k = \frac{\alpha}{k+1} \, \forall \, k \geq 1$. If $\alpha \geq \frac{1}{\mu}$, then $\frac{k+1}{\alpha} \geq \frac{k+2}{\alpha} - \mu$, i.e., $\frac{1}{\alpha_k} \geq \frac{1}{\alpha_{k+1}} - \mu$. Hence, the parameters given in (3.35) satisfy the condition in Proposition 3.6, thus (3.19) holds and, with the specified parameters, becomes

$$(A.5) \qquad \mathbb{E}\|\mathbf{z}^{t+1}\|^2 \leq \frac{2\rho}{\log(K+1)} \left(\frac{2}{\alpha} - \mu \right) \|\mathbf{x}^1 - \mathbf{x}^*\|^2 + 4\|\mathbf{z}^*\|^2$$
$$+ \sum_{k=1}^{t} \frac{4\alpha\rho}{(k+1)\log(K+1)} \left(\sigma^2 + 2\beta^2 F^2 G^2 + \frac{2G^2}{M} \mathbb{E}\|\mathbf{z}^k\|^2 \right).$$

Note that for any $t \leq K$,

$$(A.6) \qquad \sum_{k=1}^{t} \frac{1}{k+1} \leq \int_1^{t+1} \frac{1}{x} dx = \log(t+1) \leq \log(K+1).$$

Hence, (A.5) implies

$$(A.7) \qquad \mathbb{E}\|\mathbf{z}^{t+1}\|^2 \leq C_3 + \sum_{k=1}^{t} \frac{4\alpha\rho}{(k+1)\log(K+1)} \frac{2G^2}{M} \mathbb{E}\|\mathbf{z}^k\|^2.$$

Now we show (3.36) by induction. When $k = 1$, it obviously holds since $\mathbf{z}^1 = \mathbf{0}$. Assume (3.36) holds for any $k \leq t \leq K$. Then, from (A.6) and (A.7), it follows that

$$\mathbb{E}\|\mathbf{z}^{t+1}\|^2 \leq C_3 + \frac{8\alpha\rho G^2}{M} \frac{C_3}{1 - \frac{8\alpha\rho G^2}{M}} = \frac{C_3}{1 - \frac{8\alpha\rho G^2}{M}},$$

which completes the proof.

A.5. Proof of Proposition 4.1. We first prove (4.2). Since $\mathbf{D}_k = \mathrm{diag}(\mathbf{s}^k) + \frac{\mathbf{I}}{\alpha_k}$ and $\alpha_k = \frac{\alpha}{\sqrt{K}} \forall k$, we have for any $1 \leq t \leq K$ that

$$\sum_{k=1}^{t} \left(\|\mathbf{x}^k - \mathbf{x}\|_{\mathbf{D}_k}^2 - \|\mathbf{x}^{k+1} - \mathbf{x}\|_{\mathbf{D}_k}^2 \right)$$
$$\leq \|\mathbf{x}^1 - \mathbf{x}\|_{\mathbf{D}_1}^2 + \sum_{k=1}^{t-1} \left\langle \mathbf{x}^{k+1} - \mathbf{x}, (\mathbf{s}^{k+1} - \mathbf{s}^k) \odot (\mathbf{x}^{k+1} - \mathbf{x}) \right\rangle$$
$$\leq \frac{1}{\alpha_1} \|\mathbf{x}^1 - \mathbf{x}\|^2 + B^2 \|\mathbf{s}^1\|_1 + B^2 \sum_{k=1}^{t-1} \|\mathbf{s}^{k+1} - \mathbf{s}^k\|_1$$
$$(A.8) \qquad = \frac{1}{\alpha_1} \|\mathbf{x}^1 - \mathbf{x}\|^2 + B^2 \|\mathbf{s}^t\|_1,$$

where $B = \max_{\mathbf{x}_1, \mathbf{x}_2 \in X} \|\mathbf{x}_1 - \mathbf{x}_2\|_\infty$, and we have used the fact $s_i^k \geq 0$ and $s_i^{k+1} \geq s_i^k$ *for all* i and k to have the last equality. By the Cauchy–Schwarz inequality $\langle \mathbf{s}^t, \mathbf{1} \rangle \leq \|\mathbf{1}\| \cdot \|\mathbf{s}^t\| = \sqrt{n}\|\mathbf{s}^t\|$ and also noting $\|\mathbf{s}^t\| \leq \eta\sqrt{t}$ due to the scaling in Setting 2, we have from (A.8) that

$$(A.9) \qquad \sum_{k=1}^{t} \left(\|\mathbf{x}^k - \mathbf{x}\|_{\mathbf{D}_k}^2 - \|\mathbf{x}^{k+1} - \mathbf{x}\|_{\mathbf{D}_k}^2 \right) \leq \frac{1}{\alpha_1} \|\mathbf{x}^1 - \mathbf{x}\|^2 + \eta B^2 \sqrt{nt}.$$

Hence, (4.2) holds.

Now let $(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^*, \mathbf{z}^*)$ in (3.10) and sum it up from $k = 1$ through $t \leq K$. Note that the last three expectation terms in (3.10) vanish when $(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^*, \mathbf{z}^*)$. Then by (2.4) and Lemma 3.1, and also since $\beta \geq \rho_k = \frac{\rho}{\sqrt{K}} \forall k$, we have

$$\frac{1}{2} \sum_{k=1}^{t} \mathbb{E} \|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{\mathbf{D}_k}^2 + \frac{\sqrt{K}}{2\rho} \mathbb{E} \|\mathbf{z}^{t+1} - \mathbf{z}^*\|^2$$

$$\leq \frac{1}{2} \sum_{k=1}^{t} \mathbb{E} \|\mathbf{x}^k - \mathbf{x}^*\|_{\mathbf{D}_k}^2 + \frac{\sqrt{K}}{2\rho} \|\mathbf{z}^1 - \mathbf{z}^*\|^2 + \sum_{k=1}^{t} \alpha_k \left(\sigma^2 + 2\beta^2 F^2 G^2 + \frac{2G^2}{M} \mathbb{E} \|\mathbf{z}^k\|^2 \right),$$

which together with (A.9) by letting $\mathbf{x} = \mathbf{x}^*$ implies

$$\frac{\sqrt{K}}{2\rho} \mathbb{E} \|\mathbf{z}^{t+1} - \mathbf{z}^*\|^2 \leq \frac{1}{2\alpha_1} \|\mathbf{x}^1 - \mathbf{x}^*\|^2 + \frac{\eta B^2 \sqrt{nt}}{2} + \frac{\sqrt{K}}{2\rho} \|\mathbf{z}^1 - \mathbf{z}^*\|^2$$

$$+ \sum_{k=1}^{t} \alpha_k \left(\sigma^2 + 2\beta^2 F^2 G^2 + \frac{2G^2}{M} \mathbb{E} \|\mathbf{z}^k\|^2 \right).$$

Since $\alpha_k = \frac{\alpha}{\sqrt{K}} \forall k$ and $\mathbf{z}^1 = \mathbf{0}$, multiplying $\frac{2\rho}{\sqrt{K}}$ to the above inequality and noting $t \leq K$ gives

$$\mathbb{E} \|\mathbf{z}^{t+1} - \mathbf{z}^*\|^2 \leq \frac{\rho}{\alpha} \|\mathbf{x}^1 - \mathbf{x}^*\|^2 + \rho \eta B^2 \sqrt{n} + \|\mathbf{z}^*\|^2$$

$$+ 2\alpha\rho \left(\sigma^2 + 2\beta^2 F^2 G^2 + \frac{2G^2}{MK} \sum_{k=1}^{t} \mathbb{E} \|\mathbf{z}^k\|^2 \right).$$

Hence, by the Young's inequality $\|\mathbf{z}^{t+1}\|^2 \leq 2\|\mathbf{z}^{t+1} - \mathbf{z}^*\|^2 + 2\|\mathbf{z}^*\|^2$, we have from the above inequality that

$$\mathbb{E} \|\mathbf{z}^{t+1}\|^2 \leq \frac{2\rho}{\alpha} \|\mathbf{x}^1 - \mathbf{x}^*\|^2 + 2\rho \eta B^2 \sqrt{n} + 4\|\mathbf{z}^*\|^2$$

$$+ 4\alpha\rho \left(\sigma^2 + 2\beta^2 F^2 G^2 + \frac{2G^2}{MK} \sum_{k=1}^{t} \mathbb{E} \|\mathbf{z}^k\|^2 \right).$$

Then following the same arguments as those in the end of the proof of Proposition 3.7, we can show the results in (4.3).

Appendix B. Proofs of a few lemmas.

B.1. Proof of Lemma 3.3. Note $\nabla_{z_j} \psi_\beta(f_j(\mathbf{x}), z_j) = \max(-\frac{z_j}{\beta}, f_j(\mathbf{x}))$. Then the update of $\mathbf{z}$ can be written in the compact form

(B.1) $$\mathbf{z}^{k+1} = \mathbf{z}^k + M\rho_k \mathbf{e}_{j_k} \odot \nabla_{\mathbf{z}} \Psi(\mathbf{x}^k, \mathbf{z}^k),$$

where $\odot$ denotes componentwise product. Hence,

(B.2) $$\frac{1}{\rho_k} \langle \mathbf{z}^k - \mathbf{z}, \mathbf{z}^{k+1} - \mathbf{z}^k \rangle = \langle \mathbf{z}^k - \mathbf{z}, \nabla_{\mathbf{z}} \Psi(\mathbf{x}^k, \mathbf{z}^k) \rangle$$

$$+ \langle \mathbf{z}^k - \mathbf{z}, M\mathbf{e}_{j_k} \odot \nabla_{\mathbf{z}} \Psi(\mathbf{x}^k, \mathbf{z}^k) - \nabla_{\mathbf{z}} \Psi(\mathbf{x}^k, \mathbf{z}^k) \rangle.$$

Let

$$J_+^k = \left\{ j \in [M] : \beta f_j(\mathbf{x}^k) + z_j^k \geq 0 \right\}, \quad J_-^k = [M] \setminus J_+^k.$$

Note that for $\mathbf{z} \geq \mathbf{0}$ and any $j \in J^k_-$, it holds that $z_j\big(f_j(\mathbf{x}^k) + \frac{z^k_j}{\beta}\big) \leq 0$. Then from the definition of Ψ_β in (1.2), one can directly verify that

$$- \Psi_\beta(\mathbf{x}^k, \mathbf{z}^k) + \frac{1}{M}\sum_{j=1}^{M} z_j f_j(\mathbf{x}^k) + \big\langle \mathbf{z}^k - \mathbf{z}, \nabla_\mathbf{z}\Psi(\mathbf{x}^k, \mathbf{z}^k)\big\rangle$$

$$= -\frac{1}{M}\sum_{j \in J^k_+} \frac{\beta}{2}\big[f_j(\mathbf{x}^k)\big]^2 - \frac{1}{M}\sum_{j \in J^k_-}\left[\frac{(z^k_j)^2}{2\beta} - z_j\big(f_j(\mathbf{x}^k) + \frac{z^k_j}{\beta}\big)\right]$$

(B.3)
$$\leq -\frac{1}{M}\sum_{j \in J^k_+} \frac{\beta}{2}\big[f_j(\mathbf{x}^k)\big]^2 - \frac{1}{M}\sum_{j \in J^k_-} \frac{(z^k_j)^2}{2\beta}.$$

In addition, note

$$-\frac{1}{M}\sum_{j \in J^k_+} \frac{\beta}{2}\big[f_j(\mathbf{x}^k)\big]^2 - \frac{1}{M}\sum_{j \in J^k_-} \frac{(z^k_j)^2}{2\beta} = -\frac{\beta}{2\rho^2_k}\mathbb{E}\big[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 \,\big|\, \mathcal{H}^k\big].$$

Hence, we have the desired result by adding (B.2) to (B.3) and using

$$\big\langle \mathbf{z}^k - \mathbf{z}, \mathbf{z}^{k+1} - \mathbf{z}^k\big\rangle = \frac{1}{2}\big[\|\mathbf{z}^{k+1} - \mathbf{z}\|^2 - \|\mathbf{z}^k - \mathbf{z}\|^2 - \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2\big].$$

B.2. Proof of Lemma 3.4. For any $j \in [M]$, we have for some $\tilde{\nabla}f_j(\mathbf{x}) \in \partial f_j(\mathbf{x})$ that

$$\tilde{\nabla}_\mathbf{x}\psi_\beta(f_j(\mathbf{x}), z_j) = [\beta f_j(\mathbf{x}) + z_j]_+\tilde{\nabla}f_j(\mathbf{x}).$$

From Assumption 2, note that $\|\tilde{\nabla}f_j(\mathbf{x})\| \leq G$ and $[\beta f_j(\mathbf{x}) + z_j]^2_+ \leq 2\beta^2 F^2 + 2(z_j)^2$. Hence,

$$\|\tilde{\nabla}_\mathbf{x}\psi_\beta(f_j(\mathbf{x}), z_j)\|^2 \leq [\beta f_j(\mathbf{x}) + z_j]^2_+ \|\tilde{\nabla}f_j(\mathbf{x})\|^2 \leq 2G^2(\beta^2 F^2 + (z_j)^2),$$

which implies the desired result.

B.3. Proof of Lemma 3.9. First note that $\mathbb{E}\big[\mathbf{g}^k_0 - \tilde{\nabla}f_0(\mathbf{x}^k)\,\big|\,\mathcal{H}_k\big] = \mathbf{0}$. Hence, if $\mathbf{x}$ is deterministic, then the result in (3.26) trivially holds, and similarly if $(\mathbf{x}, \mathbf{z})$ is deterministic, then the results in (3.27) and (3.28) hold. Next, we prove the results for the stochastic case.

Let $\tilde{\mathbf{x}}^1 = \mathbf{x}^1$ and $\tilde{\mathbf{x}}^{k+1} = \tilde{\mathbf{x}}^k + \alpha_k(\mathbf{g}^k_0 - \tilde{\nabla}f_0(\mathbf{x}^k))$ for $1 \leq k \leq K$. Then $\mathbb{E}\big[\langle\mathbf{x}^k - \tilde{\mathbf{x}}^k, \mathbf{g}^k_0 - \tilde{\nabla}f_0(\mathbf{x}^k)\rangle\,\big|\,\mathcal{H}_k\big] = 0$. Hence,

(B.4)
$$-\sum_{k=1}^{K} \alpha_k\mathbb{E}\left\langle\mathbf{x}^k - \mathbf{x}, \mathbf{g}^k_0 - \tilde{\nabla}f_0(\mathbf{x}^k)\right\rangle = -\sum_{k=1}^{K}\alpha_k\mathbb{E}\left\langle\tilde{\mathbf{x}}^k - \mathbf{x}, \mathbf{g}^k_0 - \tilde{\nabla}f_0(\mathbf{x}^k)\right\rangle.$$

In addition, by the definition of $\{\tilde{\mathbf{x}}^k\}$, we have

$$-\sum_{k=1}^{K}\alpha_k\left\langle\tilde{\mathbf{x}}^k - \mathbf{x}, \mathbf{g}^k_0 - \tilde{\nabla}f_0\left(\mathbf{x}^k\right)\right\rangle = \sum_{k=1}^{K}\left\langle\tilde{\mathbf{x}}^k - \mathbf{x}, \tilde{\mathbf{x}}^k - \tilde{\mathbf{x}}^{k+1}\right\rangle$$

$$= \frac{1}{2}\left[\|\tilde{\mathbf{x}}^1 - \mathbf{x}\|^2 - \|\tilde{\mathbf{x}}^{K+1} - \mathbf{x}\|^2 + \sum_{k=1}^{K}\|\tilde{\mathbf{x}}^k - \tilde{\mathbf{x}}^{k+1}\|^2\right]$$

$$\leq \frac{1}{2}\left[\|\mathbf{x}^1 - \mathbf{x}\|^2 + \sum_{k=1}^{K}\alpha^2_k\|\mathbf{g}^k_0 - \tilde{\nabla}f_0(\mathbf{x}^k)\|^2\right],$$

where we have used the fact $\tilde{\mathbf{x}}^1 = \mathbf{x}^1$. Substituting the above inequality into (B.4) gives

(B.5)
$$-\sum_{k=1}^{K} \alpha_k \mathbb{E}\left[\left\langle \mathbf{x}^k - \mathbf{x}, \mathbf{g}_0^k - \tilde{\nabla} f_0\left(\mathbf{x}^k\right)\right\rangle\right] \leq \frac{1}{2}\mathbb{E}\left[\|\mathbf{x}^1 - \mathbf{x}\|^2 + \sum_{k=1}^{K} \alpha_k^2 \|\mathbf{g}_0^k - \tilde{\nabla} f_0(\mathbf{x}^k)\|^2\right].$$

By Assumption 2 and the fact $\mathbb{E}\|\xi - \mathbb{E}\xi\|^2 \leq \mathbb{E}\|\xi\|^2$ for any random vector ξ, we have $\mathbb{E}\|\mathbf{g}_0^k - \tilde{\nabla} f_0(\mathbf{x}^k)\|^2 \leq \sigma^2$, and thus (B.5) implies (3.26).

By essentially the same arguments, we can show (3.27) by noting $\mathbb{E}\|\mathbf{h}^k\|^2 \leq 2\beta^2 F^2 G^2 + \frac{2G^2}{M}\mathbb{E}\|\mathbf{z}^k\|^2$ from (3.15), and also we can show (3.28) by noting from Assumption 2 that

$$\mathbb{E}\|M\mathbf{e}_{j_k} \odot \nabla_\mathbf{z}\Psi(\mathbf{x}^k, \mathbf{z}^k)\|^2 = \mathbb{E}\left|\max\left(-\frac{z_{j_k}}{\beta}, f_{j_k}(\mathbf{x}^k)\right)\right|^2 \leq F^2.$$

B.4. Proof of Lemma 3.13. Denote $\Delta_\mathbf{z}^k = M\mathbf{e}_{j_k} \odot \nabla_\mathbf{z}\Psi(\mathbf{x}^k, \mathbf{z}^k) - \nabla_\mathbf{z}\Psi(\mathbf{x}^k, \mathbf{z}^k)$. Let $\tilde{\mathbf{z}}^1 = \mathbf{z}^1$ and $\tilde{\mathbf{z}}^{k+1} = \tilde{\mathbf{z}}^k - \rho_k \Delta_\mathbf{z}^k \, \forall k \geq 1$. Then $\mathbb{E}\langle \mathbf{z}^k - \tilde{\mathbf{z}}^k, \Delta_\mathbf{z}^k \rangle = 0$ for any k. Note that $\rho_k = \frac{\rho}{\log(K+1)} \, \forall k$. Hence,

$$\sum_{k=1}^{K} \mathbb{E}\left\langle \mathbf{z}^k - \mathbf{z}, M\mathbf{e}_{j_k} \odot \nabla_\mathbf{z}\Psi\left(\mathbf{x}^k, \mathbf{z}^k\right) - \nabla_\mathbf{z}\Psi\left(\mathbf{x}^k, \mathbf{z}^k\right)\right\rangle$$

$$= \frac{\log(K+1)}{\rho}\sum_{k=1}^{K} \mathbb{E}\left\langle \tilde{\mathbf{z}}^k - \mathbf{z}, \rho_k \Delta_\mathbf{z}^k\right\rangle$$

$$= \frac{\log(K+1)}{\rho}\sum_{k=1}^{K} \mathbb{E}\left\langle \tilde{\mathbf{z}}^k - \mathbf{z}, \tilde{\mathbf{z}}^k - \tilde{\mathbf{z}}^{k+1}\right\rangle$$

$$= \frac{\log(K+1)}{2\rho}\mathbb{E}\left[\|\tilde{\mathbf{z}}^1 - \mathbf{z}\|^2 - \|\tilde{\mathbf{z}}^{K+1} - \mathbf{z}\|^2 + \sum_{k=1}^{K}\|\tilde{\mathbf{z}}^k - \tilde{\mathbf{z}}^{k+1}\|^2\right]$$

$$= \frac{\log(K+1)}{2\rho}\mathbb{E}\left[\|\mathbf{z}^1 - \mathbf{z}\|^2 - \|\tilde{\mathbf{z}}^{K+1} - \mathbf{z}\|^2 + \sum_{k=1}^{K}\rho_k^2\|\Delta_\mathbf{z}^k\|^2\right],$$

where we have used $\tilde{\mathbf{z}}^1 = \mathbf{z}^1$. Since $\mathbb{E}\|\Delta_\mathbf{z}^k\|^2 \leq \mathbb{E}\|M\mathbf{e}_{j_k} \odot \nabla_\mathbf{z}\Psi(\mathbf{x}^k, \mathbf{z}^k)\|^2 = \frac{1}{\rho_k^2}\mathbb{E}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2$, we have (3.38) from the above inequality.

REFERENCES

[1] M. Baes, M. Bürgisser, and A. Nemirovski, *A randomized mirror-prox method for solving structured large-scale matrix saddle-point problems*, SIAM J. Optim., 23 (2013), pp. 934–962.
[2] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons, New York, 2013.

[3] G. CALAFIORE AND M. C. CAMPI, *Uncertain convex programs: Randomized solutions and confidence levels*, Math. Program., 102 (2005), pp. 25–46.

[4] M. C. CAMPI AND S. GARATTI, *A sampling-and-discarding approach to chance-constrained optimization: Feasibility and optimality*, J. Optim. Theory Appl., 148 (2011), pp. 257–280.

[5] Y. CHEN, G. LAN, AND Y. OUYANG, *Optimal primal-dual methods for a class of saddle point problems*, SIAM J. Optim., 24 (2014), pp. 1779–1814.

[6] A. COTTER, M. GUPTA, AND J. PFEIFER, *A light touch for heavily constrained SGD*, in Proceedings of the Conference on Learning Theory, 2016, pp. 729–771.

[7] A. JUDITSKY, A. NEMIROVSKI, AND C. TAUVEL, *Solving variational inequalities with stochastic mirror-prox algorithm*, Stoch. Syst., 1 (2011), pp. 17–58.

[8] G. LAN AND Z. ZHOU, *Algorithms for stochastic optimization with function or expectation constraints*, Comput. Optim. Appl., to appear.

[9] Q. LIN, S. NADARAJAH, AND N. SOHEILI, *A level-set method for convex optimization with a feasible solution path*, SIAM J. Optim., 28 (2018), pp. 3290–3311.

[10] J. LUEDTKE AND S. AHMED, *A sample approximation approach for optimization with probabilistic constraints*, SIAM J. Optim., 19 (2008), pp. 674–699.

[11] M. MAHDAVI, T. YANG, R. JIN, S. ZHU, AND J. YI, *Stochastic gradient descent with only one projection*, in Advances in Neural Information Processing Systems, 2012, pp. 494–502.

[12] A. NEDIĆ AND A. OZDAGLAR, *Subgradient methods for saddle-point problems*, J. Optim. Theory Appl., 142 (2009), pp. 205–228.

[13] A. NEMIROVSKI, A. JUDITSKY, G. LAN, AND A. SHAPIRO, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609.

[14] A. NEMIROVSKI AND A. SHAPIRO, *Scenario approximations of chance constraints*, in Probabilistic and Randomized Methods for Design Under Uncertainty, Springer, New York, 2006, pp. 3–47.

[15] B. PALANIAPPAN AND F. BACH, *Stochastic variance reduction methods for saddle-point problems*, in Advances in Neural Information Processing Systems, 2016, pp. 1416–1424.

[16] C. J. PANG, *Set intersection problems: Supporting hyperplanes and quadratic programming*, Math. Program., 149 (2015), pp. 329–359.

[17] P. RIGOLLET AND X. TONG, *Neyman-Pearson classification, convexity and stochastic constraints*, J. Mach. Learn. Res., 12 (2011), pp. 2831–2855.

[18] R. T. ROCKAFELLAR, *A dual approach to solving nonlinear programming problems by unconstrained optimization*, Math. Program., 5 (1973), pp. 354–373.

[19] R. T. ROCKAFELLAR, *The multiplier method of hestenes and powell applied to convex programming*, J. Optim. Theory Appl., 12 (1973), pp. 555–562.

[20] R. T. ROCKAFELLAR, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.

[21] E. K. RYU AND W. YIN, *Proximal-proximal-gradient method*, J. Comput. Math., 37 (2019), pp. 778–812.

[22] A. SHAPIRO, D. DENTCHEVA, AND A. RUSZCZYŃSKI, *Lectures on Stochastic Programming: Modeling and Theory*, MOS-SIAM Ser. Optim. 9, SIAM, Philadelphia, 2009.

[23] M. STOŠIĆ, J. XAVIER, AND M. DODIG, *Projection on the intersection of convex sets*, Linear Algebra Appl., 509 (2016), pp. 191–205.

[24] M. WANG AND D. P. BERTSEKAS, *Stochastic first-order methods with random constraint projection*, SIAM J. Optim., 26 (2016), pp. 681–717.

[25] M. WANG, Y. CHEN, J. LIU, AND Y. GU, *Random Multi-Constraint Projection: Stochastic Gradient Methods for Convex Optimization with Many Constraints*, preprint, https://arxiv.org/abs/1511.03760, 2015.

[26] M. WANG, E. X. FANG, AND H. LIU, *Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions*, Math. Program., 161 (2017), pp. 419–449.

[27] Y. XU, *First-Order Methods for Constrained Convex Programming Based on Linearized Augmented Lagrangian Function*, preprint, https://arxiv.org/abs/1711.08020, 2017.

[28] Y. XU, *Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming*, Math. Program., Ser. A, to appear.

[29] A. W. YU, L. HUANG, Q. LIN, R. SALAKHUTDINOV, AND J. CARBONELL, *Block-Normalized Gradient Method: An Empirical Study for Training Deep Neural Network*, preprint, https://arxiv.org/abs/1707.04822, 2017.

[30] H. YU AND M. J. NEELY, *A primal-dual type algorithm with the $O(1/t)$ convergence rate for large scale constrained convex programs*, in Proceedings of the 55th Conference on Decision and Control, IEEE, 2016, pp. 1900–1905.

[31] H. YU AND M. J. NEELY, *A Primal-Dual Parallel Method with $O(1/\epsilon)$ Convergence for Constrained Composite Convex Programs*, preprint, https://arxiv.org/abs/1708.00322, 2017.