Penalized stochastic gradient methods for stochastic convex optimization with expectation constraints

Xiantao Xiao

School of Mathematical Sciences, Dalian University of Technology,
Dalian 116023, PR China

September 12, 2019

Abstract

Stochastic gradient method and its variants are simple yet effective for minimizing an expectation function over a closed convex set. However, none of these methods are applicable to solve stochastic programs with expectation constraints, since the projection onto the feasible set is prohibitive. To deal with the expectation constrained stochastic convex optimization problems, we propose a class of penalized stochastic gradient (PSG) methods in which at each iteration a stochastic gradient step is performed along the stochastic (sub)gradients of the objective function and the expectation constraint function. We prove that the basic PSG method and the mini-batch PSG method both converge almost surely to an optimal solution under mild assumptions. The favorable convergence rates of these methods are established by carefully selecting the stepsizes. We also extend PSG to solve the optimization problem with multiple expectation constraints. The efficiency of the proposed methods is validated through numerical experiments in a variety of practical instances.

Keywords: Stochastic convex optimization, expectation constraints, stochastic approximation, convergence analysis, numerical experiments.

1 Introduction

In this paper we consider the following stochastic optimization problem

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & f(x) := \mathbb{E}[F(x,\xi)] \\ \text{s.t.} & g(x) := \mathbb{E}[G(x,\xi)] \le 0, \\ & x \in \mathcal{C}, \end{array} \tag{1.1}$$

where $\mathcal{C} \subseteq \mathbb{R}^n$ is a nonempty closed convex set, the random vector $\xi : \Omega \to \Xi$ is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, Ξ is a measurable space, $F : \mathcal{C} \times \Xi \to \mathbb{R}$ and $G : \mathcal{C} \times \Xi \to \mathbb{R}$. Problems in this formulation are standard in stochastic programming models [52, 51] and also arise frequently in many other applications such as machine learning [53, 46], finance [49, 12] and operations research [9, 34].

Throughout this paper, we make the following standard assumptions: (i) it is possible to generate independent identically distributed samples $\xi_1, \xi_2, \ldots$ of the random vector ξ; (ii) there is an oracle, which, for given $(x, \xi) \in \mathbb{R}^n \times \Xi$, returns $F'(x,\xi) \in \partial_x F(x,\xi)$ and $G'(x,\xi) \in \partial_x G(x,\xi)$ such that $f'(x) := \mathbb{E}[F'(x,\xi)]$ and $g'(x) := \mathbb{E}[G'(x,\xi)]$ are well defined and $f'(x) \in \partial f(x), g'(x) \in \partial g(x)$ [1]. The aim of this paper is to develop a class of efficient numerical methods for solving problem (1.1) based on Monte Carlo sampling techniques.

For stochastic program (without expectation constraints) $\min_{x\in\mathcal{C}} f(x)$, there has been extensive research on its efficient numerical methods. Most of those methods are based on two fundamental approaches: *sample average approximation* (SAA) approach and *stochastic approximation* (SA) approach.

[1] Under mild conditions, it holds that $\partial f(x) = \mathbb{E}[\partial_x F(x,\xi)]$ and $\partial g(x) = \mathbb{E}[\partial_x G(x,\xi)]$, see [56, Theorem 7.52].

By generating a set of samples $\{\xi^1, \ldots, \xi^N\}$ of the random vector ξ, the SAA approach is to solve the following deterministic approximation problem: $\min_{x\in\mathcal{C}} \frac{1}{N}\sum_{i=1}^N F(x,\xi^i)$. The SA approach, which can be traced back to the work by Robbins and Monro [47], is a gradient descent-type method by using stochastic gradient $F'(x^k,\xi^k)$ at each iteration, and hence also named as *stochastic gradient* (SG) method. It is shown in [38] that a properly modified SA method can significantly outperform SAA for a class of stochastic programs. In recent years, there has been an explosive growth of interest in SA-type methods, see a review paper [5] and references therein.

For solving stochastic program with expectation constraint in the form of (1.1), SAA approach [61, 55, 14, 42] is very common, in which the expectation functions $f(x)$ and $g(x)$, respectively, are replaced with $\frac{1}{N}\sum_{i=1}^N F(x,\xi^i)$ and $\frac{1}{N}\sum_{i=1}^N G(x,\xi^i)$, and then the approximation problem is solved by traditional optimization methods for nonlinear programming. A main drawback of this approach is that when the number of samples is small, the optimal solution of the associated SAA problem maybe highly infeasible and far from optimal with respect to the original problem. Instead, when N is large, the gradient evaluation at each iteration is expensive such that the computational complexity is high. In contrast to SAA, the study of SA is still limited since the aforementioned SA-type methods cannot be trivially modified to deal with the expectation constrained programs. The main reason is that those methods typically require the projection onto the feasible set $\{x \in \mathcal{C} : g(x) \le 0\}$ which may not be possible to compute in the expectation formulation. As a first attempt to develop efficient SA-type methods to solve problem (1.1), Lan and Zhou [26] introduce a cooperative stochastic approximation (CSA) algorithm and establish the expected and high probability convergence rates in terms of both objective values and constraint violations. This CSA algorithm is a stochastic counterpart of Polyak's subgradient method [44] and recently extended by [1] to solve stochastic program with multiple expectation constraints. In [30, 31], a stochastic level-set method, which ensures a feasible solution path with high probability, is proposed and analyzed for solving expectation constrained (or finite-sum constrained) optimization problems. In [66, 63], the authors develop some online algorithms to solve online convex optimization with expectation constraints and analyze the expected and high probability rates of regret and constraint violations. In addition to the above algorithms, it is worth mentioning that there are a considerable amount of research on SA-type algorithms for stochastic program with functional constraints, e.g., [36, 62, 43, 65, 37].

In this paper, we propose a penalized stochastic (sub)gradient (PSG) method and its several extensions for solving expectation constrained stochastic programs. PSG can be roughly viewed as a hybrid of the classical quadratic penalty method [40] for nonlinear programming and the stochastic quasi-gradient method [59] for stochastic composition problem. Unlike the traditional penalty method in which at each iteration an unconstrained minimization problem is required to be solved, in PSG at each iteration only a penalized stochastic gradient step is computed which makes it easy to be implemented. In CSA [26], at each point x^k it performs a stochastic gradient (or stochastic subgradient) step along $F'(x^k,\xi^k)$ if x^k is feasible and along $G'(x^k,\xi^k)$ otherwise. When the initial point is highly infeasible, it may require a large number of iterations along the stochastic gradient $G'(x^k,\xi^k)$ to decrease the constraint function value (meanwhile the objective value may increase), and then switch to the direction $F'(x^k,\xi^k)$ to reduce the objective value. However, to guarantee the convergence of CSA, a key observation is that it requires a diminishing stepsize sequence. Therefore, when k is large, the stepsize is usually too small to reduce the objective value. In contrast, at each iteration in PSG we use the stochastic gradient step along both $F'(x^k,\xi^k)$ and $G'(x^k,\xi^k)$ and then force f and g to decrease simultaneously, see, e.g., Figure 1 for the performances of CSA and PSG. Compared the framework of Algorithm 1 in this paper with [66, Algorithm 1], we can observe that these two algorithms share some similarities. However, the intuition, the targeted problems and the analysis of convergence results in this work are quite different from that of [66]. Moreover, in the aforementioned related work, the almost sure global convergence has not been investigated.

The contributions of this paper exist in the following several aspects. Firstly, we present a basic PSG method for solving convex problem (1.1). At the k-th iteration, by generating two samples ξ^k and η^k, it performs a stochastic subgradient step along $F'(x^k,\xi^k)$ and $[t_{k+1}]_+G'(x^k,\eta^k)$, where t_{k+1} is a stochastic approximation to $g(x^k)$. We prove that, under a global error bound condition, when the three diminishing stepsizes $\{\alpha_k\}, \{\beta_k\}$ and $\{\gamma_k\}$ satisfy certain assumptions, the sequence $\{x^k\}$ is shown to converge almost surely to some optimal solution x^* by using the classical supermartingale convergence

theorem. In addition, we analyze the optimality convergence rates of the averaged iterates by controlling the stepsizes for both convex and restricted strongly convex problems. Secondly, motivated by the efficiency of the mini-batch SG methods in machine learning, we propose a mini-batch PSG method, in which at each iteration a better stochastic gradient is used by taking a mini-batch of samples. If the objective function f is differentiable and its gradient is Lipschitz continuous, we show that the choice of the stepsizes in this mini-batch algorithm is more flexible. We prove that this method converges almost surely to an optimal solution by using a coupled supermartingale convergence argument. By carefully selecting the stepsizes, we are able to get faster convergence rates for both convex and restricted strongly convex problems than basic PSG. The mini-batch PSG method is shown to be effectively applicable in several practical instances arise in machine learning, risk management, chance constrained programs, etc. Thirdly, we extend the mini-batch PSG method to solve stochastic program with multiple expectation constraints. At each iteration, it takes a stochastic gradient step to minimize the objective and to avoid the violations of only a mini-batch of expectation constraints. We show that the almost sure convergence and the expected convergence rates can be established in a similar way.

The rest of this paper is organized as follows. In Section 2, we introduce some properties of the Bregman prox-mapping and two supermartingale convergence theorems. In Section 3, we analyze the basic PSG method and prove its almost sure convergence and expected convergence rates. In Section 4 we present the mini-batch PSG method for problems with smooth objective, and establish its convergence results. We extend the mini-batch PSG method to stochastic optimization with multiple expectation constraints in Section 5. Finally, in Section 6 the preliminary numerical results in several practical applications are presented.

1.1 Notations.

Let $[t]_+ := \max(t, 0)$ for all $t \in \mathbb{R}$. Let us denote $\phi(x) := [g(x)]_+^2$, then the constraint $g(x) \leq 0$ is equivalent to $\phi(x) = 0$. Since $[\cdot]_+^2$ is nondecreasing and convex, we have that ϕ is convex if g is convex. Let X^* be the set of optimal solutions, and f^* be the optimal objective value. We denote the feasible set of problem (1.1) by $\Phi := \{x \in \mathcal{C} : g(x) \leq 0\}$. For a set $\mathcal{C} \in \mathbb{R}^n$, the characteristic function is given by

$$\mathbb{1}_{\mathcal{C}}(x) := \begin{cases} 1, & \text{if } x \in \mathcal{C}, \\ 0, & \text{otherwise.} \end{cases}$$

Let $\|\cdot\|$ be a general norm on $\mathbb{R}^n$ induced by a given inner product $\langle \cdot, \cdot \rangle$ and $\|x\|_* = \sup_{\|y\| \leq 1} \langle y, x \rangle$ be its dual norm. The projection onto a closed convex set $\mathcal{C}$ is defined by $\Pi_{\mathcal{C}}(x) := \arg\min_{u \in \mathcal{C}} \|u - x\|$. Then, the distance of x to $\mathcal{C}$ is given by $\text{dist}(x, \mathcal{C}) := \|x - \Pi_{\mathcal{C}}(x)\|$. The function f is said to be *strongly convex* with modulus $\lambda > 0$, if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\lambda}{2}\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n.$$

For every $a \in \mathbb{R}$, let $\lfloor a \rfloor$ be the floor function that returns the largest integer less than or equal to a, and $\lceil a \rceil$ be the ceil function that returns the least integer greater than or equal to a. Let $|S|$ be the cardinality of a set S. The abbreviation "w.p.1" means "with probability 1", while the abbreviation "i.i.d." means "independent identically distributed".

2 Preliminaries

We say that a function $h : \mathcal{C} \to \mathbb{R}$ is a *distance-generating function* with modulus $\lambda > 0$ with respect to $\|\cdot\|$, if h is continuously differentiable and λ-strongly convex over $\mathcal{C}$ with respect to the norm $\|\cdot\|$. The *Bregman distance* associated with h is defined by

$$V(x, y) := h(y) - [h(x) + \langle \nabla h(x), y - x \rangle].$$

From the definition, it follows that $\nabla_y V(x, y) = \nabla h(y) - \nabla h(x)$ and

$$V(x, y) \geq \frac{\lambda}{2}\|x - y\|^2, \quad \forall x, y \in \mathcal{C}. \tag{2.1}$$

As a simple example, $V(x,y)$ recovers the Euclidean distance $\frac{1}{2}\|x-y\|_2^2$ if we take $h(x)=\frac{1}{2}\|x\|_2^2$.

For a given point $x\in\mathcal{C}$, let us define the following *prox-mapping*,

$$\mathcal{P}_x(g):=\arg\min_{y\in\mathcal{C}}\left\{\langle g,y\rangle+V(x,y)\right\},\quad \forall g\in\mathbb{R}^n.$$

Throughout this paper, we assume that the set $\mathcal{C}$ is simple enough such that this mapping is easy to compute. For $h(x)=\frac{1}{2}\|x\|_2^2$, we have $\mathcal{P}_{x^k}(f'(x^k))=\Pi_{\mathcal{C}}(x^k-f'(x^k))$, which reduces to the traditional projected gradient map.

We introduce some well studied properties of $\mathcal{P}_x(g)$ in the following lemmas. First, it is shown that the prox-mapping is Lipschitz continuous, see also in [16, Lemma 2].

Lemma 2.1. *Given $x\in\mathcal{C}$. For all $g_1,g_2\in\mathbb{R}^n$,*

$$\|\mathcal{P}_x(g_1)-\mathcal{P}_x(g_2)\|\le\frac{1}{\lambda}\|g_1-g_2\|_*.$$

The following lemma is well known, see, e.g., [38, Lemma 2.1].

Lemma 2.2. *Given $x\in\mathcal{C}$ and $g\in\mathbb{R}^n$. Let $x^+=\mathcal{P}_x(g)$. Then, for every $u\in\mathcal{C}$,*

$$V(x^+,u)\le V(x,u)-\langle g,x-u\rangle+\frac{1}{2\lambda}\|g\|_*^2. \tag{2.2}$$

Finally, a relation of the distances of three reference points is presented in Lemma 2.3.

Lemma 2.3. *Given $x\in\mathcal{C}$ and $g\in\mathbb{R}^n$. Let $x^+=\mathcal{P}_x(g)$. Then, for every $u\in\mathcal{C}$,*

$$V(x^+,u)\le V(x,u)-V(x,x^+)-\langle g,x^+-u\rangle. \tag{2.3}$$

Proof. From the definition of $\mathcal{P}_x(g)$, one has

$$\langle g+\nabla h(x^+)-\nabla h(x),u-x^+\rangle\ge 0,$$

and thus by using the following three-points identity [10, Lemma 3.1]

$$V(x,u)-V(x^+,u)-V(x,x^+)=\langle\nabla h(x^+)-\nabla h(x),u-x^+\rangle,$$

we get the claim. □

The following supermartingale convergence theorem of Robbins and Sigmund [48] is a fundamental tool to derive the almost sure convergence of stochastic algorithms.

Theorem 2.4. *Let $\{\zeta_k\},\{u_k\},\{a_k\}$ and $\{b_k\}$ be nonnegative adapted processes with respect to the filtration $\{\mathcal{F}_k\}$ such that $\sum_k a_k<\infty,\sum_k b_k<\infty$ with probability 1, and for all k,*

$$\mathbb{E}[\zeta_{k+1}|\mathcal{F}_k]\le(1+a_k)\zeta_k-u_k+b_k,\quad w.p.1.$$

Then, $\{\zeta_k\}$ converges almost surely to a nonnegative finite random variable and $\sum_k u_k<\infty$ with probability 1.

In [58], a coupled supermartingale convergence theorem is proved, which generalizes the classical Robbins and Sigmund theorem. This theorem is useful to derive the almost sure convergence of two coupled stochastic processes.

Theorem 2.5. *Let $\{\theta_k\},\{\zeta_k\},\{u_k\},\{\bar{u}_k\},\{a_k\},\{b_k\},\{d_k\},\{\mu_k\}$ and $\{\nu_k\}$ be nonnegative adapted processes with respect to the filtration $\{\mathcal{F}_k\}$ such that*

$$\sum_k a_k<\infty,\quad\sum_k b_k<\infty,\quad\sum_k\mu_k<\infty,\quad\sum_k\nu_k<\infty,\quad w.p.1,$$

and for all k,

$$\mathbb{E}[\theta_{k+1}|\mathcal{F}_k]\le(1+a_k)\theta_k-u_k+d_k\zeta_k+\mu_k,$$
$$\mathbb{E}[\zeta_{k+1}|\mathcal{F}_k]\le(1-d_k)\zeta_k-\bar{u}_k+b_k\theta_k+\nu_k,$$

with probability 1. Then, $\{\theta_k\}$ and $\{\zeta_k\}$ converge almost surely to nonnegative finite random variables and

$$\sum_k u_k<\infty,\quad\sum_k\bar{u}_k<\infty,\quad\sum_k d_k\zeta_k<\infty,\quad w.p.1.$$

3 A basic penalized stochastic subgradient method

In this section, we will restrict our attention to the case that f and g in problem (1.1) are both convex but not differentiable. We propose a basic penalized stochastic subgradient algorithm for solving this problem.

3.1 Assumptions, algorithmic framework and auxiliary lemmas

Throughout this section, we impose a set of assumptions on the random functions $F(x,\xi)$ and $G(x,\xi)$ as follows.

Assumption 1. *We assume:*

(A.1) *For almost every $\omega \in \Omega$ the functions $F(\cdot,\xi) \equiv F(\cdot,\xi(\omega))$ and $G(\cdot,\xi) \equiv G(\cdot,\xi(\omega))$ are both convex on $\mathcal{C}$.*

(A.2) *There exist constants $C_f > 0$ and $C_t > 0$ such that for all $x \in \mathcal{C}$,*

$$\mathbb{E}[\|F'(x,\xi)\|_*^2] \le C_f^2, \quad \mathbb{E}[[G(x,\xi)]_+^2] \le C_t^2,$$

where $F'(x,\xi) \in \partial_x F(x,\xi)$.

(A.3) *There exists a measurable function $c_g : \Xi \to \mathbb{R}_+$ with $C_g := \mathbb{E}[c_g(\xi)]$ and $\hat{C}_g := \sqrt{\mathbb{E}[c_g^2(\xi)]} < \infty$, such that for all $x, y \in \mathcal{C}$ and for almost every $\omega \in \Omega$,*

$$|G(x,\xi) - G(y,\xi)| \le c_g(\xi)\|x - y\|.$$

Assumption (A.1) implies that the expectation functions f and g are convex on $\mathcal{C}$. Moreover, we have ϕ is convex on $\mathcal{C}$.

In Assumption (A.2), the first condition $\mathbb{E}[\|F'(x,\xi)\|_*^2] \le C_f^2$ is common in the stochastic optimization literature. Together with Jensen's inequality, it gives that

$$\|f'(x)\|_*^2 \le \mathbb{E}[\|F'(x,\xi)\|_*^2] \le C_f^2, \quad \forall x \in \mathcal{C}.$$

The latter condition in Assumption (A.2) is obviously weaker than the moment bounded condition $\mathbb{E}[G^2(x,\xi)] \le C_t^2$ since $\mathbb{E}[[G(x,\xi)]_+^2] \le \mathbb{E}[G^2(x,\xi)]$. Roughly speaking, the latter condition says that the expectation constraint function is bounded above.

Assumption (A.3) indicates that g is Lipschitz continuous with parameter C_g, $\|g'(x)\|_* \le C_g$ and $\mathbb{E}[\|G'(x,\xi)\|_*^2] \le \hat{C}_g^2$. From Jensen's inequality, it follows that $C_g \le \hat{C}_g$. For every $x^* \in X^*$, we denote the following variance by

$$\mathcal{V}_g(x^*) := \mathbb{E}[(G(x^*,\xi) - g(x^*))^2].$$

Then, for every $x \in \mathcal{C}$, it yields that

$$\begin{aligned}
&\mathbb{E}[(G(x,\xi) - g(x))^2] \\
&\le 3\mathbb{E}[(G(x,\xi) - G(x^*,\xi))^2] + 3\mathbb{E}[(G(x^*,\xi) - g(x^*))^2] + 3(g(x) - g(x^*))^2 \\
&\le 3(C_g^2 + \hat{C}_g^2)\|x - x^*\|^2 + 3\mathcal{V}_g(x^*) \\
&\le \tfrac{6(C_g^2+\hat{C}_g^2)}{\lambda} V(x,x^*) + 3\mathcal{V}_g(x^*).
\end{aligned} \tag{3.1}$$

Let us mention that, with slight modification, the convergence results in this section can also be established if Assumption (A.3) is relaxed to $\mathbb{E}[\|G'(x,\xi)\|_*^2] \le \hat{C}_g^2$ and the latter condition in Assumption (A.2) is replaced with $\mathbb{E}[G^2(x,\xi)] \le C_t^2$.

The detail of the proposed basic PSG method is described in Algorithm 1, in which we assume that the stepsizes $\{\alpha_k\}, \{\beta_k\}, \{\gamma_k\}$ are diminishing sequences and $\beta_k \in (0,1)$. Our basic idea is to use the classical stochastic subgradient method to solve the penalty problem

$$\min_{x \in \mathcal{C}} f(x) + \frac{\tau}{2}\phi(x),$$

Algorithm 1: Basic penalized stochastic subgradient method

1 Initialization: Choose the initial point $x^0 \in \mathcal{C}$ and the stepsizes sequence $\{\alpha_k, \beta_k, \gamma_k\}$. Set $k = 0$ and $t_0 = 0$.

2 **while** *did not converge* **do**

3 Generate i.i.d. samples ξ^k and η^k of ξ and compute

$$x^{k+1} = \mathcal{P}_{x^k}\left(\alpha_k F'(x^k, \xi^k) + \gamma_k [t_{k+1}]_+ G'(x^k, \eta^k)\right), \tag{3.2}$$

where t_{k+1} is computed by

$$t_{k+1} = (1 - \beta_k) t_k + \beta_k G(x^k, \xi^k).$$

4 Set $k \leftarrow k + 1$.

where $\tau > 0$ is a penalty parameter and $\phi(x) = [g(x)]_+^2 = [\mathbb{E}[G(x, \xi)]]_+^2$. At first glance, the step (3.2) in Algorithm 1 seems a bit cumbersome and a more natural choice would be

$$x^{k+1} = \mathcal{P}_{x^k}\left(\alpha_k F'(x^k, \xi^k) + \gamma_k [G(x^k, \xi^k)]_+ G'(x^k, \eta^k)\right).$$

Unfortunately, as opposed to the first term $F'(x^k, \xi^k)$, the second term $[G(x^k, \xi^k)]_+ G'(x^k, \eta^k)$ is not an unbiased subgradient of $\frac{1}{2}\phi(x^k)$ which is a standard requirement in SA-type algorithms. Motivated by the stochastic compositional gradient algorithm in [59], we replace $G(x^k, \xi^k)$ in the second term with an iterative weighted average t_{k+1}. As will be shown later, t_{k+1} is a good stochastic approximation to $g(x^k)$, which is one of the key techniques to guarantee the convergence of Algorithm 1.

We shall study the convergence of the stochastic process $\{x^k\}$ generated by Algorithm 1 with respect to the following filtrations (σ-fields)

$$\mathcal{F}_k := \sigma(\xi^0, \eta^0, \ldots, \xi^{k-1}, \eta^{k-1}), \quad \hat{\mathcal{F}}_k := \sigma(\xi^0, \eta^0, \ldots, \xi^{k-1}, \eta^{k-1}, \xi^k).$$

The filtration $\mathcal{F}_k$ represents the information collected up to iteration $k-1$, and the filtration $\hat{\mathcal{F}}_k$ corresponds to the information collected up to iteration $k-1$ plus the sample ξ^k. It is not difficult to see that $x^k \in \mathcal{F}_k$ and $t_{k+1} \in \hat{\mathcal{F}}_k$.

We present some auxiliary results in the following lemma.

Lemma 3.1. *Under Assumption 1, it holds with probability 1 that, for all $k \geq 0$,*

(i) $\mathbb{E}[[t_{k+1}]_+^2 | \mathcal{F}_k] \leq C_t^2$;

(ii) $\mathbb{E}[\|[t_{k+1}]_+ G'(x^k, \eta^k)\|_*^2 | \mathcal{F}_k] \leq C_t^2 \hat{C}_g^2$;

(iii) $\mathbb{E}[\|x^{k+1} - x^k\|^2 | \mathcal{F}_k] \leq \frac{2C_f^2}{\lambda^2}\alpha_k^2 + \frac{2C_t^2 \hat{C}_g^2}{\lambda^2}\gamma_k^2$.

Proof. From $t_{k+1} = (1 - \beta_k) t_k + \beta_k G(x^k, \xi^k)$, by using the convexity of $[\cdot]_+^2$ we have that

$$\begin{aligned}
\mathbb{E}[[t_{k+1}]_+^2 | \mathcal{F}_k] &\leq \mathbb{E}[(1 - \beta_k)[t_k]_+^2 + \beta_k [G(x^k, \xi^k)]_+^2 | \mathcal{F}_k] \\
&\leq (1 - \beta_k)[t_k]_+^2 + \beta_k \mathbb{E}[[G(x^k, \xi^k)]_+^2 | \mathcal{F}_k], \\
&\leq (1 - \beta_k)[t_k]_+^2 + \beta_k C_t^2,
\end{aligned}$$

which verifies item (i) by induction. Item (ii) is immediately from

$$\mathbb{E}[\|[t_{k+1}]_+ G'(x^k, \eta^k)\|_*^2 | \mathcal{F}_k] = \mathbb{E}[[t_{k+1}]_+^2 \mathbb{E}[\|G'(x^k, \eta^k)\|_*^2 | \hat{\mathcal{F}}_k] | \mathcal{F}_k] \leq C_t^2 \hat{C}_g^2.$$

For item (iii), applying (2.1) and substituting $x^+ = x^{k+1}$ and $u = x^k$ in (2.2) we have

$$\frac{\lambda}{2}\|x^{k+1} - x^k\|^2 \leq V(x^{k+1}, x^k) \leq \frac{1}{2\lambda}\|\alpha_k F'(x^k, \xi^k) + \gamma_k [t_{k+1}]_+ G'(x^k, \eta^k)\|_*^2$$

and hence

$$\begin{aligned}\mathbb{E}[\|x^{k+1}-x^k\|^2|\mathcal{F}_k] &\leq \tfrac{2\alpha_k^2}{\lambda^2}\mathbb{E}[\|F'(x^k,\xi^k)\|_*^2|\mathcal{F}_k] + \tfrac{2\gamma_k^2}{\lambda^2}\mathbb{E}[\|t_{k+1}]_+G'(x^k,\eta^k)\|_*^2|\mathcal{F}_k]\\ &\leq \tfrac{2C_f^2}{\lambda^2}\alpha_k^2 + \tfrac{2C_t^2\hat{C}_g^2}{\lambda^2}\gamma_k^2.\end{aligned}$$

The proof is completed. □

We briefly state the main steps of the convergence analysis for Algorithm 1. The first step is to establish a (conditional expected) relationship between $V(x^{k+1},x^*)$ and $V(x^k,x^*)$. To achieve this, from (2.2) (let $x^+=x^{k+1}$ and $u=x^*$) and Lemma 3.1 we have

$$\begin{aligned}&\mathbb{E}[V(x^{k+1},x^*)|\mathcal{F}_k]\\ &\leq V(x^k,x^*) + \mathbb{E}[\tfrac{1}{\lambda}\|\alpha_k F'(x^k,\xi^k)\|_*^2|\mathcal{F}_k] + \mathbb{E}[\tfrac{1}{\lambda}\|\gamma_k[t_{k+1}]_+G'(x^k,\eta^k)\|_*^2|\mathcal{F}_k]\\ &\quad +\mathbb{E}[-\langle\alpha_k F'(x^k,\xi^k),x^k-x^*\rangle|\mathcal{F}_k] + \mathbb{E}[-\langle\gamma_k[t_{k+1}]_+G'(x^k,\eta^k),x^k-x^*\rangle|\mathcal{F}_k]. \qquad (3.3)\\ &\leq V(x^k,x^*) + \tfrac{1}{\lambda}(\alpha_k^2C_f^2+\gamma_k^2C_t^2\hat{C}_g^2) + \mathbb{E}[-\langle\alpha_k F'(x^k,\xi^k),x^k-x^*\rangle|\mathcal{F}_k]\\ &\quad +\mathbb{E}[-\langle\gamma_k[t_{k+1}]_+G'(x^k,\eta^k),x^k-x^*\rangle|\mathcal{F}_k], \quad \text{w.p.1.}\end{aligned}$$

Therefore, it is sufficient to derive the bounds of the last two terms on the right-hand side of the above inequality. Once the relationship is obtained, the second step is to get the almost sure global convergence by applying the supermartingale convergence theorem. Finally, by carefully choosing the stepsizes, we investigate the convergence rates of the averaged iterates for both convex and strongly convex cases.

In the following two lemmas, we get the bounds of the last two terms in (3.3), respectively.

Lemma 3.2. *Let x^* be any point in X^*. Under Assumption 1, it holds with probability 1 that,*

$$\begin{aligned}&\mathbb{E}[-\gamma_k\langle[t_{k+1}]_+G'(x^k,\eta^k),x^k-x^*\rangle|\mathcal{F}_k]\\ &\leq -\tfrac{\gamma_k}{2}\phi(x^k) + \left(\tfrac{C_g^2}{2\lambda}\cdot\tfrac{\gamma_k^2}{\beta_k} + \tfrac{6(C_g^2+\hat{C}_g^2)}{\lambda}\beta_k^2\right)V(x^k,x^*) + C_g^2\|x^k-x^{k-1}\|^2 \qquad (3.4)\\ &\quad +3\mathcal{V}_g(x^*)\beta_k^2 + \beta_k(g(x^{k-1})-t_k)^2.\end{aligned}$$

Proof. We first write

$$\begin{aligned}&-\gamma_k\langle[t_{k+1}]_+G'(x^k,\eta^k),x^k-x^*\rangle\\ &= -\gamma_k\langle g(x^k)]_+G'(x^k,\eta^k),x^k-x^*\rangle + \gamma_k\langle G'(x^k,\eta^k),x^k-x^*\rangle([g(x^k)]_+-[t_{k+1}]_+).\end{aligned}$$

Let $g'(x^k)=\mathbb{E}[G'(x^k,\eta^k)|\hat{\mathcal{F}}_k]$. By taking conditional expectation on both sides of the above equation, it holds with probability 1 that

$$\begin{aligned}&\mathbb{E}[-\gamma_k\langle[t_{k+1}]_+G'(x^k,\eta^k),x^k-x^*\rangle|\hat{\mathcal{F}}_k]\\ &= -\gamma_k\langle[g(x^k)]_+g'(x^k),x^k-x^*\rangle + \gamma_k\langle g'(x^k),x^k-x^*\rangle([g(x^k)]_+-[t_{k+1}]_+)\\ &\leq -\tfrac{\gamma_k}{2}\langle\phi'(x^k),x^k-x^*\rangle + \tfrac{\gamma_k^2}{4\beta_k}\|g'(x^k)\|_*^2\cdot\|x^k-x^*\|^2 + \beta_k([g(x^k)]_+-[t_{k+1}]_+)^2\\ &\leq -\tfrac{\gamma_k}{2}\phi(x^k) + \tfrac{C_g^2}{2\lambda}\cdot\tfrac{\gamma_k^2}{\beta_k}V(x^k,x^*) + \beta_k(g(x^k)-t_{k+1})^2.\end{aligned}$$

In the first inequality above we use the facts that $\phi'(x^k):=2[g(x^k)]_+g'(x^k)\in\partial\phi(x^k)$ and $ab\leq\frac{1}{4\beta_k}a^2+\beta_kb^2$. The convexity of ϕ and $\phi(x^*)=0$ are used in the second inequality. Thus, by using $\mathbb{E}[\mathbb{E}[\cdot|\hat{F}_k]|\mathcal{F}_k]=\mathbb{E}[\cdot|\mathcal{F}_k]$ we get

$$\begin{aligned}&\mathbb{E}[-\gamma_k\langle[t_{k+1}]_+G'(x^k,\eta^k),x^k-x^*\rangle|\mathcal{F}_k]\\ &\leq -\tfrac{\gamma_k}{2}\phi(x^k) + \tfrac{C_g^2}{2\lambda}\cdot\tfrac{\gamma_k^2}{\beta_k}V(x^k,x^*) + \beta_k\mathbb{E}[(g(x^k)-t_{k+1})^2|\mathcal{F}_k], \quad \text{w.p.1.}\end{aligned} \qquad (3.5)$$

Finally, we derive the claimed result by noticing that the last term above is bounded as follows,

$$\begin{aligned}
&\mathbb{E}[(g(x^k)-t_{k+1})^2|\mathcal{F}_k] = \mathbb{E}[((1-\beta_k)(g(x^k)-t_k)+\beta_k(g(x^k)-G(x^k,\xi^k)))^2|\mathcal{F}_k]\\
&\le (1-\beta_k)(g(x^k)-t_k)^2+\beta_k\mathbb{E}[(g(x^k)-G(x^k,\xi^k))^2|\mathcal{F}_k]\\
&\le (1-\beta_k)(1+1/\beta_k)(g(x^k)-g(x^{k-1}))^2+(1-\beta_k)(1+\beta_k)(g(x^{k-1})-t_k)^2\\
&\quad +\beta_k[\tfrac{6(C_g^2+\hat{C}_g^2)}{\lambda}V(x^k,x^*)+3\mathcal{V}_g(x^*)]\\
&\le \tfrac{1}{\beta_k}C_g^2\|x^k-x^{k-1}\|^2+(g(x^{k-1})-t_k)^2+\beta_k\left[\tfrac{6(C_g^2+\hat{C}_g^2)}{\lambda}V(x^k,x^*)+3\mathcal{V}_g(x^*)\right], \quad \text{w.p.1.}
\end{aligned}$$

Let us note that, the second inequality above is from the fact that $(a+b)^2 \le (1+1/\beta_k)a^2+(1+\beta_k)b^2$ and (3.1), in the third inequality we use the Lipschitz continuity of g and the fact $1-\beta_k^2 \le 1$. □

Lemma 3.3. *Under Assumption 1, for every $x^* \in X^*$ and $u \in \mathcal{C}$, we get*

$$\mathbb{E}[-\alpha_k\langle F'(x^k,\xi^k), x^k-x^*\rangle|\mathcal{F}_k] \le -\alpha_k(f(u)-f^*)+\eta C_f^2\alpha_k^2+\frac{1}{4\eta}\|x^k-u\|^2$$

with probability 1, where $\eta>0$ is an arbitrary scalar.

Proof. In view of $f'(x^k) := \mathbb{E}[F'(x^k,\xi^k)|\mathcal{F}_k] \in \partial f(x^k)$, we get

$$\mathbb{E}[-\alpha_k\langle F'(x^k,\xi^k), x^k-x^*\rangle|\mathcal{F}_k] \le -\alpha_k(f(x^k)-f^*), \quad \text{w.p.1.} \tag{3.6}$$

Further, by using the convexity of f and $\|f'(u)\|_* \le C_f$ we have

$$-\alpha_k(f(x^k)-f(u)) \le -\alpha_k\langle f'(u), x^k-u\rangle \le \eta C_f^2\alpha_k^2+\frac{1}{4\eta}\|x^k-u\|^2.$$

Substituting the above inequality into (3.6), we get the claim. □

In the following subsections, we shall establish the almost sure global convergence and expected convergence rates of Algorithm 1 based on the previous auxiliary lemmas.

3.2 Almost sure convergence

In the sequel, we need the following global error bound condition that is a commonly used assumption for analyzing the algorithms for convex inequality constrained programs.

Assumption 2. *We suppose that there exists a constant $C_{eb}>0$ such that*

$$dist^2(x,\Phi) \le C_{eb}\cdot[g(x)]_+^2, \quad \forall x \in \mathcal{C}. \tag{3.7}$$

Since the pioneering work of Hoffman [18], error bound conditions have been extensively studied, both in variational analysis [20, 6] (especially in connection with metric regularity and subregularity as well as calmness and weak sharp minima) and in numerical analysis, such as the establishment of linear convergence of first order methods [32] and the convergence of alternating projection methods [36]. In [28, 33] the authors investigated many sufficient conditions for convex inequality systems to possess a global error bound, e.g., when the Slater condition holds and $\mathrm{cl}(\mathcal{C}\cap g^{-1}(0))$ is bounded.

Denote $\bar{x}^k := \Pi_\Phi(x^k)$, then the global error bound condition (3.7) is rewritten by

$$\phi(x^k) \ge \frac{1}{C_{eb}}\|x^k-\bar{x}^k\|^2. \tag{3.8}$$

We also have that the parameter satisfies $C_{eb} \ge 1/C_g^2$, by noticing that $g(\bar{x}^k)\le 0$ and hence

$$\phi(x^k) = [g(x^k)]_+^2 \le (g(x^k)-g(\bar{x}^k))^2 \le C_g^2\|x^k-\bar{x}^k\|^2.$$

Let us also mention that in Assumption 2 the set Φ can be replaced with $\mathcal{G} := \{x\in\mathbb{R}^n | g(x)\le 0\}$. It follows that, if g is convex and continuous, the set $\mathcal{G}$ is closed and convex. Then, for all $x\in\mathcal{C}$, it yields that $\mathrm{dist}(x,\mathcal{G}) = \mathrm{dist}(x,\Phi)$.

So far, we have not imposed any severe assumptions on the stepsizes $\alpha_k, \beta_k, \gamma_k$. It is well known that the stepsize requirement plays a significant role in the convergence analysis of SA-type methods (see for example, [47, 24, 38, 59]). To obtain the almost sure convergence of Algorithm 1, it is required that the stepsize sequences satisfy the following conditions.

Assumption 3. *Let the diminishing stepsize sequences $\{\alpha_k\}$, $\{\beta_k\}$ and $\{\gamma_k\}$ be such that*

$$0 < \beta_k < 1, \quad \frac{\gamma_{k-1}}{\beta_k} \to 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \left(\beta_k^2 + \frac{\gamma_k^2}{\beta_{k+1}} + \frac{\alpha_k^2}{\gamma_k} \right) < \infty.$$

In addition, we assume $\gamma_k \geq 4C_{eb}\alpha_k$.

We mention that, when Assumption 3 holds true, it implicitly implies that $\alpha_k < \gamma_k < \beta_{k+1}$ (for sufficiently large k) and

$$\frac{\alpha_{k-1}}{\beta_k} \to 0, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty, \quad \sum_{k=0}^{\infty} \gamma_k^2 < \infty, \quad \sum_{k=1}^{\infty} \frac{\alpha_{k-1}^2}{\beta_k} < \infty, \quad \sum_{k=0}^{\infty} \frac{\gamma_k^2}{\beta_k} < \infty. \tag{3.9}$$

It is not difficult to verify that, a simple example of the stepsizes that satisfies Assumption 3 is:

$$\alpha_k = \alpha \cdot k^{-(\frac{7}{8}+\varepsilon)}, \quad \beta_{k+1} = \beta \cdot k^{-(\frac{1}{2}+\varepsilon)}, \quad \gamma_k = \gamma \cdot k^{-(\frac{3}{4}+\varepsilon)}, \quad \forall k \geq 1, \tag{3.10}$$

where $\alpha, \beta, \gamma > 0$ are proper constants, the parameter $\varepsilon \in (0, \frac{1}{8})$ and $\alpha_0, \beta_0, \beta_1, \gamma_0$ are carefully selected[2]. In this case, for all $K \geq 2$ one has

$$\sum_{k=\lfloor K/2 \rfloor}^{K-1} \alpha_k \geq \alpha \int_{\lfloor K/2 \rfloor}^{K} u^{-(\frac{7}{8}+\varepsilon)} \mathrm{d}u \geq 2\alpha c(\varepsilon) K^{\frac{1}{8}-\varepsilon}, \tag{3.11}$$

where $c(\varepsilon)$ is defined by

$$c(\varepsilon) := \frac{1 - 2^{\varepsilon - \frac{1}{8}}}{1/8 - \varepsilon}.$$

In Lemma 3.4 we get the convergence property of the sequence $\{(g(x^k) - t_{k+1})^2\}$, which demonstrates that t_{k+1} is a good stochastic approximation to $g(x^k)$.

Lemma 3.4. *Suppose that Assumptions 1 and 3 hold. Then, the sequence $\{(g(x^{k-1}) - t_k)^2\}$ converges almost surely to zero.*

Proof. Under Assumption 1, from [59, Lemma 2], for $k \geq 1$ we have

$$\mathbb{E}[(g(x^k) - t_{k+1})^2 | \mathcal{F}_k] \leq (1 - \beta_k)(g(x^{k-1}) - t_k)^2 + \beta_k^{-1} C_g^2 \|x^k - x^{k-1}\|^2 + 2\mathcal{V}_g(x^*)\beta_k^2 \tag{3.12}$$

with probability 1. Denote

$$w_k := \beta_k^{-1} C_g^2 \|x^k - x^{k-1}\|^2 + 2\mathcal{V}_g(x^*)\beta_k^2.$$

To apply Lemma A.1, we need to show that $\sum_k w_k < \infty$ with probability 1 and $\mathbb{E}[w_k]/\beta_k \to 0$.

From item (iii) in Lemma 3.1, it follows that

$$\mathbb{E}[\|x^k - x^{k-1}\|^2] \leq \frac{2C_f^2}{\lambda^2} \alpha_{k-1}^2 + \frac{2C_t^2 \hat{C}_g^2}{\lambda^2} \gamma_{k-1}^2, \tag{3.13}$$

which, together with Assumption 3, yields that

$$\sum_{k=1}^{\infty} \mathbb{E}[\beta_k^{-1} C_g^2 \|x^k - x^{k-1}\|^2] \leq \sum_{k=1}^{\infty} \frac{2C_g^2}{\lambda^2} \left(C_f^2 \frac{\alpha_{k-1}^2}{\beta_k} + C_t^2 \hat{C}_g^2 \frac{\gamma_{k-1}^2}{\beta_k} \right) < \infty.$$

[2]We remark that (3.10) with $\varepsilon = 1/8$ also satisfies Assumption 3. However, for simplicity we omit this case in the following analysis since the discussion is different from the case $\varepsilon \in (0, \frac{1}{8})$.

Therefore, we get

$$\mathbb{E}\left[\sum_{k=1}^{\infty}\beta_k^{-1}C_g^2\|x^k-x^{k-1}\|^2\right]<\infty,$$

which, together with Markov's inequality [56], indicates

$$\sum_{k=1}^{\infty}\beta_k^{-1}C_g^2\|x^k-x^{k-1}\|^2<\infty,\quad \text{w.p.1.} \tag{3.14}$$

Due to $\sum_k\beta_k^2<\infty$ under Assumption 3, together with (3.14) we get $\sum_k w_k<\infty$ with probability 1.

By using (3.13) again, we have that

$$\frac{\mathbb{E}[w_k]}{\beta_k}\le\frac{2C_f^2}{\lambda^2}\cdot\frac{\alpha_{k-1}^2}{\beta_k^2}+\frac{2C_t^2\hat{C}_g^2}{\lambda^2}\cdot\frac{\gamma_{k-1}^2}{\beta_k^2}+2\mathcal{V}_g(x^*)\beta_k.$$

It follows from Assumption 3 that $\mathbb{E}[w_k]/\beta_k\to 0$.

Finally, by using Lemma A.1 to (3.12), we get the result. □

In what follows, we establish an important recursive relation in order to use the supermartingale convergence theorem to obtain the required almost sure convergence.

Lemma 3.5. *Suppose that Assumptions 1-3 hold true. Let x^* be any given optimal solution to problem (1.1). Then, we have*

$$\mathbb{E}[\zeta_{k+1}|\mathcal{F}_k]\le(1+a_k)\zeta_k-u_k+b_k,\quad w.p.1, \tag{3.15}$$

where $\zeta_k:=V(x^k,x^*)+(g(x^{k-1})-t_k)^2$,

$$a_k:=\frac{C_g^2}{2\lambda}\cdot\frac{\gamma_k^2}{\beta_k}+\frac{6(C_g^2+\hat{C}_g^2)}{\lambda}\beta_k^2,\quad u_k:=\alpha_k(f(\bar{x}^k)-f^*+\|x^k-\bar{x}^k\|^2)$$

and

$$b_k:=C_{eb}C_f^2\frac{\alpha_k^2}{\gamma_k}+\frac{C_f^2}{\lambda}\alpha_k^2+5\mathcal{V}_g(x^*)\beta_k^2+\frac{\hat{C}_g^2C_t^2}{\lambda}\gamma_k^2+(1+1/\beta_k)C_g^2\|x^k-x^{k-1}\|^2.$$

Proof. Substituting the results in Lemma 3.2 and Lemma 3.3 (take $u=\bar{x}^k$ and $\eta=C_{eb}/\gamma_k$) into (3.3), using the error bound condition (3.8) and rearranging, we have

$$\begin{aligned}&\mathbb{E}[V(x^{k+1},x^*)|\mathcal{F}_k]\\&\le\left(1+\tfrac{C_g^2}{2\lambda}\cdot\tfrac{\gamma_k^2}{\beta_k}+\tfrac{6(C_g^2+\hat{C}_g^2)}{\lambda}\beta_k^2\right)V(x^k,x^*)-\alpha_k(f(\bar{x}^k)-f^*)-\tfrac{\gamma_k}{4C_{eb}}\|x^k-\bar{x}^k\|^2\\&\quad+\beta_k(g(x^{k-1})-t_k)^2+C_{eb}C_f^2\tfrac{\alpha_k^2}{\gamma_k}+\tfrac{C_f^2}{\lambda}\alpha_k^2+3\mathcal{V}_g(x^*)\beta_k^2+\tfrac{\hat{C}_g^2C_t^2}{\lambda}\gamma_k^2+C_g^2\|x^k-x^{k-1}\|^2\end{aligned}$$

with probability 1. Furthermore, from $-\frac{\gamma_k}{4C_{eb}}\le-\alpha_k$ it follows that

$$\begin{aligned}&\mathbb{E}[V(x^{k+1},x^*)|\mathcal{F}_k]\\&\le\left(1+\tfrac{C_g^2}{2\lambda}\cdot\tfrac{\gamma_k^2}{\beta_k}+\tfrac{6(C_g^2+\hat{C}_g^2)}{\lambda}\beta_k^2\right)V(x^k,x^*)-\alpha_k(f(\bar{x}^k)-f^*+\|x^k-\bar{x}^k\|^2)+\beta_k(g(x^{k-1})-t_k)^2\\&\quad+C_{eb}C_f^2\tfrac{\alpha_k^2}{\gamma_k}+\tfrac{C_f^2}{\lambda}\alpha_k^2+3\mathcal{V}_g(x^*)\beta_k^2+\tfrac{\hat{C}_g^2C_t^2}{\lambda}\gamma_k^2+C_g^2\|x^k-x^{k-1}\|^2\end{aligned}$$

with probability 1. Thus, by combining with (3.12), we derive the claim. □

We now present the almost sure convergence result in the following theorem. The proof is partially inspired by [2, Proposition 6.4.8] and [59, Theorem 1].

Theorem 3.6. *Let Assumptions 1-3 hold. Then the sequence $\{x^k\}$ converges almost surely to a point in X^*.*

Proof. From Assumption 3, (3.14) and the definitions of a_k, b_k, it is easy to get that $\sum_k a_k < \infty$ and $\sum_k b_k < \infty$ with probability 1. Let x^* be an arbitrary point in X^*. By applying the supermartingale convergence theorem to (3.15), we obtain that $\{V(x^k, x^*) + (g(x^{k-1}) - t_k)^2\}$ converges almost surely to a finite random variable, and

$$\sum_{k=1}^{\infty} \alpha_k (f(\bar{x}^k) - f^* + \|x^k - \bar{x}^k\|^2) < \infty, \quad \text{w.p.1.}$$

Since it has been proved in Lemma 3.4 that $\{(g(x^{k-1}) - t_k)^2\}$ converges almost surely to zero, we have that $\{V(x^k, x^*)\}$ and consequently $\{\|x^k - x^*\|^2\}$ converge almost surely. Hence, $\{x^k\}$ is bounded with probability 1. Moreover, in view of $\sum_k \alpha_k = \infty$, we have

$$\liminf_{k\to\infty} (f(\bar{x}^k) - f^* + \|x^k - \bar{x}^k\|^2) = 0, \quad \text{w.p.1.}$$

Let us now define the following set $\bar{\Omega} \subset \Omega$: $\omega \in \bar{\Omega}$ if and only if the trajectory $\{x^k\} \equiv \{x^k(\omega)\}$ satisfies that the sequence $\{\|x^k(\omega) - x^*\|^2\}$ converges for every $x^* \in X^*$, $\{x^k(\omega)\}$ is bounded and $\liminf_{k\to\infty}(f(\bar{x}^k(\omega)) - f^* + \|x^k(\omega) - \bar{x}^k(\omega)\|^2) = 0$, where $\bar{x}^k(\omega) := \Pi_\Phi(x^k(\omega))$. It can be shown that the set $\bar{\Omega}$ is measurable and $\mathbb{P}(\bar{\Omega}) = 1$, see [59, Theorem 1] for example. Therefore, it suffices to prove that for every $\omega \in \bar{\Omega}$ the trajectory $\{x^k(\omega)\}$ converges to some point in X^*.

For a fixed $\omega \in \bar{\Omega}$. Since $\{x^k(\omega)\}$ is bounded and $f(\bar{x}^k(\omega)) - f^* \geq 0$, together with

$$\liminf_{k\to\infty} (f(\bar{x}^k(\omega)) - f^* + \|x^k(\omega) - \bar{x}^k(\omega)\|^2) = 0,$$

it follows that there exists a subsequence $\{x^{k_j}(\omega)\}$ such that $\|x^{k_j}(\omega) - \bar{x}^{k_j}(\omega)\|^2$ converges to zero and $\{\bar{x}^{k_j}(\omega)\}$ converges to a point $\hat{x}$ in X^*, hence $\hat{x}$ is a cluster point of $\{x^k(\omega)\}$. Combining with that $\{\|x^k(\omega) - \hat{x}\|^2\}$ converges, we obtain that $\{x^k(\omega)\}$ converges to $\hat{x}$. □

3.3 Expected convergence rates

For a positive number $K \in \mathbb{N}$ and $K \geq 2$, we define

$$c_K(x^*) := \sup_{\lfloor \frac{K}{2} \rfloor \leq k \leq K-1} \mathbb{E}[\zeta_k], \quad c_K^* := \inf_{x^* \in X^*} c_K(x^*),$$

where ζ_k is defined in Lemma 3.5. We have proved in Theorem 3.6 that $\{\zeta_k\}$ converges to a finite random variable with probability 1, hence $c_K(x^*)$ is bounded and $c_\infty(x^*)$ is finite. Define

$$\hat{c}_K := c_K^* \sum_{k=\lfloor K/2 \rfloor}^{K-1} a_k + \sum_{k=\lfloor K/2 \rfloor}^{K-1} \mathbb{E}[b_k].$$

Under Assumption 3, owing to $\sum_k a_k < \infty$ and $\sum_k b_k < \infty$ with probability 1, we have that $\hat{c}_K$ is finite for all $K \geq 2$. Let us define the following weighted averages

$$\hat{x}^K := \frac{\sum_{k=\lfloor K/2 \rfloor}^{K-1} \alpha_k x^k}{\sum_{k=\lfloor K/2 \rfloor}^{K-1} \alpha_k}, \quad \hat{u}^K := \frac{\sum_{k=\lfloor K/2 \rfloor}^{K-1} \alpha_k \bar{x}^k}{\sum_{k=\lfloor K/2 \rfloor}^{K-1} \alpha_k}. \tag{3.16}$$

Recall that the sets $\mathcal{C}$ and Φ are both convex, we have $\hat{x}^K \in \mathcal{C}$ and $\hat{u}^K \in \Phi$. We now present the convergence rates of Algorithm 1.

Theorem 3.7. *Let Assumptions 1 and 2 hold and the stepsizes be chosen as in (3.10). Then, it gives*

$$\mathbb{E}[f(\hat{u}^K) - f^* + \|\hat{x}^K - \hat{u}^K\|^2] = \mathcal{O}\left(\frac{1}{K^{1/8-\varepsilon}}\right).$$

Proof. From (3.15), for $\lfloor \frac{K}{2} \rfloor \le k \le K-1$ we have

$$\alpha_k \mathbb{E}[(f(\bar{x}^k) - f^* + \|x^k - \bar{x}^k\|^2)] \le \mathbb{E}[\zeta_k] - \mathbb{E}[\zeta_{k+1}] + c_K^* a_k + \mathbb{E}[b_k].$$

Summing up from $k = \lfloor K/2 \rfloor$ to $K-1$ we get

$$\sum_{k=\lfloor K/2 \rfloor}^{K-1} \alpha_k \mathbb{E}[(f(\bar{x}^k) - f^* + \|x^k - \bar{x}^k\|^2)] \le \mathbb{E}[\zeta_{\lfloor K/2 \rfloor}] + \hat{c}_K \le c_K^* + \hat{c}_K.$$

By using the definitions of $\hat{x}^K, \hat{u}^K$ and the convexity of f, it follows

$$\mathbb{E}[f(\hat{u}^K) - f^* + \|\hat{x}^K - \hat{u}^K\|^2] \le \frac{\sum_{k=\lfloor K/2 \rfloor}^{K-1} \alpha_k \mathbb{E}[(f(\bar{x}^k) - f^* + \|x^k - \bar{x}^k\|^2)]}{\sum_{k=\lfloor K/2 \rfloor}^{K-1} \alpha_k},$$

and thus from (3.11),

$$\mathbb{E}[f(\hat{u}^K) - f^* + \|\hat{x}^K - \hat{u}^K\|^2] \le \frac{c_K^* + \hat{c}_K}{4\alpha c(\varepsilon)} \cdot \frac{1}{K^{1/8-\varepsilon}}.$$

The proof is completed. □

We remark that, let ε be close to zero, it roughly yields

$$\mathbb{E}[f(\hat{u}^K) - f^* + \|\hat{x}^K - \hat{u}^K\|^2] \approx \mathcal{O}\left(\frac{1}{K^{1/8}}\right).$$

Here, $f(\hat{u}^K) - f^*$ presents the distance to the optimal value and $\|\hat{x}^K - \hat{u}^K\|^2$ evaluates the violation of constraint.

3.4 Convergence under restricted strong convexity

In this subsection, we investigate the convergence of Algorithm 1 when problem (1.1) satisfies the following restricted strongly convex condition.

Assumption 4. *We assume:*

(i) There exist $x^* \in X^*$ and a penalty parameter $\rho > 0$ such that x^* is an optimal solution to the following penalty problem

$$\min_{x \in \mathcal{C}} f(x) + \rho\phi(x). \tag{3.17}$$

(ii) The objective function $f + \rho\phi(x)$ is restricted strongly convex (or has a quadratic growth) over the set $\mathcal{C}$ with modulus $\kappa > 0$, i.e.,

$$f(x) + \rho\phi(x) \ge f^* + \frac{\kappa}{2}\|x - x^*\|_2^2, \quad \forall x \in \mathcal{C}. \tag{3.18}$$

Using penalty function to move the functional constraints to objective is a simple and popular technique to solve functional constrained programs. Here, we do not intend to solve the penalty problem (3.17). Instead, we only require that the original problem has such an equivalent reformulation. For item (ii) in Assumption 4, we remark that in (3.18) the term $\rho\phi(x^*)$ is omitted on the right-hand side since $\phi(x^*) = 0$. This restricted strongly convex condition, also named as quadratic growth condition, plays a significant role in the convergence analysis of various convex optimization methods [60, 13, 35, 37]. The quadratic growth condition (in the local form) is also an important concept in the stability analysis of (nonconvex) optimization problems [3, 20, 11]. Under certain constraint qualifications, the quadratic growth condition is equivalent to the standard second order sufficient condition [4].

For simplicity we consider the Euclidean distance $V(x, y) = \frac{1}{2}\|x - y\|_2^2$ in this subsection. Consequently, we have $\lambda = 1$ and the step (3.2) is reduced to

$$x^{k+1} = \Pi_{\mathcal{C}}\left(x^k - \alpha_k F'(x^k, \xi^k) - \gamma_k [t_{k+1}]_+ G'(x^k, \eta^k)\right).$$

Consider the following stepsizes policy

$$\alpha_k = \alpha \cdot k^{-(\frac{3}{4}+2\varepsilon)}, \quad \beta_{k+1} = \beta \cdot k^{-(\frac{1}{2}+\varepsilon)}, \quad \gamma_k = \gamma \cdot k^{-(\frac{3}{4}+\varepsilon)}, \quad \forall k \geq 1, \tag{3.19}$$

where $\alpha, \beta, \gamma > 0$ are proper constants such that $\alpha_k < \min\{\frac{2}{\kappa}, \frac{\beta_k}{\kappa}, \frac{\gamma_k}{2\rho}\}$, the parameter $\varepsilon \in (0, \frac{1}{8})$ and $\alpha_0, \beta_0, \beta_1, \gamma_0$ are carefully selected. Then, it is not difficult to verify that

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \gamma_k^2 < \infty, \quad \sum_{k=0}^{\infty} \beta_k^2 < \infty, \quad \sum_{k=1}^{\infty} \frac{\gamma_{k-1}^2}{\beta_k} < \infty.$$

In the next theorem, we present the global convergence of $\{x^k\}$ and the convergence rates of the averaged iterates for the restricted strongly convex problem.

Theorem 3.8. *Suppose that Assumptions 1 and 4 hold. If the diminishing stepsizes are given by (3.19), then we have that $\{\|x^k - x^*\|_2^2\}$ and $\{(g(x^k) - t_{k+1})^2\}$ converge almost surely to zero and*

$$\mathbb{E}[\|\hat{x}^K - x^*\|_2^2] \approx \mathcal{O}\left(\frac{1}{K^{1/4}}\right).$$

Proof. We first define $v_k := \frac{1}{2}\|x^k - x^*\|_2^2 + \delta_k(g(x^{k-1}) - t_k)^2$ and

$$\delta_k := \frac{\beta_k}{\beta_k - \frac{\kappa}{2}\alpha_k}, \quad u_k := C_f^2\alpha_k^2 + \hat{C}_g^2 C_t^2 \gamma_k^2 + 6C_g^2(C_t^2\hat{C}_g^2\frac{\gamma_{k-1}^2}{\beta_k} + C_f^2\frac{\alpha_{k-1}^2}{\beta_k}) + 6\mathcal{V}_g(x^*)\beta_k^2.$$

From $\beta_k \geq \kappa\alpha_k$ and the definition of δ_k, by some calculations (see Lemma A.2 in the appendix), for k large enough we have

$$(\delta_k + \beta_k)(1 - \beta_k) \leq (1 - \frac{\kappa}{2}\alpha_k)\delta_k, \quad \delta_k \leq 2, \quad \delta_{k+1} \leq \delta_k, \quad \delta_k > 1. \tag{3.20}$$

We next get a recursive relation by using a similar argument as in Lemma 3.5. Taking expectation on both sides of (3.3), together with (3.5) and (3.6), we have

$$\begin{aligned}\mathbb{E}[\tfrac{1}{2}\|x^{k+1} - x^*\|_2^2|\mathcal{F}_k] \quad &\leq \tfrac{1}{2}\|x^k - x^*\|_2^2 + C_g^2\tfrac{\gamma_k^2}{4\beta_k}\|x^k - x^*\|_2^2 - \alpha_k(f(x^k) - f^*) - \tfrac{\gamma_k}{2}\phi(x^k) + C_f^2\alpha_k^2 \\ &\quad + \hat{C}_g^2 C_t^2\gamma_k^2 + \beta_k\mathbb{E}[(g(x^k) - t_{k+1})^2|\mathcal{F}_k], \quad \text{w.p.1},\end{aligned}$$

which, together with $\frac{\gamma_k}{2} \geq \rho\alpha_k$, $\frac{C_g^2\gamma_k^2}{4\beta_k} \leq \frac{\kappa}{4}\alpha_k$ (which holds when k is large enough) and assumption 3.18, gives

$$\begin{aligned}\mathbb{E}[\tfrac{1}{2}\|x^{k+1} - x^*\|_2^2|\mathcal{F}_k] \quad &\leq \big(1 - \tfrac{\kappa}{2}\alpha_k\big) \cdot \tfrac{1}{2}\|x^k - x^*\|_2^2 + C_f^2\alpha_k^2 + \hat{C}_g^2 C_t^2\gamma_k^2 \\ &\quad + \beta_k\mathbb{E}[(g(x^k) - t_{k+1})^2|\mathcal{F}_k], \quad \text{w.p.1}.\end{aligned}$$

By multiplying (3.12) with $(\delta_k + \beta_k)$ and adding it to the above inequality, we get

$$\begin{aligned}&\mathbb{E}[\tfrac{1}{2}\|x^{k+1} - x^*\|_2^2 + \delta_k(g(x^k) - t_{k+1})^2|\mathcal{F}_k] \\ &\leq (1 - \tfrac{\kappa}{2}\alpha_k) \cdot \tfrac{1}{2}\|x^k - x^*\|_2^2 + (\delta_k + \beta_k)(1 - \beta_k)(g(x^{k-1}) - t_k)^2 + C_f^2\alpha_k^2 + \hat{C}_g^2 C_t^2\gamma_k^2 \\ &\quad + (\delta_k + \beta_k)\beta_k^{-1}C_g^2\|x^k - x^{k-1}\|_2^2 + 2\mathcal{V}_g(x^*)(\delta_k + \beta_k)\beta_k^2, \quad \text{w.p.1},\end{aligned}$$

which, together with (3.20), item (iii) in Lemma 3.1 and the definition of v_k, implies

$$\mathbb{E}[v_{k+1}|\mathcal{F}_k] \leq \mathbb{E}[\|x^{k+1} - x^*\|_2^2/2 + \delta_k(g(x^k) - t_{k+1})^2|\mathcal{F}_k] \leq (1 - \frac{\kappa}{2}\alpha_k)v_k + u_k \quad \text{w.p.1}.$$

It is easy to check that

$$\frac{\kappa}{2}\alpha_k < 1, \quad \sum_k \alpha_k = \infty, \quad \sum_k u_k < \infty, \quad \frac{u_k}{\alpha_k} \to 0.$$

Therefore, by using Lemma A.1, it follows that $\{v_k\}$ converges almost surely to zero, which obviously yields that $\{\|x^k - x^*\|_2^2\}$ converges almost surely to zero. Due to $\delta_k > 1$ when k is large enough, we also have that $\{(g(x^k) - t_{k+1})^2\}$ converges almost surely to zero.

From the previous result $\mathbb{E}[v_{k+1}|\mathcal{F}_k] \leq (1 - \frac{\kappa}{2}\alpha_k)v_k + u_k$, one has

$$\frac{\kappa}{2}\sum_{k=\lfloor K/2\rfloor}^{K-1} \alpha_k \mathbb{E}[v_k] \leq \mathbb{E}[v_{\lfloor K/2\rfloor}] + \sum_{k=\lfloor K/2\rfloor}^{K-1} u_k < \infty.$$

Then, from the convexity of $\|\cdot\|_2^2$, we have

$$\mathbb{E}[\|\hat{x}^K - x^*\|_2^2] \leq \frac{\sum_{k=\lfloor K/2\rfloor}^{K-1} \alpha_k \mathbb{E}[\|x^k - x^*\|_2^2]}{\sum_{k=\lfloor K/2\rfloor}^{K-1} \alpha_k} \leq \frac{\sum_{k=\lfloor K/2\rfloor}^{K-1} 2\alpha_k \mathbb{E}[v_k]}{\sum_{k=\lfloor K/2\rfloor}^{K-1} \alpha_k} \leq \mathcal{O}\left(\frac{1}{K^{1/4-2\varepsilon}}\right),$$

by using a similar result of (3.11). The proof is completed. □

Let us mention that the global error bound condition (Assumption 2) is not required for this restricted strongly convex setting. Comparing with the convex setting, we observe that the convergence rate is (approximately) improved from $\mathcal{O}(\frac{1}{K^{1/8}})$ to $\mathcal{O}(\frac{1}{K^{1/4}})$.

4 Penalized mini-batch stochastic gradient method

In this section, we consider an extension of the basic PSG method, in which we use the mini-batch stochastic gradient associated with the objective function. We will show that, when the objective function f is differentiable and its gradient is Lipschitz continuous, the convergence rates of this mini-batch PSG can be improved and the choice of stepsizes is more flexible. Moreover, it has been witnessed in practice that the mini-batch algorithm is usually more efficient than the basic algorithm.

4.1 Assumptions, algorithmic framework and lemmas

We make the following assumption, in which (A.4) implies that the gradient $\nabla f(x) := \mathbb{E}[\nabla_x F(x,\xi)]$ is Lipschitz continuous with parameter L_f.

Assumption 5. *In addition to (A.1), (A.2) and (A.3) in Assumption 1, we also assume:*

(A.4) *For almost every $\omega \in \Omega$, the function $F(\cdot,\xi) \equiv F(\cdot,\xi(\omega))$ is differentiable on $\mathcal{C}$ and there exists a measurable function $l_f : \Xi \to \mathbb{R}_+$ with $L_f := \mathbb{E}[l_f(\xi)]$ and $\hat{L}_f := \sqrt{\mathbb{E}[l_f^2(\xi)]} < \infty$, such that for every $x, y \in \mathcal{C}$,*

$$\|\nabla_x F(x,\xi) - \nabla_x F(y,\xi)\| \leq l_f(\xi)\|x - y\|.$$

The detail of the proposed mini-batch algorithm is described in Algorithm 2, in which N_k is the batch size of ξ^k, and η^k is still a single i.i.d. sample as in Algorithm 1. We make the following remarks on Algorithm 2. Firstly, the reason why the stochastic (sub)gradient $G'(x^k,\eta^k)$ is not replaced with its mini-batch counterpart is that theoretically we can not derive a better convergence rate even if $G(x,\xi)$ is assumed to be differentiable and Lipschitz gradient continuous with respect to x. However, in practice we suggest using the mini-batch (sub)gradient to replace $G'(x^k,\eta^k)$. Secondly, we recall that, in the definition of the prox-mapping $\mathcal{P}_x(\cdot)$ it is required that the reference point x lies in the set $\mathcal{C}$. Therefore, we have to compute the prox-mapping twice at each iteration such that $y^k \in \mathcal{C}$. Particularly, if we use Euclidean distance, these two prox-mapping steps, respectively, can be replaced with

$$y^k = x^k - \alpha_k \nabla \mathcal{F}^k, \quad x^{k+1} = \Pi_{\mathcal{C}}\left(y^k - \gamma_k [t_{k+1}]_+ G'(x^k,\eta^k)\right). \tag{4.1}$$

In this setting, only one prox-mapping (projection) is computed.

Before proceeding with the convergence analysis, we first introduce some notations. For every $x^* \in X^*$, we define two variances by

$$\mathcal{V}_{f'}(x^*) := \mathbb{E}[\|\nabla_x F(x^*,\xi) - \nabla f(x^*)\|_*^2], \quad \mathcal{V}_g(x^*) := \mathbb{E}[(G(x^*,\xi) - g(x^*))^2].$$

The stochastic errors are denoted by

$$\varepsilon_{f'}^k := \nabla \mathcal{F}^k - \nabla f(x^k), \quad \varepsilon_g^k := \mathcal{G}_k - g(x^k).$$

Algorithm 2: Penalized mini-batch stochastic gradient method

1 Initialization: Choose the initial point $x^0 \in \mathcal{C}$, the batch size sequence $\{N_k\}$ and the stepsizes sequence $\{\alpha_k, \beta_k, \gamma_k\}$. Set $k = 0$ and $t_0 = 0$.

2 **while** *did not converge* **do**

3 Generate i.i.d. samples $\xi^k := \{\xi_1^k, \xi_2^k, \ldots, \xi_{N_k}^k\}$ and η^k of ξ, and compute

$$y^k = \mathcal{P}_{x^k}\left(\alpha_k \nabla \mathcal{F}^k\right), \quad x^{k+1} = \mathcal{P}_{y^k}\left(\gamma_k [t_{k+1}]_+ G'(x^k, \eta^k)\right).$$

where

$$\nabla \mathcal{F}^k := \frac{1}{N_k}\sum_{i=1}^{N_k} \nabla_x F(x^k, \xi_i^k), \quad t_{k+1} = (1-\beta_k)t_k + \beta_k \mathcal{G}_k,$$

and

$$\mathcal{G}_k := \frac{1}{N_k}\sum_{i=1}^{N_k} G(x^k, \xi_i^k).$$

4 Set $k \leftarrow k+1$.

We also define the following prox-mapping with full gradient as

$$\hat{y}^k := \mathcal{P}_{x^k}(\alpha_k \nabla f(x^k)),$$

which is important in the convergence analysis but not used in practice. It is worth mentioning that $\hat{y}^k \in \mathcal{F}_k$ and $y^k \notin \mathcal{F}_k$. From the Lipschitz continuity of the prox-mapping (Lemma 2.1), it follows

$$\|\hat{y}^k - y^k\| \leq \frac{\alpha_k}{\lambda}\|\varepsilon_{f'}^k\|_*. \tag{4.2}$$

In the following lemma, we estimate the bounds of the stochastic errors.

Lemma 4.1. *Suppose that Assumptions (A.3) and (A.4) are satisfied, then*

$$\mathbb{E}[\|\varepsilon_{f'}^k\|_*^2 | \mathcal{F}_k] \leq \frac{1}{N_k}\left[6\frac{(L_f^2 + \hat{L}_f^2)}{\lambda} V(x^k, x^*) + 3\mathcal{V}_{f'}(x^*)\right]$$

and

$$\mathbb{E}[(\varepsilon_g^k)^2 | \mathcal{F}_k] \leq \frac{1}{N_k}\left[\frac{6(C_g^2 + \hat{C}_g^2)}{\lambda} V(x^k, x^*) + 3\mathcal{V}_g(x^*)\right]$$

hold with probability 1.

Proof. By looking at (3.1), under Assumption (A.4), for every $x \in \mathcal{C}$ we have

$$\mathbb{E}[\|\nabla_x F(x,\xi) - \nabla f(x)\|_*^2] \leq \frac{6(L_f^2 + \hat{L}_f^2)}{\lambda} V(x, x^*) + 3\mathcal{V}_{f'}(x^*).$$

Taking into account the facts that $\{\xi_i^k\}$ are i.i.d. and $\mathbb{E}[\nabla_x F(x^k, \xi_i^k)|\mathcal{F}_k] = \nabla f(x^k)$ for all $i = 1, \ldots, N_k$, we get

$$\begin{aligned}
\mathbb{E}[\|\varepsilon_{f'}^k\|_*^2 | \mathcal{F}_k] &= \mathbb{E}[\|\tfrac{1}{N_k}\textstyle\sum_{i=1}^{N_k}(\nabla_x F(x^k, \xi_i^k) - \nabla f(x^k))\|_*^2 | \mathcal{F}_k] \\
&= \tfrac{1}{N_k^2}\textstyle\sum_{i=1}^{N_k} \mathbb{E}[\|\nabla_x F(x^k, \xi_i^k) - \nabla f(x^k)\|_*^2 | \mathcal{F}_k] \\
&= \tfrac{1}{N_k}\mathbb{E}[\|\nabla_x F(x^k, \xi_1^k) - \nabla f(x^k)\|_*^2 | \mathcal{F}_k] \\
&\leq \tfrac{1}{N_k}\left[6\tfrac{(L_f^2 + \hat{L}_f^2)}{\lambda} V(x^k, x^*) + 3\mathcal{V}_{f'}(x^*)\right]
\end{aligned}$$

with probability 1. The case for ε_g^k can be proved similarly. □

By noticing that

$$\mathbb{E}[[\mathcal{G}_k]_+^2|\mathcal{F}_k] \le \frac{1}{N_k}\sum_{i=1}^{N_k}\mathbb{E}[[G(x^k,\xi_i^k)]_+^2|\mathcal{F}_k] \le C_t^2, \quad \text{w.p.1},$$

we can get the following results in an analogous way as Lemma 3.1.

Lemma 4.2. *Under Assumption 5, it holds with probability 1 that, for all $k \ge 0$,*

(i) $\mathbb{E}[[t_{k+1}]_+^2|\mathcal{F}_k] \le C_t^2$;

(ii) $\mathbb{E}[\|\nabla\mathcal{F}^k\|_*^2|\mathcal{F}_k] \le C_f^2$;

(iii) $\mathbb{E}[\|[t_{k+1}]_+G'(x^k,\eta^k)\|_*^2|\mathcal{F}_k] \le C_t^2\hat{C}_g^2$;

(iv) $\mathbb{E}[\|x^{k+1}-x^k\|^2|\mathcal{F}_k] \le \frac{2C_f^2}{\lambda^2}\alpha_k^2 + \frac{2C_t^2\hat{C}_g^2}{\lambda^2}\gamma_k^2$.

Proof. In view of the proof of Lemma 3.1, items (i)-(iii) are obvious. From Lemma 2.2, it follows that

$$\begin{aligned}\|x^{k+1}-x^k\|^2 &\le 2\|x^{k+1}-y^k\|^2 + 2\|y^k-x^k\|^2\\ &\le \tfrac{4}{\lambda}V(x^{k+1},y^k) + \tfrac{4}{\lambda}V(y^k,x^k)\\ &\le \tfrac{2}{\lambda^2}\|\alpha_k\nabla\mathcal{F}^k\|_*^2 + \tfrac{2}{\lambda^2}\|\gamma_k[t_{k+1}]_+G'(x^k,\eta^k)\|_*^2,\end{aligned}$$

and hence item (iv) is proved. □

The core of the convergence analysis is to get the relationship between $V(x^{k+1},x^*)$ and $V(x^k,x^*)$. Substituting $x^+ = x^{k+1}$ and $u = x^*$ in (2.3), we have

$$V(x^{k+1},x^*) \le V(y^k,x^*) - V(y^k,x^{k+1}) - \langle\gamma_k[t_{k+1}]_+G'(x^k,\eta^k), x^{k+1}-x^*\rangle.$$

Therefore, to get the relationship, it is sufficient to establish the relation between $V(y^k,x^*)$ and $V(x^k,x^*)$ (which will be done in Lemma 4.3) and derive the bound of the last term above (which will be done in Lemma 4.4). The proofs of Lemmas 4.3 and 4.4 are provided in Appendix B.

Lemma 4.3. *Under Assumption 5, for every $x^* \in X^*$ and $u \in \mathcal{C}$, we get*

$$\begin{aligned}\mathbb{E}[V(y^k,x^*)|\mathcal{F}_k] \le{}& \left[1 + \tfrac{L_f^2}{2\lambda\eta}\cdot\tfrac{\alpha_k^2}{\gamma_k} + 6(L_f^2+\hat{L}_f^2)\tfrac{\alpha_k^2}{\lambda^2N_k}\right]V(x^k,x^*) - \alpha_k(f(u)-f^*) + 3\mathcal{V}_{f'}(x^*)\tfrac{\alpha_k^2}{\lambda N_k}\\ &+\tfrac{C_f^2}{4\eta}\tfrac{\alpha_k^2}{\gamma_k} + (\alpha_kL_f + 3\eta\gamma_k)\|u-x^k\|^2 - \tfrac{\lambda}{4}\mathbb{E}[\|y^k-x^k\|^2|\mathcal{F}_k]\end{aligned}$$

with probability 1, where $\eta > 0$ is arbitrary and γ_k is chosen such that $2\eta\gamma_k \le \lambda/4$.

Following the same line of argument as in Lemma 3.2, we can derive the conditional expectation bound for $\langle\gamma_k[t_{k+1}]_+G'(x^k,\eta^k), x^k-x^*\rangle$.

Lemma 4.4. *Let x^* be an arbitrary point in X^*. Under Assumption 5, it holds with probability 1 that,*

$$\begin{aligned}\mathbb{E}[-\langle\gamma_k[t_{k+1}]_+G'(x^k,\eta^k), x^k-x^*\rangle|\mathcal{F}_k] \le{}& -\tfrac{\gamma_k}{2}\phi(x^k) + [C_g^2\tfrac{\gamma_k^2}{2\lambda\beta_k} + 6(C_g^2+\hat{C}_g^2)\tfrac{\beta_k^2}{\lambda N_k}]V(x^k,x^*)\\ &+3\mathcal{V}_g(x^*)\tfrac{\beta_k^2}{N_k} + C_g^2\|x^k-x^{k-1}\|^2 + \beta_k(g(x^{k-1})-t_k)^2.\end{aligned}$$

In the following we give a recursive relation of the sequence $\{(g(x^k)-t_{k+1})^2\}$. See Appendix B for its proof.

Lemma 4.5. *Let Assumption 5 hold. Then, we have*

$$\begin{aligned}&\mathbb{E}[(g(x^k)-t_{k+1})^2|\mathcal{F}_k]\\ &\le (1-\beta_k)(g(x^{k-1})-t_k)^2 + \tfrac{12\beta_k^2}{\lambda N_k}(C_g^2+\hat{C}_g^2)V(x^k,x^*) + \tfrac{6\mathcal{V}_g(x^*)\beta_k^2}{N_k} + \tfrac{C_g^2}{\beta_k}\|x^k-x^{k-1}\|^2\end{aligned}$$

with probability 1.

4.2 Almost sure convergence

The almost sure convergence of Algorithm 2 is investigated under the following stepsize assumption.

Assumption 6. *Let the diminishing stepsize sequences $\{\alpha_k\}$, $\{\beta_k\}$ and $\{\gamma_k\}$ be such that*

$$0 < \beta_k < 1, \quad \sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \left(\frac{\gamma_k^2}{\beta_{k+1}} + \frac{\alpha_k^2}{\gamma_k} + \frac{\alpha_k^2 + \beta_k^2}{N_k} \right) < \infty.$$

In addition, we assume $\frac{5}{4}\lambda C_{eb} \geq \gamma_k \geq 5(1 + L_f)C_{eb}\alpha_k$.

We mention that, when Assumption 6 holds true, it follows that

$$\sum_{k=0}^{\infty} \gamma_k^2 < \infty, \quad \sum_{k=1}^{\infty} \frac{\alpha_{k-1}^2}{\beta_k} < \infty.$$

Moreover, we let $\eta = \frac{1}{10C_{eb}}$ in the rest of this section, then Assumption 6 implies that

$$2\eta\gamma_k \leq \frac{\lambda}{4}, \qquad -\frac{\gamma_k}{2C_{eb}} + 3\eta\gamma_k + \alpha_k L_f \leq -\alpha_k. \tag{4.3}$$

Let us give a few remarks on Assumption 6, which show that the stepsizes choice is more flexible when using mini-batch stochastic gradient. If the batch size is constant, i.e., $N_k \equiv m$, then the stepsizes should satisfy

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \beta_k^2 < \infty, \quad \sum_{k=1}^{\infty} \frac{\gamma_{k-1}^2}{\beta_k} < \infty, \quad \sum_{k=0}^{\infty} \frac{\alpha_k^2}{\gamma_k} < \infty.$$

Otherwise, if the batch size is dynamically increased such that $\sum_k 1/N_k < \infty$ (which implies $N_k \to \infty$), then we can select the stepsizes as

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \quad \beta_k \equiv \beta < 1, \quad \sum_{k=1}^{\infty} \gamma_{k-1}^2 < \infty, \quad \sum_{k=0}^{\infty} \frac{\alpha_k^2}{\gamma_k} < \infty.$$

In this case, we could set the stepsizes $\alpha_k, \beta_k, \gamma_k$ much larger than the previous constant batch size case.

In what follows we present two critical coupled recursive relations in order to invoke the coupled supermartingale convergence theorem. See Appendix B for the detailed proof.

Lemma 4.6. *Suppose that Assumptions 2, 5 and 6 hold true. Let x^* be any given optimal solution to problem (1.1). Then, we have*

$$\mathbb{E}[\theta_{k+1}|\mathcal{F}_k] \leq (1 + a_k)\theta_k - u_k + d_k\zeta_k + \mu_k, \quad w.p.1 \tag{4.4}$$

and

$$\mathbb{E}[\zeta_{k+1}|\mathcal{F}_k] \leq (1 - d_k)\zeta_k - \bar{u}_k + b_k\theta_k + \nu_k, \quad w.p.1 \tag{4.5}$$

where $\theta_k := V(x^k, x^*)$, $\zeta_k := (g(x^{k-1}) - t_k)^2$, $d_k := \beta_k$, $\bar{u}_k := 0$,

$$a_k := \tfrac{L_f^2}{2\lambda\eta} \cdot \tfrac{\alpha_k^2}{\gamma_k} + 6(L_f^2 + \hat{L}_f^2)\tfrac{\alpha_k^2}{\lambda^2 N_k} + \tfrac{C_g^2}{2\lambda} \cdot \tfrac{\gamma_k^2}{\beta_k} + 6(C_g^2 + \hat{C}_g^2)\tfrac{\beta_k^2}{\lambda N_k},$$
$$u_k := \alpha_k(f(\bar{x}^k) - f^* + \|x^k - \bar{x}^k\|^2),$$
$$b_k := \tfrac{12\beta_k^2}{\lambda N_k}(C_g^2 + \hat{C}_g^2), \quad \nu_k := \tfrac{6\mathcal{V}_g(x^*)\beta_k^2}{N_k} + \tfrac{C_g^2}{\beta_k}\|x^k - x^{k-1}\|^2$$

and

$$\mu_k := \frac{C_f^2}{4\eta} \cdot \frac{\alpha_k^2}{\gamma_k} + \frac{2\hat{C}_g^2 C_t^2}{\lambda}\gamma_k^2 + 3\mathcal{V}_{f'}(x^*)\frac{\alpha_k^2}{\lambda N_k} + 3\mathcal{V}_g(x^*)\frac{\beta_k^2}{N_k} + C_g^2\|x^k - x^{k-1}\|^2.$$

We now present the almost sure convergence of the mini-batch algorithm for solving problem (1.1).

Theorem 4.7. *Let Assumptions 2, 5 and 6 hold. Then the sequence $\{x^k\}$ generated by Algorithm 2 converges almost surely to a point in X^*.*

Proof. In order to apply the coupled martingale convergence theorem, we first verify that the conditions of Theorem 2.5 are satisfied. Indeed, by using the same argument for deriving (3.14), from item (iv) in Lemma 4.2 we obtain

$$\sum_{k=1}^{\infty} \beta_k^{-1} \|x^k - x^{k-1}\|^2 < \infty, \quad \text{w.p.1.}$$

Then, from the definitions of a_k, b_k, μ_k, ν_k, it is easy to see that

$$\sum_k a_k < \infty, \quad \sum_k b_k < \infty, \quad \sum_k \mu_k < \infty, \quad \sum_k \nu_k < \infty, \quad \text{w.p.1.}$$

Take x^* be an arbitrary point in X^*. Let us now apply the coupled supermartingale convergence theorem to (4.4) and (4.5). Then we obtain that $\{\|x^k - x^*\|^2\}$ and $\{(g(x^{k-1}) - t_k)^2\}$ converges almost surely to finite random variables, and

$$\sum_{k=1}^{\infty} \alpha_k (f(\bar{x}^k) - f^* + \|x^k - \bar{x}^k\|^2) < \infty, \quad \text{w.p.1.}$$

Hence, $\{x^k\}$ is bounded with probability 1. Moreover, as $\sum_k \alpha_k = \infty$, we obtain

$$\liminf_{k\to\infty} (f(\bar{x}^k) - f^* + \|x^k - \bar{x}^k\|^2) = 0, \quad \text{w.p.1.}$$

The rest of the proof can be proceeded in the same way as that of Theorem 3.6. □

We remark that, if we additionally assume that $\sum_k \beta_k = \infty$, it can also be shown that the sequence $\{(g(x^k) - t_k)^2\}$ converges almost surely to zero. Indeed, in the proof of Theorem 4.7 we have shown that $\{(g(x^{k-1}) - t_k)^2\}$ converges almost surely to a finite random variable. From Theorem 2.5 we have $\sum_k d_k \zeta_k < \infty$ with probability 1, i.e., $\sum_k \beta_k (g(x^k) - t_k)^2 < \infty$ with probability 1, which, together with $\sum_k \beta_k = \infty$, further implies that $\liminf_{k\to\infty} (g(x^k) - t_k)^2 = 0$ with probability 1. Hence, we get the claim.

4.3 Expected convergence rates

Let $\hat{x}^K$ and $\hat{u}^K$ be the averaged iterates defined by (3.16). In the following theorem we present the convergence rates of Algorithm 2 with two different stepsize choices.

Theorem 4.8. *Let Assumptions 2 and 5 hold.*

(i) If the batch size $N_k \equiv m$ is constant and the stepsizes are chosen as in (3.10), then

$$\mathbb{E}[f(\hat{u}^K) - f^* + \|\hat{x}^K - \hat{u}^K\|^2] \approx \mathcal{O}\left(\frac{1}{K^{1/8}}\right).$$

(ii) If the batch size is set as $N_k = m \cdot \lceil k^{1+\varepsilon} \rceil$ and the stepsizes are chosen as

$$\beta_k \equiv \beta, \quad \alpha_k = \alpha \cdot k^{-(\frac{3}{4}+\varepsilon)}, \quad \gamma_k = \gamma \cdot k^{-(\frac{1}{2}+\varepsilon)}, \tag{4.6}$$

where m, α, β, γ are proper constants, then

$$\mathbb{E}[f(\hat{u}^K) - f^* + \|\hat{x}^K - \hat{u}^K\|^2] \approx \mathcal{O}\left(\frac{1}{K^{1/4}}\right).$$

Proof. See Appendix B for the detailed proof. □

4.4 Convergence under restricted strong convexity

To conclude this section, we consider the restricted strongly convex case. We now take the Euclidean distance $V(x,y) = \frac{1}{2}\|x-y\|_2^2$.

Theorem 4.9. *Let Assumptions 4 and 5 hold. Let $\hat{x}^K$ be defined as in (3.16).*

(i) If the batch size $N_k \equiv m$ is constant and the stepsizes are chosen as in (3.19), then

$$\mathbb{E}[\|\hat{x}^K - x^*\|_2^2] \approx \mathcal{O}\left(\frac{1}{K^{1/4}}\right).$$

(ii) If the batch size is set as $N_k = m \cdot \lceil k^{1+\varepsilon} \rceil$ and the stepsizes are chosen as

$$\beta_k \equiv \beta, \quad \alpha_k = \alpha \cdot k^{-(\frac{1}{2}+2\varepsilon)}, \quad \gamma_k = \gamma \cdot k^{-(\frac{1}{2}+\varepsilon)}, \tag{4.7}$$

where m, α, β, γ are proper constants and $\varepsilon \in (0, 1/4)$, then

$$\mathbb{E}[\|\hat{x}^K - x^*\|_2^2] \approx \mathcal{O}\left(\frac{1}{K^{1/2}}\right).$$

(iii) In both settings of item (i) and item (ii), we have that $\{\|x^k - x^*\|_2^2\}$ and $\{(g(x^k) - t_{k+1})^2\}$ converge almost surely to zero.

Proof. See Appendix B for the detailed proof. □

5 Mini-batch algorithm for problems with multiple expectation constraints

In this section, we consider a variant of the mini-batch PSG to solve the following stochastic program with multiple expectation constraints,

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & f(x) \\ \text{s.t.} & g_i(x) := \mathbb{E}[G_i(x,\xi)] \le 0, \quad i \in \mathcal{I}, \\ & x \in \mathcal{C}. \end{array} \tag{5.1}$$

Here, f, ξ and $\mathcal{C}$ are defined as same as in problem (1.1), and $G_i : \mathbb{R}^n \times \Xi \to \mathbb{R}$ for each $i \in \mathcal{I}$. Let us denote $\phi_i(x) := [g_i(x)]_+^2$, $i \in \mathcal{I}$. Again, let X^* be the set of optimal solutions, f^* be the optimal objective value, and $\Phi := \{x \in \mathcal{C} : \phi_i(x) = 0, i \in \mathcal{I}\}$ be the feasible set.

The detail of the proposed algorithm for solving problem (5.1) is described in Algorithm 3. At each iteration, it contains two main steps: $y^k = \mathcal{P}_{x^k}\left(\alpha_k \nabla \mathcal{F}^k\right)$ and $x^{k+1} = \mathcal{P}_{y^k}\left(\gamma_k d^k\right)$. The first step is exactly the same as that in Algorithm 2. We now give some explanations for the second step. Instead of considering the whole set $\mathcal{I}$, we first randomly select a subset $\mathcal{I}_k$ with $1 \le |\mathcal{I}_k| \le |\mathcal{I}|$ such that

$$\mathbb{E}\left[\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k} g_i(x)\right] = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} g_i(x).$$

Then, we perform the second step which can be rewritten as

$$x^{k+1} = \mathcal{P}_{y^k}\left(\frac{\gamma_k}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k} [t_{k+1}^i]_+ G_i'(x^k,\eta^k)\right), \tag{5.3}$$

where $t_{k+1}^i, i \in \mathcal{I}$ are updated by (5.2). Note that, if there is only a single constraint in problem (5.1), i.e., $|\mathcal{I}| = 1$, then obviously $\mathcal{I}_k \equiv \mathcal{I}$ and Algorithm 3 is reduced to Algorithm 2. Moreover, we can see

Algorithm 3: Multiple penalized mini-batch stochastic gradient method

1 Initialization: Choose the initial point $x^0 \in \mathcal{C}$, the batch size sequence $\{N_k\}$, the constant $M \leq |\mathcal{I}|$ and the stepsizes sequence $\{\alpha_k, \beta_k, \gamma_k\}$ with $\beta_k < 1$. Set $k = 0$ and $t_0^i = 0$ for all $i \in \mathcal{I}$.

2 **while** *did not converge* **do**

3 Generate i.i.d. samples $\xi^k := \{\xi_1^k, \xi_2^k, \ldots, \xi_{N_k}^k\}$. Set

$$\nabla \mathcal{F}^k := \frac{1}{N_k} \sum_{j=1}^{N_k} \nabla_x F(x^k, \xi_j^k),$$

and

$$\mathcal{G}_k^i := \frac{1}{N_k} \sum_{j=1}^{N_k} G_i(x^k, \xi_j^k), \quad t_{k+1}^i = (1-\beta_k) t_k^i + \beta_k \mathcal{G}_k^i, \quad \forall i \in \mathcal{I}. \tag{5.2}$$

4 Generate i.i.d. sample η^k of ξ and select randomly an index set $\mathcal{I}_k \subset \mathcal{I}$ with $|\mathcal{I}_k| = M$. Set

$$d^k := \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} d^{k,i}, \quad \text{where } d^{k,i} := [t_{k+1}^i]_+ G_i'(x^k, \eta^k), \quad \forall i \in \mathcal{I}_k.$$

5 Then, compute

$$y^k = \mathcal{P}_{x^k}\left(\alpha_k \nabla \mathcal{F}^k\right), \quad x^{k+1} = \mathcal{P}_{y^k}\left(\gamma_k d^k\right).$$

6 Set $k \leftarrow k+1$.

that, in the second step (5.3), we only require $\{t_{k+1}^i, i \in \mathcal{I}_k\}$ but not the full set $\{t_{k+1}^i, i \in \mathcal{I}\}$. Therefore, from a practical perspective, it is better to replace (5.2) with the following cheaper update rule:

$$\begin{cases} \mathcal{G}_k^i = \frac{1}{N_k} \sum_{j=1}^{N_k} G_i(x^k, \xi_j^k), \quad t_{k+1}^i = (1-\beta_k) t_k^i + \beta_k \mathcal{G}_k^i, & i \in \mathcal{I}_k, \\ t_{k+1}^i = t_k^i, & i \in \mathcal{I} \backslash \mathcal{I}_k. \end{cases} \tag{5.4}$$

However, by doing this, it is unclear how to guarantee the global convergence since t_{k+1}^i may not be a good approximation to $g_i(x^k)$ anymore (even for $i \in \mathcal{I}_k$) when k is large.

Note that, there are three random items ξ^k, η^k and $\mathcal{I}_k$ at each iteration in Algorithm 3. Therefore, we shall study the convergence of the stochastic process $\{x^k\}$ with respect to the following three filtrations

$$\mathcal{F}_k := \sigma(\xi^0, \eta^0, \mathcal{I}_0, \ldots, \xi^{k-1}, \eta^{k-1}, \mathcal{I}_{k-1}), \quad \hat{\mathcal{F}}_k := \sigma(\xi^0, \eta^0, \mathcal{I}_0, \ldots, \xi^{k-1}, \eta^{k-1}, \mathcal{I}_{k-1}, \xi^k)$$

and

$$\tilde{\mathcal{F}}_k := \sigma(\xi^0, \eta^0, \mathcal{I}_0, \ldots, \xi^{k-1}, \eta^{k-1}, \mathcal{I}_{k-1}, \xi^k, \eta^k).$$

It is not difficult to verify that $x^k \in \mathcal{F}_k$, $y^k \in \hat{\mathcal{F}}_k$, $t_{k+1}^i \in \hat{\mathcal{F}}_k$ and $d^{k,i} \in \tilde{\mathcal{F}}_k$. The following relation will be used several times in the analysis,

$$\mathbb{E}[\cdot|\mathcal{F}_k] = \mathbb{E}[\mathbb{E}[\cdot|\hat{\mathcal{F}}_k]\mathcal{F}_k] = \mathbb{E}[\mathbb{E}[\mathbb{E}[\cdot|\tilde{\mathcal{F}}_k]\hat{\mathcal{F}}_k]\mathcal{F}_k].$$

For every $x^* \in X^*$, we denote the following variances by

$$\mathcal{V}_{f'}(x^*) := \mathbb{E}[\|\nabla_x F(x^*, \xi) - \nabla f(x^*)\|_*^2], \quad \mathcal{V}_g(x^*) := \max_{i \in \mathcal{I}} \mathbb{E}[(G_i(x^*, \xi) - g_i(x^*))^2].$$

Let us also define the following stochastic errors:

$$\varepsilon_{f'}^k := \nabla \mathcal{F}^k - \nabla f(x^k), \quad \varepsilon_{g_i}^k := \mathcal{G}_k^i - g_i(x^k), \quad \forall i \in \mathcal{I}.$$

The convergence analysis of Algorithm 3 is carried out very similarly as that in the preceding section. The following two assumptions are made in this section.

Assumption 7. *We assume:*

(B.1) *For almost every $\omega \in \Omega$, the function $F(\cdot,\xi) \equiv F(\cdot,\xi(\omega))$ is convex and differentiable on $\mathcal{C}$, and the functions $G_i(\cdot,\xi) \equiv G_i(\cdot,\xi(\omega)), i \in \mathcal{I}$ are convex on $\mathcal{C}$.*

(B.2) *There exist constants $C_f > 0$ and $C_t > 0$ such that for all $x \in \mathcal{C}$,*

$$\mathbb{E}[\|\nabla_x F(x,\xi)\|_*^2] \le C_f^2, \quad \mathbb{E}[[G_i(x,\xi)]_+^2] \le C_t^2, \quad \forall i \in \mathcal{I}.$$

(B.3) *There exists a measurable function $c_g : \Xi \to \mathbb{R}_+$ with $C_g := \mathbb{E}[c_g(\xi)]$ and $\hat{C}_g := \sqrt{\mathbb{E}[c_g^2(\xi)]} < \infty$, such that for all $x, y \in \mathcal{C}$ and for almost every $\omega \in \Omega$,*

$$|G_i(x,\xi) - G_i(y,\xi)| \le c_g(\xi)\|x - y\|, \quad \forall i \in \mathcal{I}.$$

(B.4) *There exists a measurable functions $l_f : \Xi \to \mathbb{R}_+$ with $L_f := \mathbb{E}[l_f(\xi)]$ and $\hat{L}_f := \sqrt{\mathbb{E}[l_f^2(\xi)]} < \infty$, such that for all $x, y \in \mathcal{C}$ and for almost every $\omega \in \Omega$,*

$$\|\nabla_x F(x,\xi) - \nabla_x F(y,\xi)\| \le l_f(\xi)\|x - y\|.$$

Assumption 8. *We suppose that, for any given $x^* \in X^*$, there exists a constant $C_{eb} > 0$ such that for all $k \ge 0$,*

$$\|x^k - \bar{x}^k\|^2 \le 2C_{eb} \cdot \mathbb{E}\left[\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\langle [g_i(x^k)]_+ G_i'(x^k,\eta^k), x^k - x^*\rangle | \mathcal{F}_k\right], \quad w.p.1, \tag{5.5}$$

where $\bar{x}^k$ is the projection of x^k onto the feasible set Φ.

By noticing that

$$\mathbb{E}\left[\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\langle [g_i(x^k)]_+ G_i'(x^k,\eta^k), x^k - x^*\rangle | \tilde{\mathcal{F}}_k\right] = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\langle [g_i(x^k)]_+ G_i'(x^k,\eta^k), x^k - x^*\rangle$$

and

$$\mathbb{E}\left[\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\langle [g_i(x^k)]_+ G_i'(x^k,\eta^k), x^k - x^*\rangle | \hat{\mathcal{F}}_k\right] = \frac{1}{2|\mathcal{I}|}\sum_{i\in\mathcal{I}}\langle \phi_i'(x^k), x^k - x^*\rangle \ge \frac{1}{2|\mathcal{I}|}\sum_{i\in\mathcal{I}}\phi_i(x^k)$$

hold with probability 1, we get that Assumption 8 is slightly weaker than the following global error bound condition,

$$C_{eb} \cdot \phi_i(x^k) \ge \mathrm{dist}^2(x^k, \Phi), \quad \forall i \in \mathcal{I}.$$

We present the almost sure convergence of Algorithm 3 in Theorem 5.1 and give its proof in Appendix C.

Theorem 5.1. *Let Assumptions 7 and 8 hold. Suppose that the stepsizes and the batch sizes satisfy Assumption 6. Then the sequence $\{x^k\}$ generated by Algorithm 3 converges almost surely to a point in X^*.*

By using a similar argument as in Theorem 4.8 and Theorem 4.9, we can establish the same convergence rates of Algorithm 3. We omit the details to avoid repetition.

6 Numerical experiments

In this section, we conduct numerical experiments on four applications: (i) Neyman-Pearson classification, (ii) portfolio optimization with Conditional Value-at-Risk constraint, (iii) chance constrained program and (iv) second-order stochastic dominance constrained portfolio problem. All numerical experiments are carried out using MATLAB R2019a on a desktop computer with Intel(R) Core(TM) i7-8700T 2.40GHz and 16GB memory. All reported time is wall-clock time in seconds.

Let us recall that, we have proposed at least three types of penalized stochastic gradient methods (basic PSG, mini-batch PSG with two different stepsizes) for solving problem (1.1). In our experience, the mini-batch method with constant batch size ($N_k \equiv m$ with $m > 1$ in Algorithm 2) usually performs better than other variants, such as basic PSG (Algorithm 1 or $N_k \equiv 1$ in Algorithm 2) and adaptive batch sizes PSG ($N_k = \lceil k^{1+\varepsilon} \rceil$ in Algorithm 2). This phenomenon is also observed in various stochastic gradient-type algorithms for solving optimization problems without expectation constraints. Therefore, in the following numerical experiments, we pay our attention to present the performance of the mini-batch algorithm with constant batch size, which is simply denoted by "PSG". The stepsizes $\alpha_k, \beta_k, \gamma_k$ are set as in (3.10) or (3.19) dependently.

6.1 Neyman-Pearson classification

Unlike conventional binary classification, the Neyman-Pearson (NP) classification paradigm is developed to learn a classifier by minimizing type II error with type I error being below a user-specified level $\alpha \in (0, 1)$, see [57] and references therein. For a classifier h to predict 1 and -1, let us define the type I error (misclassifying class -1 as 1) and type II error (misclassifying class 1 as -1) respectively by

$$\text{type I error} := \mathbb{E}[\varphi(-bh(a))|b = -1], \quad \text{type II error} := \mathbb{E}[\varphi(-bh(a))|b = 1],$$

where φ is some merit function. For a given class $\mathcal{H}$ of classifiers, the NP classification is to solve the following problem

$$\begin{array}{ll} \min\limits_{h\in\mathcal{H}} & \mathbb{E}[\varphi(-bh(a))|b = 1] \\ \text{s.t.} & \mathbb{E}[\varphi(-bh(a))|b = -1] \le \alpha. \end{array}$$

In what follows, we consider its empirical risk minimization counterpart. Suppose that a labelled training dataset $\{a_i\}_{i=1}^N$ consists of the positive set $\{a_i^0\}_{i=1}^{N_0}$ and the negative set $\{a_i^1\}_{i=1}^{N_1}$. The associated empirical NP classification problem is

$$\begin{array}{ll} \min\limits_{x} & f(x) := \frac{1}{N_0} \sum_{i=1}^{N_0} \ell(x^T a_i^0) \\ \text{s.t.} & g(x) := \frac{1}{N_1} \sum_{i=1}^{N_1} \ell(-x^T a_i^1) - \alpha \le 0, \\ & x \in \mathcal{C} := \{x \in \mathbb{R}^n : \|x\|_2 \le \iota\}. \end{array}$$

Here, the constraint $\|x\|_2 \le \iota$ is to avoid overfitting, and $\ell(\cdot)$ is a loss function, e.g., logistic loss $\ell(y) := \log(1 + \exp(-y))$ or smoothed hinge loss

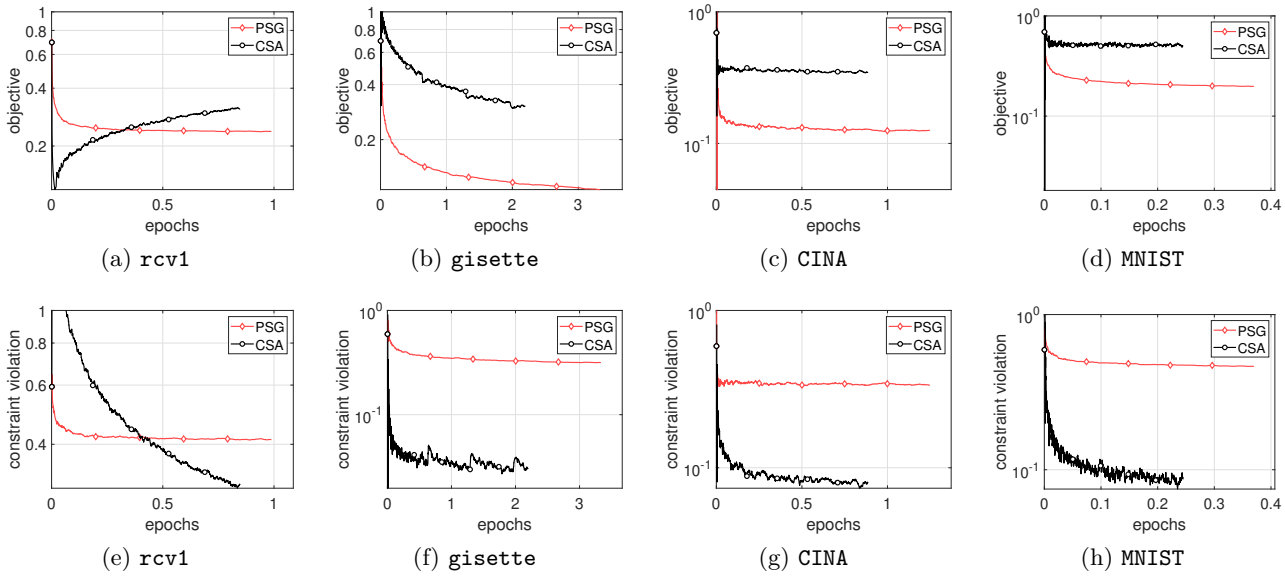
$$\ell(y) := \begin{cases} \frac{1}{2} - y, & \text{if } x <= 0, \\ \frac{1}{2}(y-1)^2, & \text{if } 0 < x \le 1, \\ 0, & \text{if } x > 1. \end{cases}$$

In this numerical experiment, we intend to show the performance differences between PSG and CSA [26]. The datasets tested are summarized in Table 1. We choose $x^0 = 0$ as the initial point and use the following parameters $\alpha = 0.1$ and $\iota = 5$. We let both two methods run 1,000 iterations. The constant batch size is set to $N_k = 10$. The results are averaged over 50 independent runs.

In Figure 1 and Figure 2, we show the results for the four datasets with logistic loss and hinge loss, respectively. Changes of objective value and expectation constraint value are reported with respect to *epochs*. We observe that PSG and CSA have similar convergence speed. However, the directions of the changes of objective value and constraint value are very different. For CSA, it prefers to decrease the constraint value, while the objective value may increase. For PSG, the constraint value and the

Table 1: Datasets used in Neyman-Pearson classification

Dataset	Data N	Variable n	Density	Reference
`rcv1`	20242	47236	0.16%	[29]
`gisette`	6000	5000	12.97%	[17]
`CINA`	16033	132	29.56%	[64]
`MNIST`	60000	784	19.12%	[27]



(a) `rcv1` (b) `gisette` (c) `CINA` (d) `MNIST`

(e) `rcv1` (f) `gisette` (g) `CINA` (h) `MNIST`

Figure 1: Results of PSG and CSA for Neyman-Pearson classification problem with *logistic* loss. (Average of 50 runs).

objective value always decrease simultaneously. We also witness this behavior in other experiments. In the subsequent experiments, we will not compare PSG with CSA anymore and focus on the comparison with other non-SA-type methods.

6.2 Portfolio optimization with Conditional Value-at-Risk constraint

We now consider the following Conditional Value-at-Risk (CVaR) constrained portfolio optimization problem

$$\begin{aligned} \min_{x} \quad & \mathbb{E}[-\xi^T x] \\ \text{s.t.} \quad & \mathrm{CVaR}_{1-\alpha}(-\xi^T x) \le \beta, \\ & x \in \mathcal{C}_x := \{x \in \mathbb{R}^n : \textstyle\sum_{i=1}^n x_i = 1, \quad x \ge 0\}, \end{aligned}$$

where ξ is the random return of assets, $1 > \alpha > 0$ is a probability threshold,

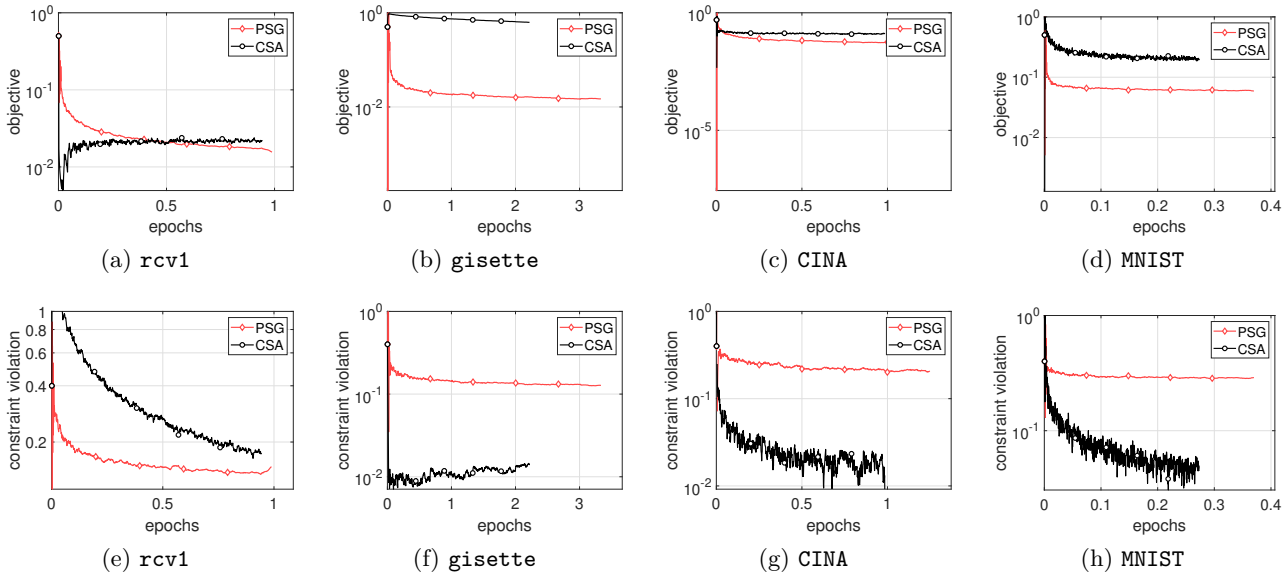
$$\mathrm{CVaR}_{1-\alpha}(z) := \inf_{\tau \in \mathbb{R}} \left\{ \tau + \frac{1}{\alpha} \mathbb{E}\left[[z-\tau]_+\right] \right\}$$

is a convex risk measure introduced in the pioneering work [49] and $\beta > 0$ is a user-specified loss tolerance. It is shown in [50] that this portfolio problem can be rewritten as follows:

$$\begin{aligned} \min_{x,\tau} \quad & \mathbb{E}[-\xi^T x] \\ \text{s.t.} \quad & \tau + \tfrac{1}{\alpha}\mathbb{E}[[-\xi^T x - \tau]_+] - \beta \le 0, \\ & x \in \mathcal{C}_x. \end{aligned}$$

Suppose that we know the mean and the covariance: $\bar{\xi} := \mathbb{E}[\xi]$ and $\Sigma := \mathrm{Cov}(\xi)$. Let $R > 0$ be the desired return and $Y := \{x \in \mathcal{C}_x : \xi^T x \ge R\}$. Then, in [25] the authors claim that τ is restricted to

$$\mathcal{C}_\tau := \left[\underline{\mu} - \bar{\sigma}\sqrt{\frac{\alpha}{1-\alpha}}, \ \bar{\mu} + \bar{\sigma}\sqrt{\frac{\alpha}{1-\alpha}} \right],$$



(a) `rcv1` (b) `gisette` (c) `CINA` (d) `MNIST`

(e) `rcv1` (f) `gisette` (g) `CINA` (h) `MNIST`

Figure 2: Results of PSG and CSA for Neyman-Pearson classification problem with *hinge* loss. (Average of 50 runs).

where

$$\underline{\mu} := \min_{y\in Y} \left\{-\bar{\xi}^T y\right\}, \quad \bar{\mu} := \max_{y\in Y} \left\{-\bar{\xi}^T y\right\}, \quad \bar{\sigma} = \max_i \Sigma_{ii}.$$

Therefore, it is equivalent to solve the following problem

$$\begin{array}{ll} \min\limits_{x,\tau} & f(x) := \mathbb{E}[-\xi^T x] \\ \text{s.t.} & g(x) := \tau + \frac{1}{\alpha}\mathbb{E}[[-\xi^T x - \tau]_+] - \beta \le 0, \\ & (x,\tau) \in \mathcal{C} := \mathcal{C}_x \times \mathcal{C}_\tau, \end{array} \tag{6.1}$$

which is in the form of (1.1). The test problems in this experiment are generated as follows.

Problem 6.1. The return ξ with its mean, standard deviation and correlation is generated by using

load('port5.mat','Correlation', 'stdDev_return','mean_return')

in MATLAB. The number of assets is $n = 225$. The desired return is set as $R = 0.002$.

Problem 6.2. The mean of the return ξ is randomly generated by

rng(0,'twister'); a = -0.1; b = 0.4; mean_return = a + (b-a).*rand(n,1);

while the standard deviation is randomly generated by

a = -0.08; b = 0.6; mean_return = a + (b-a).*rand(n,1);

and the correlation is generated by using gallery('randcorr',n) in MATLAB. The desired return is set as $R = 0.15$.

To get the numerical problems in the formulation of (6.1) by using the data in Problem 6.1 and Problem 6.2, we first derive the covariance matrix by

Covariance = Correlation .* (stdDev_return * steDev_return')

in MATLAB, and then compute the required elements to obtain $\mathcal{C}_\tau$. In both two problems, the probability threshold is set to $\alpha = 0.05$, and the tolerance loss β is set by an approximate upper bound of $g(x)$.

We compare PSG with SAA approach (denoted by "SAA-CVX"). In PSG, the batch size is set to 100 constantly at each iteration. In SAA-CVX, we use the package CVX[3] with solver Gurobi[4] to solve the following SAA problem

$$\begin{array}{ll}\min\limits_{x,\tau,z} & -\langle\bar{\xi},x\rangle\\ \text{s.t.} & \frac{1}{N}\sum_{j=1}^{N} y_j \le \alpha(\beta-\tau),\\ & z_j \ge -\xi_j^T x-\tau, \quad z_j\ge 0, \quad j=1,\ldots,N,\\ & (x,\tau)\in\mathcal{C}:=\mathcal{C}_x\times\mathcal{C}_\tau,\end{array}$$

where $\{\xi_1,\ldots,\xi_N\}$ are i.i.d. samples of ξ.

The initial point x^0 is set to $(1/n,1/n,\ldots,1/n)^T$. We record the number of iterations (by "iter"), the objective function value of f (by "obj"), the constraint function value of g (by "cons") and the elapsed time. The results are averaged over 20 independent runs. In addition, obj_{max} and obj_{min} denote the maximum and the minimum of objective values in 20 runs, respectively.

The results for Problem 6.1 and Problem 6.2, respectively, are shown in Table 2 and Table 3. The approximate optimal objective value of Problem 6.1 is -2.8e-3. We summarize the observations drawn from these two tables as follows. In view of cputime, PSG obviously outperforms SAA-CVX. PSG always converges to an approximate solution with a few (500 or 1000) iterations. However, it can be seen that the accuracy is not improved apparently as the iteration process goes further owing to the drawback of SA-type methods. Looking at the maximum and the minimum of the objective values, when N is small, the objective values of SAA-CVX vary widely (and hence not stable). On the other hand, as the sample size N increases, the range (maximum minus minimum) of the objective values becomes small but the cputime grows rapidly.

Table 2: Results of Problem 6.1. The approx. optimal obj. value is -2.8e-3.

PSG						SAA-CVX				
iter	obj	obj_{max}	obj_{min}	cons	time	N	obj	obj_{max}	obj_{min}	time
100	-2.29e-3	-2.02e-3	-2.41e-3	1.3e-2	0.16	1000	-2.77e-3	-2.25e-3	-3.31e-3	0.32
500	-2.63e-3	-2.48e-3	-2.76e-3	1.4e-2	0.59	5000	-2.78e-3	-2.52e-3	-2.99e-3	3.21
1000	-2.76e-3	-2.67e-3	-2.83e-3	1.3e-2	1.10	10000	-2.77e-3	-2.61e-3	-2.90e-3	10.9
2000	-2.83e-3	-2.69e-3	-2.90e-3	1.4e-2	2.14	15000	-2.80e-3	-2.66e-3	-2.99e-3	23.9
3000	-2.89e-3	-2.79e-3	-2.95e-3	1.4e-2	3.22	20000	-2.77e-3	-2.62e-3	-2.87e-3	41.7

6.3 Chance constrained programs

In this subsection, we consider the following (joint) chance constrained program (CCP)

$$\begin{array}{ll}\min & \mathbb{E}[F(x,\xi)]\\ \text{s.t.} & \mathbb{P}\{G_i(x,\xi)\le 0,\ i=1,\ldots,m\}\ge 1-\alpha,\\ & x\in\mathcal{C}\subseteq\mathbb{R}^n.\end{array}$$

Here $G_i:\mathbb{R}^n\times\Xi\to\mathbb{R}, i=1,\ldots,m$ are random functions, and $\alpha>0$ is a probability bound that is usually set as 0.01, 0.05 or 0.1. Obviously, if we let $G(x,\xi):=\max_{1\le i\le m}G_i(x,\xi)$, the chance constraint can be rewritten as $\mathbb{P}\{G(x,\xi)\le 0\}\ge 1-\alpha$. This constraint can also be transformed to an expectation constraint as

$$\mathbb{E}[\mathbb{1}_{[0,\infty)}(G(x,\xi))]\le\alpha.$$

Problem (CCP) can now be cast into the form of (1.1), but a couple of challenges still remain to handle.

[3]See `http://cvxr.com/cvx`.

[4]See `http://www.gurobi.com`.

Table 3: Results of Problem 6.2 with $n = 200, 1000, 2000, 5000$.

PSG				SAA-CVX		
$n = 200$						
iter	obj	cons	time	N	obj	time
500	-1.7264e-1	1.4819e-0	0.86	2000	-2.5911e-1	3.58
1000	-2.1888e-1	1.9670e-1	1.69	5000	-2.7243e-1	13.75
2000	-2.6990e-1	3.4623e-2	3.36	10000	-2.7602e-1	42.54
$n = 1000$						
iter	obj	cons	time	N	obj	time
500	-3.4405e-1	5.4791e-1	5.13	2000	-3.5511e-1	16.15
1000	-3.5092e-1	5.2894e-1	10.24	5000	-3.4650e-1	89.36
2000	-3.5567e-1	5.0755e-1	20.41	10000	-3.4820e-1	221.4
$n = 2000$						
iter	obj	cons	time	N	obj	time
500	-3.5136e-1	7.0267e-1	14.07	2000	-3.7339e-1	33.26
1000	-3.6720e-1	7.0783e-1	28.09	5000	-3.6945e-1	175.2
2000	-3.7652e-1	6.8650e-1	56.12	10000	-3.6837e-1	613.5
$n = 5000$						
iter	obj	cons	time	N	obj	time
500	-3.2474e-1	9.3906e-1	51.07	2000	-3.9027e-1	68.29
1000	-3.4848e-1	9.0624e-1	101.7	5000	-3.8869e-1	332.4
2000	-3.6643e-1	8.6893e-1	203.3	10000	-3.8781e-1	1159.4

First, as the characteristic function $\mathbb{1}_{[0,\infty)}(\cdot)$ is discontinuous, we replace it with its smooth approximation function $\phi : \mathbb{R} \to \mathbb{R}$ to obtain the following problem

$$\begin{aligned} \min \quad & f(x) := \mathbb{E}[F(x,\xi)] \\ \text{s.t.} \quad & g(x) := \mathbb{E}[\phi(G(x,\xi))] - \alpha \le 0, \\ & x \in \mathcal{C}. \end{aligned} \tag{6.2}$$

See [54, 15] for detailed discussion on this smooth approximation approach. In this experiment, at the k-th iteration we use the following smooth approximation function

$$\phi_k(y) := \frac{1}{1 + \exp(-y/s_k)},$$

where s_k is the smoothing parameter and updated by $s_{k+1} = 0.999 \cdot s_k$. Second, it is known that the chance constraint is usually nonconvex. Fortunately, a "good" initial point can be derived by solving the following CVaR constrained problem

$$\begin{aligned} \min \quad & \mathbb{E}[F(x,\xi)] \\ \text{s.t.} \quad & \mathrm{CVaR}_{1-\alpha}(G(x,\xi)) \le 0, \\ & x \in \mathcal{C}. \end{aligned} \tag{6.3}$$

In [39] it is shown that this CVaR constrained problem is the best conservative convex approximation to (CCP).

Based on the above discussion, we apply a two-stage method to solve problem (CCP): we first use PSG to solve problem (6.3) (as done in the preceding subsection) to get an initial point; then we use PSG again to solve the approximation problem (6.2). Without any risk of confusion, we still denote this method by "PSG". Let us mention that, in a very recent work [22] the authors also propose a SA-type method for estimating the efficient frontier of (CCP). However, they require that the projection onto the level set $\{x \in \mathcal{C} : \mathbb{E}[F(x,\xi)] \le t\}$ is easy to compute.

We compare PSG with the following two well studied methods:

- SCENARIO. The scenario approximation approach [7, 8] is a simple but popular method for solving (CCP), which takes i.i.d. samples $\xi_1, \ldots, \xi_N$ and then solves the deterministic convex program

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & G(x, \xi_i) \leq 0, , \quad i = 1, \ldots, N, \\ & x \in \mathcal{C}. \end{aligned}$$

  We set $N = 2000$ and use MATLAB function *fmincon* to solve this nonlinear program. It is known that, when N is large enough, this scenario approach usually yields a feasible solution to (CCP).

- SCA-DC. In [19], a sequential convex approximation (SCA) method is proposed for solving the DC (difference-of-convex) formulation of (CCP). At each iteration, we let x^{k+1} be the solution of a (convex) nonlinear program in the following form:

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_1(x) - [g_2(x^k) + \nabla g_2(x^k)^T(x - x^k)] \leq \varepsilon\alpha, \\ & x \in \mathcal{C}. \end{aligned}$$

  Here, $\varepsilon > 0$ is a small approximation parameter, y is fixed, g_1 and g_2 are both convex functions given by

$$g_1(x) := \mathbb{E}[[G(x, \xi) + \varepsilon]_+], \quad g_2(x) := \mathbb{E}[[G(x, \xi)]_+].$$

  In practice, it is suggested in [19] to use *fmincon* to solve the associated SAA (sample size 1000) problem. We stop this method when $|f(x^k) - f(x^{k-1})|/|f(x^*)| \leq 1e-3$.

The following norm optimization problem is constructed in [19] and frequently used to evaluate the performance of numerical algorithms for problem (CCP):

$$\begin{aligned} \min \quad & -\textstyle\sum_{j=1}^n x_j \\ \text{s.t.} \quad & \mathbb{P}\left\{\textstyle\sum_{j=1}^n \xi_{ij}^2 x_j^2 \leq u^2, \ i = 1, \ldots, m\right\} \geq 1 - \alpha, \\ & x \in \mathbb{R}_n^+. \end{aligned} \tag{6.4}$$

We test the following two numerical problems.

Problem 6.3 (iid case). In problem (6.4), ξ_{ij}, $i = 1, \ldots, m$ and $j = 1, \ldots, n$ are i.i.d. standard normal random variables.

Problem 6.4 (noniid case). In problem (6.4), ξ_{ij}, $i = 1, \ldots, m$ and $j = 1, \ldots, n$ are dependent normal random variables with mean j/n, variance 1, and $\mathrm{Cov}(\xi_{ij}, \xi_{i'j}) = 0.5$ when $i \neq i'$ and $\mathrm{Cov}(\xi_{ij}, \xi_{i'j'}) = 0$ when $j \neq j'$.

A wonderful property of Problem 6.3 is that its optimal solution is explicit, which is $x_1^* = \ldots = x_n^* = u/\sqrt{\mathcal{F}_{\chi_n^2}^{-1}(1-\beta)}$ where $\beta = 1 - (1-\alpha)^{1/m}$ and $\mathcal{F}_{\chi_n^2}^{-1}(\cdot)$ denotes the inverse of a chi-square distribution with n degrees of freedom.

The numerical results are presented in Table 4 and Table 5. Looking at the columns of cputime, we obtain that PSG is obviously faster than the other two methods. In view of "obj" and "cons", we observe that PSG and SCA-DC always generate better solutions than SCENARIO.

6.4 Portfolio optimization with second-order stochastic dominance constraint

To investigate the efficiency of Algorithm 3, we consider the second-order stochastic dominance (SSD) constrained portfolio optimization problem which is formulated as:

$$\begin{aligned} \min \quad & \mathbb{E}[-\xi^T x] \\ \text{s.t.} \quad & \mathbb{E}[[\eta - \xi^T x]_+] \leq \mathbb{E}[[\eta - Y]_+], \quad \forall \eta \in \mathbb{R}, \\ & x \in \mathcal{C} := \{x \in \mathbb{R}^n : \textstyle\sum_{i=1}^n x_i = 1, \quad x \geq 0\}, \end{aligned}$$

Table 4: Results of Problem 6.3 (**iid case**). Set $u = 100$ and $\alpha = 0.1$.

PSG			SCENARIO			SCA-DC		
n=10, m=10, optim=-2.0818e+1								
obj	cons	time	obj	cons	time	obj	cons	time
-2.0693e+1	-6.0e-3	0.38	-1.6157e+1	-1.0e-1	18.12	-2.0601e+1	-1.4e-2	18.23
n=100, m=10, optim=-8.5907e+1								
obj	cons	time	obj	cons	time	obj	cons	time
-8.5700e+1	-2.5e-3	0.87	-8.2200e+1	-1.0e-1	52.44	-8.6127e+1	-1.9e-2	20.96
n=100, m=100, optim=-8.1878e+2								
obj	cons	time	obj	cons	time	obj	cons	time
-8.5700e+1	-2.5e-3	0.87	-7.9928e+2	-9.8e-2	150.59	-8.6127e+1	-1.9e-2	20.96

Table 5: Results of Problem 6.4 (**noniid case**). Set $u = 100$ and $\alpha = 0.1$.

PSG			SCENARIO			SCA-DC		
n=10, m=10								
obj	cons	time	obj	cons	time	obj	cons	time
-1.8719e+1	-1.2e-1	0.05	-1.5507e+1	-5.0e-3	2.41	-1.8397e+1	-1.0e-1	1.12
n=100, m=100								
obj	cons	time	obj	cons	time	obj	cons	time
-6.7262e+2	-8.0e-3	4.92	-6.9597e+2	-4.0e-3	49.03	-6.9580e+2	-8.0e-3	51.55

where Y stands for the random return of a benchmark portfolio dominated by the target portfolio in the SSD sense. This SSD model was first introduced by Dentcheva and Ruszczyński [12]. Due to their attractive properties, both CVaR and SSD are widely used to control risk in various stochastic programs, especially in financial portfolio, see [41].

If Y has a discrete distribution $\{y_1, y_2, \ldots, y_N\}$, the SSD constrained portfolio problem is reduced to (see [23])

$$\begin{aligned} \min \quad & \mathbb{E}[-\xi^T x] \\ \text{s.t.} \quad & \mathbb{E}[[y_i - \xi^T x]_+] - \mathbb{E}[[y_i - Y]_+] \le 0, \quad i = 1, \ldots, N, \\ & x \in \mathcal{C} := \{x \in \mathbb{R}^n : \textstyle\sum_{i=1}^n x_i = 1, \quad x \ge 0\}, \end{aligned} \tag{6.5}$$

which can be written in the form of problem (5.1).

We use the following four datasets[5]

$$\{\text{“}\texttt{Dax\_26\_3046}\text{”}, \text{“}\texttt{DowJones\_29\_3020}\text{”}, \text{“}\texttt{SP100\_90\_3020}\text{”}, \text{“}\texttt{DowJones\_76\_30000}\text{”}\}$$

in [23]. Take “`DawJones_29_3020`” for example, “`DawJones`” stands for Dow Jones Index, 29 is the number of stocks and 3020 is the number of scenarios, i.e., $n = 29, N = 3020$ in (6.5).

We test the performance of Algorithm 3 for solving problem (6.5) in the empirical formulation. We let “m-PSG-1” denote the multiple penalized mini-batch stochastic gradient method (Algorithm 3) with constant batch size, while “m-PSG-2” stands for a modified version of Algorithm 3 where step (5.2) is replaced with (5.4). In Table 6, the columns with headers “obj”, “max(cons)” and “time” record the final objective value, the maximum of the values of N constraints, and the elapsed time, respectively. The column “optim” shows the optimal value. We run both two algorithms 500 iterations and the initial point is set to zero. All results are averaged over 20 independent runs.

The results in Table 6 show that, despite absence of theoretical convergence guarantee, m-PSG-2 is very promising and much more efficient than m-PSG-1 as expected, in particular when the number of constraints N is large.

[5]The datasets are available at `https://www.ise.ufl.edu/uryasev/research/testproblems/financial_engineering/portfolio-optimization-with-second-orders-stochastic-dominance-constraints`.

Table 6: Results for SSD constrained portfolio problem.

		m-PSG-1			m-PSG-2		
data	optim	obj	max(cons)	time	obj	max(cons)	time
DAX_26_3046	-6.5701e-4	-6.3941e-4	8.3335e-5	7.51	-6.3850e-4	1.0627e-4	0.12
DowJones_29_3020	-3.3469e-4	-3.2835e-4	0	7.55	-3.2777e-4	0	0.10
SP100_90_3020	-8.6528e-4	-8.4214e-4	5.4401e-4	8.56	-8.4246e-4	5.5448e-4	0.13
DowJones_76_30000	-1.8653e-2	-1.7028e-2	1.1700e-6	78.29	-1.7030e-2	1.1591e-6	0.12

7 Conclusion

Inspired by the success of stochastic approximation algorithms in machine learning and many other fields, in this paper we develop a class of SA-type methods to efficiently solve the stochastic programs with expectation constraints. We present a comprehensive convergence analysis including almost sure convergence and expected convergence rates, and some preliminary numerical experiments to verify the effectiveness of these methods.

There are several interesting future research directions. First, it is unclear how to accelerate the proposed methods by combining with some popular techniques, such as Nesterov's accelerated technique, the accelerated technique in [60] and the variance reduction technique [21]. Second, as witnessed in the experiments for SSD constrained problem, a variant of PSG with (5.4) is particularly effective for optimization with multiple expectation constraints. However, its convergence is not clear yet. Third, it is of both theoretical and practical interest to study the proposed methods for nonconvex problems.

Acknowledgments
This work was partially supported by the National Natural Science Foundation of China (No. 11871135) and the Fundamental Research Funds for the Central Universities (No. DUT19K46).

References

[1] K. Basu and P. Nandy. Optimal convergence for stochastic optimization with multiple expectation constraints. https://arxiv.org/abs/1906.03401, 6 2019.

[2] D. P. Bertsekas. *Convex optimization algorithms*. Athena Scientific, Belmont, MA, 2015.

[3] J. F. Bonnans and A. Ioffe. Second-order sufficiency and quadratic growth for nonisolated minima. *Math. Oper. Res.*, 20(4):801–817, 1995.

[4] J. F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer Series in Operations Research. Springer-Verlag, New York, 2000.

[5] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Rev.*, 60(2):223–311, 2018.

[6] J. V. Burke and S. Deng. Weak sharp minima revisited. II. Application to linear regularity and error bounds. *Math. Program.*, 104(2-3, Ser. B):235–261, 2005.

[7] G. Calafiore and M. C. Campi. Uncertain convex programs: randomized solutions and confidence levels. *Math. Program.*, 102(1, Ser. A):25–46, 2005.

[8] M. C. Campi and S. Garatti. A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality. *J. Optim. Theory Appl.*, 148(2):257–280, 2011.

[9] A. Charnes and W. W. Cooper. Chance-constrained programming. *Management Sci.*, 6:73–79, 1959/1960.

[10] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM J. Optim.*, 3(3):538–543, 1993.

[11] Y. Cui, C. Ding, and X. Zhao. Quadratic growth conditions for convex matrix optimization problems associated with spectral functions. *SIAM J. Optim.*, 27(4):2332–2355, 2017.

[12] D. Dentcheva and A. Ruszczyński. Optimization with stochastic dominance constraints. *SIAM J. Optim.*, 14(2):548–566, 2003.

[13] D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Math. Oper. Res.*, 43(3):919–948, 2018.

[14] Y. M. Ermoliev and V. I. Norkin. Sample average approximation method for compound stochastic optimization problems. *SIAM J. Optim.*, 23(4):2231–2263, 2013.

[15] A. Geletu, A. Hoffmann, M. Klöppel, and P. Li. An inner-outer approximation approach to chance constrained optimization. *SIAM J. Optim.*, 27(3):1834–1857, 2017.

[16] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Math. Program.*, 155(1-2, Ser. A):267–305, 2016.

[17] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror. Result analysis of the NIPS 2003 feature selection challenge. In *Adv. in Neural Inf. Process. Syst. 17*, pages 545–552. MIT Press, 2004.

[18] A. J. Hoffman. On approximate solutions of systems of linear inequalities. *J. Research Nat. Bur. Standards*, 49:263–265, 1952.

[19] L. J. Hong, Y. Yang, and L. Zhang. Sequential convex approximations to joint chance constrained programs: a Monte Carlo approach. *Oper. Res.*, 59(3):617–630, 2011.

[20] A. D. Ioffe. *Variational analysis of regular mappings*. Springer Monographs in Mathematics. Springer, Cham, 2017. Theory and applications.

[21] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Adv. in Neural Inf. Process. Syst.*, pages 315–323, 2013.

[22] R. Kannan and J. Luedtke. A stochastic approximation method for chance-constrained nonlinear programs. https://arxiv.org/abs/1812.07066, 12 2018.

[23] N. F. Keçeci, V. Kuzmenko, and S. Uryasev. Portfolios dominating indices: Optimization with second-order stochastic dominance constraints vs. minimum and mean variance portfolios. *Journal of Risk and Financial Management*, 9(4):1–14, 2016.

[24] H. J. Kushner and G. G. Yin. *Stochastic approximation algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1997.

[25] G. Lan, A. Nemirovski, and A. Shapiro. Validation analysis of mirror descent stochastic approximation method. *Math. Program.*, 134(2, Ser. A):425–458, 2012.

[26] G. Lan and Z. Zhou. Algorithms for stochastic optimization with expectation constraints. https://arxiv.org/abs/1604.03887, 04 2016.

[27] Y. LeCun, C. Cortes, and C. J. C. Burges. The mnist database of handwritten digits. `http://yann.lecun.com/exdb/mnist/`. 2010.

[28] A. S. Lewis and J.-S. Pang. Error bounds for convex inequality systems. In *Generalized convexity, generalized monotonicity: recent results (Luminy, 1996)*, volume 27 of *Nonconvex Optim. Appl.*, pages 75–110. Kluwer Acad. Publ., Dordrecht, 1998.

[29] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397, 2004.

[30] Q. Lin, R. Ma, and T. Yang. Level-set methods for finite-sum constrained convex optimization. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3112–3121, 2018.

[31] Q. Lin, S. Nadarajah, N. Soheili, and T. Yang. A data efficient and feasible level set method for stochastic convex optimization with expectation constraints. https://arxiv.org/abs/1908.03077, 08 2019.

[32] Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Ann. Oper. Res.*, 46/47(1-4):157–178, 1993. Degeneracy in optimization problems.

[33] O. L. Mangasarian. Error bounds for nondifferentiable convex inequalities under a strong Slater constraint qualification. *Math. Programming*, 83(2, Ser. A):187–194, 1998.

[34] B. L. Miller and H. M. Wagner. Chance constrained programming with joint constraints. *Operations Res.*, 13(6):930–945, 1965.

[35] I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Math. Program.*, 175(1-2, Ser. A):69–107, 2019.

[36] A. Nedić. Random algorithms for convex minimization problems. *Math. Program.*, 129(2, Ser. B):225–253, 2011.

[37] A. Nedić and I. Necoara. Random minibatch projection algorithms for convex problems with functional constraints. https://arxiv.org/abs/1903.02117, 03 2019.

[38] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2009.

[39] A. Nemirovski and A. Shapiro. Convex approximations of chance constrained programs. *SIAM J. Optim.*, 17(4):969–996, 2006.

[40] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

[41] N. Noyan. Risk-averse stochastic modeling and optimization. In *Recent Advances in Optimization and Modeling of Contemporary Problems*, chapter 10, pages 221–254. 2018.

[42] R. I. Oliveira and P. Thompson. Sample average approximation with heavier tails i: non-asymptotic bounds with weak assumptions and stochastic constraints. https://arxiv.org/abs/1705.00822, 5 2017.

[43] A. Patrascu and I. Necoara. Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *J. Mach. Learn. Res.*, 18:Paper No. 198, 42, 2017.

[44] B. T. Polyak. A general method of solving extremum problems. *Doklady Akademii Nauk SSSR*, 174(1):593–597, 1967.

[45] B. T. Polyak. *Introduction to optimization*. Translations Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York, 1987. Translated from the Russian, With a foreword by Dimitri P. Bertsekas.

[46] P. Rigollet and X. Tong. Neyman-Pearson classification, convexity and stochastic constraints. *J. Mach. Learn. Res.*, 12:2831–2855, 2011.

[47] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951.

[48] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics (Proc. Sympos., Ohio State Univ., Columbus, Ohio, 1971)*, pages 233–257. Academic Press, New York, 1971.

[49] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–41, 2000.

[50] R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, 26:1443–1471, 2002.

[51] W. Römisch. Stability of stochastic programming problems. In *Stochastic programming*, volume 10 of *Handbooks Oper. Res. Management Sci.*, pages 483–554. Elsevier Sci. B. V., Amsterdam, 2003.

[52] A. Ruszczyński and A. Shapiro. Stochastic programming models. In *Stochastic programming*, volume 10 of *Handbooks Oper. Res. Management Sci.*, pages 1–64. Elsevier Sci. B. V., Amsterdam, 2003.

[53] C. Scott and R. Nowak. A Neyman-Pearson approach to statistical learning. *IEEE Trans. Inform. Theory*, 51(11):3806–3819, 2005.

[54] F. Shan, L. Zhang, and X. Xiao. A smoothing function approach to joint chance-constrained programs. *J. Optim. Theory Appl.*, 163(1):181–199, 2014.

[55] A. Shapiro. *Sample Average Approximation*, pages 1350–1355. Springer US, Boston, MA, 2013.

[56] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming*, volume 9 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, second edition, 2014. Modeling and theory.

[57] X. Tong, Y. Feng, and A. Zhao. A survey on Neyman-Pearson classification and suggestions for future research. *Wiley Interdiscip. Rev. Comput. Stat.*, 8(2):64–81, 2016.

[58] M. Wang and D. P. Bertsekas. Stochastic first-order methods with random constraint projection. *SIAM J. Optim.*, 26(1):681–717, 2016.

[59] M. Wang, E. X. Fang, and H. Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Math. Program.*, 161(1-2, Ser. A):419–449, 2017.

[60] M. Wang, J. Liu, and E. X. Fang. Accelerating stochastic composition optimization. *J. Mach. Learn. Res.*, 18:Paper No. 105, 23, 2017.

[61] W. Wang and S. Ahmed. Sample average approximation of expected value constrained stochastic programs. *Oper. Res. Lett.*, 36(5):515–519, 2008.

[62] X. Wang, S. Ma, and Y.-x. Yuan. Penalty methods with stochastic approximation for stochastic nonlinear programming. *Math. Comp.*, 86(306):1793–1820, 2017.

[63] X. Wei, H. Yu, and M. J. Neely. Online primal-dual mirror descent under stochastic constraints. https://arxiv.org/abs/1908.00305, 08 2019.

[64] C. workbench team. A marketing dataset. `http://www.causality.inf.ethz.ch/data/CINA.html`, 09 2008.

[65] Y. Xu. Primal-dual stochastic gradient method for convex programs with many functional constraints. https://arxiv.org/abs/1802.02724, 02 2018.

[66] H. Yu, M. J. Neely, and X. Wei. Online convex optimization with stochastic constraints. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1427–1437, 04–09 Dec 2017.

Appendices

Appendix A Some useful Lemmas

In Theorem 2.4 and Theorem 2.5, we have seen some remarkable martingale convergence arguments, which are useful to prove that the stochastic sequence converges almost surely (but not necessarily to zero). We introduce another martingale convergence result in Lemma A.1 which gives a condition such that the interested stochastic process converges almost surely to zero. When the parameter sequence $\{b_k\}$ is deterministic, it reduces to the result [45, Lemma 10, Page 49].

Lemma A.1. *Let $\{\zeta_k, b_k\}$ be nonnegative adapted processes with respect to the filtration $\{\mathcal{F}_k\}$ such that*

$$\sum_k b_k < \infty, \quad w.p.1,$$

and for all $k \geq 0$,

$$\mathbb{E}[\zeta_{k+1}|\mathcal{F}_k] \leq (1 - a_k)\zeta_k + b_k, \quad w.p.1, \tag{A.1}$$

where $\{a_k\}$ is a positive deterministic scalar sequence satisfying

$$a_k \leq 1, \quad \sum_k a_k = \infty, \quad \lim_{k\to\infty} \frac{\mathbb{E}[b_k]}{a_k} = 0.$$

Then, $\{\zeta_k\}$ converges almost surely to zero.

Proof. The proof consists of three parts. Firstly, we show that $\{\zeta_k\}$ converges almost surely. Next, $\{\mathbb{E}[\xi_k]\}$ is proved to converge to zero. Finally, we get that $\{\zeta_k\}$ converges almost surely to zero.

From (A.1), it leads to

$$\mathbb{E}[\zeta_{k+1}|\mathcal{F}_k] \leq \zeta_k + b_k, \quad \text{w.p.1.}$$

As $\sum_k b_k < \infty$ with probability 1, by using Theorem 2.4, we get that $\{\zeta_k\}$ converges almost surely to a random variable ζ^*.

We use [2, Proposition A.4.3] to prove the second part. Taking expectation on both sides of (A.1), we have

$$\mathbb{E}[\zeta_{k+1}] \leq (1 - a_k)\mathbb{E}[\zeta_k] + \mathbb{E}[b_k],$$

which, together with $\sum_k a_k = \infty$ and $\mathbb{E}[b_k]/a_k \to 0$, implies that $\{\mathbb{E}[\zeta_k]\}$ converges to zero.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the corresponding probability space. We have shown that for almost every $\omega \in \Omega$, it holds that $\zeta_k(\omega)$ converges to $\zeta^*(\omega)$ and $\zeta^*(\omega) \geq 0$. Using Fatou's lemma, one has

$$\liminf_{k\to\infty} \int_\Omega \zeta_k(\omega)\mathrm{d}\mathbb{P}(\omega) \geq \int_\Omega \zeta^*(\omega)\mathrm{d}\mathbb{P}(\omega) \geq 0.$$

Since it has bee proved that $\lim_{k\to\infty} \mathbb{E}[\zeta_k] = 0$, we get $\int_\Omega \zeta^*(\omega)\mathrm{d}\mathbb{P}(\omega) = 0$, which indicates that $\xi^*(\omega) = 0$ almost everywhere. The proof is completed. □

In Lemma A.2 we introduce some auxiliary results about stepsizes α_k and β_k.

Lemma A.2. *Let $\delta_k := \frac{\beta_k}{\beta_k - \frac{\kappa}{2}\alpha_k}$, where $\kappa > 0$ is a constant, α_k and β_k are given in (3.19) or (4.7). Then, when k is large enough, we have:*

(i) $(\delta_k+\beta_k)(1-\beta_k)\le(1-\frac{\kappa}{2}\alpha_k)\delta_k$;

(ii) $\delta_k\le 2$;

(iii) $\delta_{k+1}\le\delta_k$;

(iv) $\delta_k>1$.

Proof. We only consider the case of (3.19). The case of (4.7) can be obtained identically. Note that $\beta_k\ge\frac{\kappa}{2}\alpha_k$ when $k\ge K$ for sufficiently large K. Then, item (i) is obtained by

$$\frac{(\delta_k+\beta_k)(1-\beta_k)}{(1-\frac{\kappa}{2}\alpha_k)\delta_k}=\frac{(1+\beta_k-\frac{\kappa}{2}\alpha_k)(1-\beta_k)}{1-\frac{\kappa}{2}\alpha_k}=(1-\beta_k)+\beta_k\cdot\frac{1-\beta_k}{1-\frac{\kappa}{2}\alpha_k}\le 1.$$

Item (ii) is obvious by using $\beta_k\ge\kappa\alpha_k$ for sufficiently large k. For item (iii), we can see that it is equivalent to prove that the function $\psi(k):=\frac{(k-1)^b}{k^a}$ is nonincreasing with $k\ge 2$, $b:=\frac{1}{2}+\varepsilon$, $a:=\frac{3}{4}+2\varepsilon$ and $0<\varepsilon<1/8$. The derivative of $\psi(k)$ satisfies

$$\psi'(k)=\frac{b(k-1)^{b-1}\cdot k^a-a(k-1)^b\cdot k^{a-1}}{k^{2a}}\le 0,$$

when

$$\frac{k}{k-1}\le\frac{a}{b}.$$

As $a/b>\frac{3}{2}$, the above condition holds true when $k\ge 3$. This means that $\delta_{k+1}\le\delta_k$ when $k\ge\max\{3,K\}$. Finally, noting that

$$\delta_k=\frac{1}{1-\frac{\kappa}{2}\frac{\alpha_k}{\beta_k}}=\frac{1}{1-\frac{\kappa\alpha}{2\beta}\cdot\frac{(k-1)^{1/2+\varepsilon}}{k^{3/4+2\varepsilon}}},$$

we have that $\delta_k>1$ when k is large enough. □

Appendix B Proofs in Section 4

B.1 Proof of lemma 4.3

Proof. By substituting $x^+=y^k$ and $u=x^*$ in (2.3), we have

$$V(y^k,x^*)=V(x^k,x^*)-V(x^k,y^k)-\alpha_k\langle\nabla\mathcal{F}^k,y^k-x^*\rangle. \tag{B.1}$$

We next estimate the bound of the last term in the above inequality. Write

$$-\alpha_k\langle\nabla\mathcal{F}^k,y^k-x^*\rangle=-\alpha_k\langle\varepsilon_{f'}^k,y^k-x^*\rangle-\alpha_k\langle\nabla f(x^k),y^k-x^*\rangle. \tag{B.2}$$

For the first term on the right-hand side in (B.2), by using (4.2) we see

$$\begin{aligned}-\alpha_k\langle\varepsilon_{f'}^k,y^k-x^*\rangle&=-\alpha_k\langle\varepsilon_{f'}^k,y^k-\hat{y}^k\rangle-\alpha_k\langle\varepsilon_{f'}^k,\hat{y}^k-x^*\rangle\\&\le\tfrac{\alpha_k^2}{\lambda}\|\varepsilon_{f'}^k\|_*^2-\alpha_k\langle\varepsilon_{f'}^k,\hat{y}^k-x^*\rangle.\end{aligned} \tag{B.3}$$

For the second term on the right-hand side in (B.2), we have

$$\begin{aligned}&-\alpha_k\langle\nabla f(x^k),y^k-x^*\rangle\\&=-\alpha_k\langle\nabla f(x^k)-\nabla f(u),u-x^*\rangle-\alpha_k\langle\nabla f(u),u-x^*\rangle-\alpha_k\langle\nabla f(x^k),y^k-u\rangle\\&\le\alpha_kL_f\|u-x^k\|\cdot(\|u-x^k\|+\|x^k-x^*\|)-\alpha_k(f(u)-f^*)+\tfrac{C_f^2\alpha_k^2}{4\eta\gamma_k}+\eta\gamma_k\|u-y^k\|^2\\&\le\alpha_kL_f\|u-x^k\|^2+\eta\gamma_k\|u-x^k\|^2+\tfrac{\alpha_k^2}{4\eta\gamma_k}L_f^2\|x^k-x^*\|^2-\alpha_k(f(u)-f^*)\\&\quad+\tfrac{C_f^2\alpha_k^2}{4\eta\gamma_k}+\eta\gamma_k\|u-y^k\|^2.\end{aligned} \tag{B.4}$$

Furthermore, from $2\eta\gamma_k\le\lambda/4$ one has

$$\begin{aligned}-V(x^k,y^k)+\eta\gamma_k\|u-y^k\|^2&\le-\tfrac{\lambda}{2}\|y^k-x^k\|^2+2\eta\gamma_k\|x^k-y^k\|^2+2\eta\gamma_k\|u-x^k\|^2\\&\le-\tfrac{\lambda}{4}\|y^k-x^k\|^2+2\eta\gamma_k\|u-x^k\|^2.\end{aligned} \tag{B.5}$$

Combining (B.1)-(B.5) together, we derive

$$\begin{aligned}V(y^k,x^*)&\le[1+\tfrac{L_f^2}{2\lambda\eta}\cdot\tfrac{\alpha_k^2}{\gamma_k}]V(x^k,x^*)-\alpha_k(f(u)-f^*)+C_f^2\tfrac{\alpha_k^2}{4\eta\gamma_k}\\&\quad+(\alpha_kL_f+3\eta\gamma_k)\|u-x^k\|^2+\tfrac{\alpha_k^2}{\lambda}\|\varepsilon_{f'}^k\|^2-\alpha_k\langle\varepsilon_{f'}^k,\hat{y}^k-x^*\rangle-\tfrac{\lambda}{4}\|y^k-x^k\|^2.\end{aligned}$$

By taking conditional expectation on both sides, using $\mathbb{E}[\langle\varepsilon_{f'}^k,\hat{y}^k-x^*\rangle|\mathcal{F}_k]=0$ and Lemma 4.1, we obtain the claim. □

B.2 Proof of Lemma 4.4

Proof. We first write

$$\begin{aligned}&\langle [t_{k+1}]_+ G'(x^k,\eta^k), x^k - x^*\rangle \\ &= \langle [g(x^k)]_+ G'(x^k,\eta^k), x^k - x^*\rangle - \langle G'(x^k,\eta^k), x^k - x^*\rangle([g(x^k)]_+ - [t_{k+1}]_+).\end{aligned}$$

By taking conditional expectation on both sides, it gives

$$\begin{aligned}&\mathbb{E}[-\gamma_k\langle [t_{k+1}]_+ G'(x^k,\eta^k), x^k - x^*\rangle|\hat{\mathcal{F}}_k] \\ &= -\gamma_k\langle [g(x^k)]_+ g'(x^k), x^k - x^*\rangle + \gamma_k\langle g'(x^k), x^k - x^*\rangle([g(x^k)]_+ - [t_{k+1}]_+) \\ &\le -\tfrac{\gamma_k}{2}\phi(x^k) + C_g^2\tfrac{\gamma_k^2}{4\beta_k}\|x^k - x^*\|^2 + \beta_k(g(x^k) - t_{k+1})^2, \quad \text{w.p.1},\end{aligned}$$

in which we use $g'(x^k) := \mathbb{E}[G'(x^k,\eta^k)|\hat{\mathcal{F}}_k]$ and $2[g(x^k)]_+ g'(x^k) \in \partial\phi(x^k)$. Thus, we get

$$\begin{aligned}&\mathbb{E}[-\gamma_k\langle [t_{k+1}]_+ G'(x^k,\eta^k), x^k - x^*\rangle|\mathcal{F}_k] \\ &\le -\tfrac{\gamma_k}{2}\phi(x^k) + C_g^2\tfrac{\gamma_k^2}{2\lambda\beta_k}V(x^k,x^*) + \beta_k\mathbb{E}[(g(x^k) - t_{k+1})^2|\mathcal{F}_k], \quad \text{w.p.1}.\end{aligned} \tag{B.6}$$

Finally, by combining the above inequality with the following relation

$$\begin{aligned}&\mathbb{E}[(g(x^k) - t_{k+1})^2|\mathcal{F}_k] \le (1-\beta_k)(g(x^k) - t_k)^2 + \beta_k\mathbb{E}[(\varepsilon_g^k)^2|\mathcal{F}_k] \\ &\le (1-\beta_k)(1+1/\beta_k)(g(x^k) - g(x^{k-1}))^2 + (1-\beta_k)(1+\beta_k)(g(x^{k-1}) - t_k)^2 + \beta_k\mathbb{E}[(\varepsilon_g^k)^2|\mathcal{F}_k] \\ &\le \tfrac{1}{\beta_k}C_g^2\|x^k - x^{k-1}\|^2 + (g(x^{k-1}) - t_k)^2 + \tfrac{\beta_k}{N_k}[6(C_g^2 + \hat{C}_g^2)V(x^k,x^*)/\lambda + 3\mathcal{V}_g(x^*)], \quad \text{w.p.1},\end{aligned} \tag{B.7}$$

we derive the claimed result. □

B.3 Proof of Lemma 4.5

Proof. The proof is in the same way as [59, Lemma 2]. Let $e_k = (1-\beta_k)(g(x^k) - g(x^{k-1}))$, we write

$$t_{k+1} - g(x^k) + e_k = (1-\beta_k)(t_k - g(x^{k-1})) + \beta_k\varepsilon_g^k.$$

By using $\mathbb{E}[\langle \varepsilon_g^k, t_k - g(x^{k-1})\rangle|\mathcal{F}_k] = 0$ and Lemma 4.1, we have

$$\mathbb{E}[(t_{k+1} - g(x^k) + e_k)^2|\mathcal{F}_k] \le (1-\beta_k)^2(t_k - g(x^{k-1}))^2 + \frac{\beta_k^2}{N_k}[6(C_g^2 + \hat{C}_g^2)V(x^k,x^*)/\lambda + 3\mathcal{V}_g(x^*)]$$

with probability 1. By taking conditional expectation on the following relation

$$(t_{k+1} - g(x^k))^2 \le (1+\beta_k)(t_{k+1} - g(x^k) + e_k)^2 + (1+1/\beta_k)e_k^2,$$

we obtain

$$\begin{aligned}&\mathbb{E}[(t_{k+1} - g(x^k))^2|\mathcal{F}_k] \\ &\le (1+\beta_k)(1-\beta_k)^2(t_k - g(x^{k-1}))^2 + \tfrac{(1+\beta_k)\beta_k^2}{N_k}[6(C_g^2 + \hat{C}_g^2)V(x^k,x^*)/\lambda + 3\mathcal{V}_g(x^*)] \\ &\quad + \tfrac{(1+\beta_k)(1-\beta_k)^2}{\beta_k}(g(x^k) - g(x^{k-1})^2) \\ &\le (1-\beta_k)(t_k - g(x^{k-1}))^2 + \tfrac{12\beta_k^2}{\lambda N_k}(C_g^2 + \hat{C}_g^2)V(x^k,x^*) + \tfrac{6\mathcal{V}_g(x^*)\beta_k^2}{N_k} + \tfrac{C_g^2}{\beta_k}\|x^k - x^{k-1}\|^2\end{aligned}$$

with probability 1. The proof is completed. □

B.4 Proof of Lemma 4.6

Proof. Substituting $x^+ = x^{k+1}$ and $u = x^*$ in (2.3), we have

$$V(x^{k+1},x^*) \le V(y^k,x^*) - V(y^k,x^{k+1}) - \langle \gamma_k[t_{k+1}]_+ G'(x^k,\eta^k), x^{k+1} - x^*\rangle,$$

and thus together with the following relation

$$\begin{aligned}&-V(y^k,x^{k+1}) - \langle \gamma_k[t_{k+1}]_+ G'(x^k,\eta^k), x^{k+1} - x^k\rangle \\ &\le -\tfrac{\lambda}{2}\|x^{k+1} - y^k\|^2 + \tfrac{2\gamma_k^2}{\lambda}[t_{k+1}]_+^2\|G'(x^k,\eta^k)\|_*^2 + \tfrac{\lambda}{8}\|x^{k+1} - x^k\|^2 \\ &\le \tfrac{2\gamma_k^2}{\lambda}[t_{k+1}]_+^2\|G'(x^k,\eta^k)\|_*^2 + \tfrac{\lambda}{4}\|y^k - x^k\|^2,\end{aligned}$$

it yields

$$V(x^{k+1},x^*) \le V(y^k,x^*) + \frac{2\gamma_k^2}{\lambda}[t_{k+1}]_+^2\|G'(x^k,\eta^k)\|_*^2 + \frac{\lambda}{4}\|y^k-x^k\|^2 - \langle \gamma_k[t_{k+1}]_+ G'(x^k,\eta^k), x^k-x^*\rangle.$$

Taking conditional expectation on both sides of the above inequality we get

$$\begin{aligned}\mathbb{E}[V(x^{k+1},x^*)|\mathcal{F}_k] \quad &\le \mathbb{E}[V(y^k,x^*)|\mathcal{F}_k] + 2C_t^2\hat{C}_g^2\tfrac{\gamma_k^2}{\lambda} + \tfrac{\lambda}{4}\mathbb{E}[\|y^k-x^k\|^2|\mathcal{F}_k] \\ &\quad -\gamma_k\mathbb{E}[\langle [t_{k+1}]_+ G'(x^k,\eta^k), x^k-x^*\rangle|\mathcal{F}_k], \quad \text{w.p.1.}\end{aligned} \tag{B.8}$$

In view of the above inequality, combining with Lemma 4.3 (with $u=\bar{x}^k$), Lemma 4.4 and the error bound condition (3.8), one has

$$\begin{aligned}&\mathbb{E}[V(x^{k+1},x^*)|\mathcal{F}_k] \\ &\le \Big(1+\tfrac{L_f^2}{2\lambda\eta}\cdot\tfrac{\alpha_k^2}{\gamma_k} + 6(L_f^2+\hat{L}_f^2)\tfrac{\alpha_k^2}{\lambda^2 N_k} + \tfrac{C_g^2}{2\lambda}\cdot\tfrac{\gamma_k^2}{\beta_k} + 6(C_g^2+\hat{C}_g^2)\tfrac{\beta_k^2}{\lambda N_k}\Big)V(x^k,x^*) - \alpha_k(f(\bar{x}^k)-f^*) \\ &\quad -(\tfrac{\gamma_k}{2C_{eb}} - 3\eta\gamma_k - L_f\alpha_k)\|x^k-\bar{x}^k\|^2 + \tfrac{C_f^2}{4\eta}\cdot\tfrac{\alpha_k^2}{\gamma_k} + \tfrac{2\hat{C}_g^2C_t^2}{\lambda}\gamma_k^2 \\ &\quad +3\mathcal{V}_{f'}(x^*)\tfrac{\alpha_k^2}{\lambda N_k} + 3\mathcal{V}_g(x^*)\tfrac{\beta_k^2}{N_k} + C_g^2\|x^k-x^{k-1}\|^2 + \beta_k(g(x^{k-1})-t_k)^2\end{aligned}$$

with probability 1. Furthermore, from (4.3) and the definitions of a_k,μ_k, it turns out

$$\begin{aligned}&\mathbb{E}[V(x^{k+1},x^*)|\mathcal{F}_k] \\ &\le (1+a_k)V(x^k,x^*) - \alpha_k(f(\bar{x}^k)-f^*+\|x^k-\bar{x}^k\|^2) + \beta_k(g(x^{k-1})-t_k)^2 + \mu_k\end{aligned}$$

with probability 1. Thus, we derive the relation (4.4). The relation (4.5) is obtained from Lemma 4.5 by using the definitions of $\zeta_k, d_k, \bar{u}_k, b_k, \theta_k, \nu_k$. □

B.5 Proof of Theorem 4.8

Proof. We first note that conditions (4.3) hold true when k is large enough. In view of (B.8), by combining with (B.6) and Lemma 4.3, we have

$$\begin{aligned}&\mathbb{E}[V(x^{k+1},x^*)|\mathcal{F}_k] \\ &\le \Big[1+\tfrac{L_f^2}{2\lambda\eta}\cdot\tfrac{\alpha_k^2}{\gamma_k} + 6(L_f^2+\hat{L}_f^2)\tfrac{\alpha_k^2}{\lambda^2 N_k} + \tfrac{C_g^2}{2\lambda}\cdot\tfrac{\gamma_k^2}{\beta_k}\Big]V(x^k,x^*) - \alpha_k(f(\bar{x}^k)-f^*+\|\bar{x}^k-x^k\|^2) \\ &\quad +3\mathcal{V}_{f'}(x^*)\tfrac{\alpha_k^2}{\lambda N_k} + \tfrac{C_f^2}{4\eta}\cdot\tfrac{\alpha_k^2}{\gamma_k} + 2C_t^2\hat{C}_g^2\gamma_k^2/\lambda + \beta_k\mathbb{E}[(g(x^k)-t_{k+1})^2|\mathcal{F}_k], \quad \text{w.p.1.}\end{aligned} \tag{B.9}$$

Multiplying the inequality in Lemma 4.5 with $(1+\beta_k)$ and adding it to the preceding inequality, it turns out that

$$\mathbb{E}[\zeta_{k+1}|\mathcal{F}_k] \le [1+a_k]\,\zeta_k - \alpha_k(f(\bar{x}^k)-f^*+\|\bar{x}^k-x^k\|^2) + b_k, \quad \text{w.p.1,} \tag{B.10}$$

where

$$\zeta_k := V(x^k,x^*) + (g(x^{k-1})-t_k)^2, \quad a_k := \frac{L_f^2}{2\lambda\eta}\cdot\frac{\alpha_k^2}{\gamma_k} + 6(L_f^2+\hat{L}_f^2)\frac{\alpha_k^2}{\lambda^2N_k} + \frac{C_g^2}{2\lambda}\cdot\frac{\gamma_k^2}{\beta_k} + (1+\beta_k)\frac{12\beta_k^2}{\lambda N_k}(C_g^2+\hat{C}_g^2)$$

and

$$b_k := 3\mathcal{V}_{f'}(x^*)\frac{\alpha_k^2}{\lambda N_k} + \frac{C_f^2}{4\eta}\cdot\frac{\alpha_k^2}{\gamma_k} + 2C_t^2\hat{C}_g^2\gamma_k^2/\lambda + (1+\beta_k)\frac{6\mathcal{V}_g(x^*)\beta_k^2}{N_k} + (1+\beta_k)\frac{C_g^2}{\beta_k}\|x^k-x^{k-1}\|^2.$$

It is not difficult to verify that Assumption 6 is satisfied in both settings of item (i) and item (ii), and hence we have

$$\sum_k a_k < \infty, \quad \sum_k b_k < \infty.$$

Then, from (B.10) and Theorem 2.4 it follows that ζ_k converges almost surely, and hence, for $\lfloor \frac{K}{2} \rfloor \le k \le K-1$ we have

$$\alpha_k\mathbb{E}[(f(\bar{x}^k)-f^*+\|x^k-\bar{x}^k\|^2)] \le \mathbb{E}[\zeta_k] - \mathbb{E}[\zeta_{k+1}] + a_kc_K^* + \mathbb{E}[b_k].$$

where $c_K^* := \sup_{\lfloor \frac{K}{2} \rfloor \le k \le K-1}\mathbb{E}[\zeta_k]$. Therefore, according to the proof in Theorem 3.7, the convergence rates can be derived. □

B.6 Proof of Theorem 4.9

Proof. We first establish a new relationship between y^k and x^k. Let us rewrite (B.4) by

$$\begin{aligned} -\alpha_k\langle\nabla f(x^k), y^k - x^*\rangle &= -\alpha_k\langle\nabla f(x^k), x^k - x^*\rangle - \alpha_k\langle\nabla f(x^k), y^k - x^k\rangle \\ &\le -\alpha_k(f(x^k) - f^*) + \alpha_k C_f\|x^k - y^k\|_2 \\ &\le -\alpha_k(f(x^k) - f^*) + C_f^2\alpha_k^2 + \tfrac{1}{4}\|x^k - y^k\|_2^2. \end{aligned}$$

By combining this inequality with (B.1), (B.2) and (B.3) (note that $\lambda = 1$), one has

$$\frac{1}{2}\|y^k - x^*\|_2^2 \le \frac{1}{2}\|x^k - x^*\|_2^2 - \alpha_k(f(x^k) - f^*) + \alpha_k^2\|\varepsilon_{f'}^k\|_2^2 - \alpha_k\langle\varepsilon_{f'}^k, \hat{y}^k - x^*\rangle + C_f^2\alpha_k^2 - \frac{1}{4}\|x^k - y^k\|_2^2.$$

Then, taking conditional expectation on both sides and applying Lemma 4.1, it follows

$$\begin{aligned} \mathbb{E}[\tfrac{1}{2}\|y^k - x^*\|_2^2|\mathcal{F}_k] \le &\left[1 + 6(L_f^2 + \hat{L}_f^2)\tfrac{\alpha_k^2}{N_k}\right]\cdot\tfrac{1}{2}\|x^k - x^*\|_2^2 - \alpha_k(f(x^k - f^*)) \\ &+3\mathcal{V}_{f'}(x^*)\tfrac{\alpha_k^2}{N_k} + C_f^2\alpha_k^2 - \tfrac{1}{4}\mathbb{E}[\|x^k - y^k\|^2|\mathcal{F}_k], \end{aligned} \tag{B.11}$$

with probability 1.

Let us now consider the relation between x^{k+1} and x^k. In view of (B.8), by using (B.6) and (B.11), we obtain

$$\begin{aligned} \mathbb{E}[\tfrac{1}{2}\|x^{k+1} - x^*\|_2^2|\mathcal{F}_k] \le &\left[1 + 6(L_f^2 + \hat{L}_f^2)\tfrac{\alpha_k^2}{N_k} + C_g^2\tfrac{\gamma_k^2}{2\beta_k}\right]\cdot\tfrac{1}{2}\|x^k - x^*\|_2^2 - \alpha_k(f(x^k) - f^*) - \tfrac{\gamma_k}{2}\phi(x^k) \\ &+3\mathcal{V}_{f'}(x^*)\tfrac{\alpha_k^2}{N_k} + \beta_k\mathbb{E}[(g(x^k) - t_{k+1})^2|\mathcal{F}_k] + 2C_t^2\hat{C}_g^2\gamma_k^2 + C_f^2\alpha_k^2. \end{aligned}$$

Then, from $\frac{\gamma_k}{2} > \rho\alpha_k$ (by carefully choosing α and γ) and Assumption 4 it follows that

$$\begin{aligned} \mathbb{E}[\tfrac{1}{2}\|x^{k+1} - x^*\|_2^2|\mathcal{F}_k] \le &\left[1 + 6(L_f^2 + \hat{L}_f^2)\tfrac{\alpha_k^2}{N_k} - \kappa\alpha_k + C_g^2\tfrac{\gamma_k^2}{2\beta_k}\right]\cdot\tfrac{1}{2}\|x^k - x^*\|_2^2 + 3\mathcal{V}_{f'}(x^*)\tfrac{\alpha_k^2}{N_k} \\ &+\beta_k\mathbb{E}[(g(x^k) - t_{k+1})^2|\mathcal{F}_k] + 2C_t^2\hat{C}_g^2\gamma_k^2 + C_f^2\alpha_k^2 \end{aligned} \tag{B.12}$$

with probability 1. In both settings of item (i) and item (ii), for sufficiently large k, it holds

$$6(L_f^2 + \hat{L}_f^2)\frac{\alpha_k^2}{N_k} + C_g^2\frac{\gamma_k^2}{2\beta_k} \le \frac{\kappa}{2}\alpha_k - 36(C_g^2 + \hat{C}_g^2)\frac{\beta_k^2}{N_k},$$

which, together with (B.12), yields

$$\begin{aligned} \mathbb{E}[\tfrac{1}{2}\|x^{k+1} - x^*\|_2^2|\mathcal{F}_k] \le &\left[1 - \tfrac{\kappa}{2}\alpha_k - 36(C_g^2 + \hat{C}_g^2)\tfrac{\beta_k^2}{N_k}\right]\cdot\tfrac{1}{2}\|x^k - x^*\|_2^2 + \beta_k\mathbb{E}[(g(x^k) - t_{k+1})^2|\mathcal{F}_k] \\ &+3\mathcal{V}_{f'}(x^*)\tfrac{\alpha_k^2}{N_k} + 2C_t^2\hat{C}_g^2\gamma_k^2 + C_f^2\alpha_k^2, \quad \text{w.p.1.} \end{aligned}$$

Let v_k and δ_k be defined as in the proof of Theorem 3.8. Then, from Lemma A.2, it follows

$$(\delta_k + \beta_k)(1 - \beta_k) \le (1 - \frac{\kappa}{2}\alpha_k)\delta_k, \quad \delta_k \le 2, \quad \delta_{k+1} \le \delta_k, \quad \delta_k + \beta_k \le 3.$$

By multiplying the relation in Lemma 4.5 with $(\delta_k + \beta_k)$ and adding it to the preceding conditional expected inequality, we have

$$\mathbb{E}[v_{k+1}] \le (1 - \frac{\kappa}{2}\alpha_k)\mathbb{E}[v_k] + u_k,$$

where u_k is given by

$$u_k := 18\mathcal{V}_g(x^*)\frac{\beta_k^2}{N_k} + 3C_g^2C_t^2\hat{C}_g^2\frac{\gamma_{k-1}^2}{\beta_k} + 3C_g^2C_f^2\frac{\alpha_{k-1}^2}{\beta_k} + 3\mathcal{V}_{f'}(x^*)\frac{\alpha_k^2}{N_k} + 2C_t^2\hat{C}_g^2\gamma_k^2 + C_f^2\alpha_k^2.$$

By some calculations, we can verify that, in both settings of item (i) and item (ii),

$$\sum_k \alpha_k = \infty, \quad \frac{u_k}{\alpha_k} \to 0, \quad \sum_k u_k < \infty.$$

The rest of the proof is followed by the same analysis in Theorem 3.8. □

Appendix C Proof of Theorem 5.1

The convergence can be proved almost identically to Theorem 4.7. The only difference is that we should carefully handle the random item $\mathcal{I}_k$ here.

As discussed previously, under Assumption 7, we can easily get

$$\mathbb{E}[\|d^{k,i}\|_*^2|\mathcal{F}_k] = \mathbb{E}[\|[t_{k+1}^i]_+ G_i'(x^k,\eta^k)\|_*^2|\mathcal{F}_k] \le C_t^2\hat{C}_g^2, \quad \forall i \in \mathcal{I}$$

with probability 1. This inequality also implies

$$\mathbb{E}[\|d^k\|_*^2|\mathcal{F}_k] \le \frac{1}{\mathcal{I}}\sum_{i\in\mathcal{I}}\mathbb{E}[\|d^{k,i}\|_*^2|\mathcal{F}_k] \le C_t^2\hat{C}_g^2, \quad \text{w.p.1} \tag{C.1}$$

and

$$\mathbb{E}[\|x^{k+1}-x^k\|^2|\mathcal{F}_k] \le \frac{2C_f^2}{\lambda^2}\alpha_k^2 + \frac{2C_t^2\hat{C}_g^2}{\lambda^2}\gamma_k^2, \quad \text{w.p.1}, \tag{C.2}$$

by following the same line as in the proof of Lemma 4.2.

We now start the convergence analysis with the following auxiliary lemma.

Lemma C.1. *Under Assumptions 7 and 8, it holds with probability 1 that*

$$\begin{aligned}\mathbb{E}[\langle -\gamma_k d^k, x^{k+1}-x^*\rangle|\mathcal{F}_k] \quad &\le \left[\tfrac{C_g^2}{2\lambda}\cdot\tfrac{\gamma_k^2}{\beta_k} + 6(C_g^2+\hat{C}_g^2)\tfrac{\beta_k^2}{\lambda N_k}\right]V(x^k,x^*) + \tfrac{\lambda}{8}\mathbb{E}[\|x^{k+1}-x^k\|^2|\mathcal{F}_k]\\ &\quad -\tfrac{\gamma_k}{2C_{eb}}\|x^k-\bar{x}^k\|^2 + C_g^2\|x^k-x^{k-1}\|^2 + \tfrac{\beta_k}{|\mathcal{I}|}\textstyle\sum_{i\in\mathcal{I}}(g_i(x^{k-1})-t_k^i)^2\\ &\quad +2C_t^2\hat{C}_g^2\tfrac{\gamma_k^2}{\lambda} + 3\mathcal{V}_g(x^*)\tfrac{\beta_k^2}{N_k}.\end{aligned}$$

Proof. By using (C.1), we first write

$$\begin{aligned}\mathbb{E}[\langle -\gamma_k d^k, x^{k+1}-x^*\rangle|\mathcal{F}_k] \quad &= \mathbb{E}[\langle -\gamma_k d^k, x^{k+1}-x^k\rangle|\mathcal{F}_k] + \mathbb{E}[\langle -\gamma_k d^k, x^k-x^*\rangle|\mathcal{F}_k]\\ &\le \tfrac{2\gamma_k^2}{\lambda}\mathbb{E}[\|d^k\|_*^2|\mathcal{F}_k] + \tfrac{\lambda}{8}\mathbb{E}[\|x^{k+1}-x^k\|^2|\mathcal{F}_k] + \mathbb{E}[\langle -\gamma_k d^k, x^k-x^*\rangle|\mathcal{F}_k]\\ &\le 2C_t^2\hat{C}_g^2\tfrac{\gamma_k^2}{\lambda} + \tfrac{\lambda}{8}\mathbb{E}[\|x^{k+1}-x^k\|^2|\mathcal{F}_k] + \mathbb{E}[\langle -\gamma_k d^k, x^k-x^*\rangle|\mathcal{F}_k]\end{aligned} \tag{C.3}$$

with probability 1. The rest of the proof is to get the bound of the last term in the above inequality.

It follows from Assumption 8 that

$$-\mathbb{E}\left[\frac{\gamma_k}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\langle [g_i(x^k)]_+ G_i'(x^k,\eta^k), x^k-x^*\rangle|\mathcal{F}_k\right] \le -\frac{\gamma_k}{2C_{eb}}\|x^k-\bar{x}^k\|^2, \quad \text{w.p.1},$$

and it is easy to get that

$$\begin{aligned}&\mathbb{E}\left[\tfrac{\gamma_k}{|\mathcal{I}_k|}\textstyle\sum_{i\in\mathcal{I}_k}\langle G_i'(x^k,\eta^k), x^k-x^*\rangle([g_i(x^k)]_+ - [t_{k+1}^i]_+)|\mathcal{F}_k\right]\\ &\le \mathbb{E}\left[\tfrac{1}{|\mathcal{I}_k|}\textstyle\sum_{i\in\mathcal{I}_k}\left(\tfrac{\gamma_k^2}{4\beta_k}\|G_i'(x^k,\eta^k)\|_*^2\cdot\|x^k-x^*\|^2 + \beta_k([g_i(x^k)]_+ - [t_{k+1}^i]_+)^2|\mathcal{F}_k\right)\right]\\ &\le \tfrac{1}{|\mathcal{I}|}\textstyle\sum_{i\in\mathcal{I}}\left(C_g^2\tfrac{\gamma_k^2}{4\beta_k}\|x^k-x^*\|^2\right) + \mathbb{E}[\tfrac{1}{|\mathcal{I}_k|}\textstyle\sum_{i\in\mathcal{I}_k}\beta_k([g_i(x^k)]_+ - [t_{k+1}^i]_+)^2|\mathcal{F}_k]\\ &\le C_g^2\tfrac{\gamma_k^2}{2\lambda\beta_k}V(x^k,x^*) + \mathbb{E}[\tfrac{1}{|\mathcal{I}_k|}\textstyle\sum_{i\in\mathcal{I}_k}\beta_k(g_i(x^k)-t_{k+1}^i)^2|\mathcal{F}_k], \quad \text{w.p.1}.\end{aligned}$$

Note that

$$d^k = \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}[g_i(x^k)]_+ G_i'(x^k,\eta^k) - \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}G_i'(x^k,\eta^k)([g_i(x^k)]_+ - [t_{k+1}^i]_+),$$

together with the previous two inequalities, it gives that

$$\begin{aligned}&\mathbb{E}[\langle -\gamma_k d^k, x^k-x^*\rangle|\mathcal{F}_k]\\ &\le -\tfrac{\gamma_k}{2C_{eb}}\|x^k-\bar{x}^k\|^2 + C_g^2\tfrac{\gamma_k^2}{2\lambda\beta_k}V(x^k,x^*) + \mathbb{E}[\tfrac{1}{|\mathcal{I}_k|}\textstyle\sum_{i\in\mathcal{I}_k}\beta_k(g_i(x^k)-t_{k+1}^i)^2|\mathcal{F}_k], \quad \text{w.p.1}.\end{aligned} \tag{C.4}$$

From the update rule of t_{k+1}^i, we have

$$(g_i(x^k)-t_{k+1}^i)^2 = ((1-\beta_k)(g_i(x^k)-t_k^i) - \beta_k\varepsilon_{g_i}^k)^2 \le (1-\beta_k)(g_i(x^k)-t_k^i)^2 + \beta_k(\varepsilon_{g_i}^k)^2, \quad i\in\mathcal{I}_k$$

and hence (see also the argument in (B.7))

$$\begin{aligned}
&\mathbb{E}[\tfrac{1}{|\mathcal{I}_k|}\textstyle\sum_{i\in\mathcal{I}_k}\beta_k(g_i(x^k)-t_{k+1}^i)^2|\mathcal{F}_k] \le \mathbb{E}[\tfrac{1}{|\mathcal{I}_k|}\textstyle\sum_{i\in\mathcal{I}_k}[\beta_k(1-\beta_k)(g_i(x^k)-t_k^i)^2+\beta_k^2(\varepsilon_{g_i}^k)^2]|\mathcal{F}_k]\\
&\le \mathbb{E}[\tfrac{1}{|\mathcal{I}_k|}\textstyle\sum_{i\in\mathcal{I}_k}[\beta_k(1-\beta_k)(1+1/\beta_k)(g_i(x^k)-g_i(x^{k-1}))^2+\beta_k(1-\beta_k)(1+\beta_k)(g_i(x^{k-1})-t_k^i)^2\\
&\qquad +\beta_k^2(\varepsilon_{g_i}^k)^2]|\mathcal{F}_k],\\
&\le C_g^2\|x^k-x^{k-1}\|^2+\tfrac{\beta_k}{|\mathcal{I}|}\textstyle\sum_{i\in\mathcal{I}}(g_i(x^{k-1})-t_k^i)^2+\tfrac{\beta_k^2}{N_k}[6(C_g^2+\hat{C}_g^2)V(x^k,x^*)/\lambda+3\mathcal{V}_g(x^*)]
\end{aligned}$$

with probability 1. By substituting the above relation into (C.4), together with (C.3), it gives the claim. □

As mentioned earlier, the first step $y^k=\mathcal{P}_{x^k}(\alpha_k\nabla\mathcal{F}^k)$ in Algorithm 3 is exactly the same as that in Algorithm 2, which means that the relation between y^k and x^k in Lemma 4.3 still holds. Thus, by using this relation and the preceding lemma, we derive the following important recursive relations.

Lemma C.2. *Suppose that Assumptions 6, 7 and 8 hold true. Let x^* be any given optimal solution to problem (5.1). Then, we have*

$$\mathbb{E}[\theta_{k+1}|\mathcal{F}_k]\le(1+a_k)\theta_k-u_k+d_k\zeta_k+\mu_k,\quad w.p.1 \tag{C.5}$$

and

$$\mathbb{E}[\zeta_{k+1}|\mathcal{F}_k]\le(1-d_k)\zeta_k-\bar{u}_k+b_k\theta_k+\nu_k,\quad w.p.1 \tag{C.6}$$

where $\theta_k:=V(x^k,x^*)$, $\zeta_k:=\frac{1}{\mathcal{I}}\sum_{i\in\mathcal{I}}(g_i(x^{k-1})-t_k^i)^2$, $d_k=\beta_k$, $\bar{u}_k=0$,

$$\begin{aligned}
&a_k:=\tfrac{L_f^2}{2\lambda\eta}\cdot\tfrac{\alpha_k^2}{\gamma_k}+6(L_f^2+\hat{L}_f^2)\tfrac{\alpha_k^2}{\lambda^2N_k}+\tfrac{C_g^2}{2\lambda}\cdot\tfrac{\gamma_k^2}{\beta_k}+6(C_g^2+\hat{C}_g^2)\tfrac{\beta_k^2}{\lambda N_k},\\
&u_k:=\alpha_k(f(\bar{x}^k)-f^*+\|x^k-\bar{x}^k\|^2),\\
&b_k:=\tfrac{12\beta_k^2}{\lambda N_k}(C_g^2+\hat{C}_g^2),\quad \nu_k:=\tfrac{6\mathcal{V}_g(x^*)\beta_k^2}{N_k}+\tfrac{C_g^2}{\beta_k}\|x^k-x^{k-1}\|^2
\end{aligned}$$

and

$$\mu_k:=\frac{C_f^2}{4\eta}\cdot\frac{\alpha_k^2}{\gamma_k}+\frac{2\hat{C}_g^2C_t^2}{\lambda}\gamma_k^2+3\mathcal{V}_{f'}(x^*)\frac{\alpha_k^2}{\lambda N_k}+3\mathcal{V}_g(x^*)\frac{\beta_k^2}{N_k}+C_g^2\|x^k-x^{k-1}\|^2.$$

Proof. Substituting $x^+=x^{k+1}$ and $u=x^*$ in (2.3), we have

$$V(x^{k+1},x^*)\le V(y^k,x^*)-V(y^k,x^{k+1})-\langle\gamma_kd^k,x^{k+1}-x^*\rangle.$$

By taking conditional expectation on both sides and applying Lemma 4.3 and Lemma C.1, we get

$$\begin{aligned}
&\mathbb{E}[V(x^{k+1},x^*)|\mathcal{F}_k]\\
&\le(1+a_k)V(x^k,x^*)-\alpha_k(f(\bar{x}^k)-f^*)-(\tfrac{\gamma_k}{2C_{eb}}-\alpha_kL_f-3\eta\gamma_k)\|x^k-\bar{x}^k\|^2\\
&\quad+\tfrac{\lambda}{8}\mathbb{E}[\|x^{k+1}-x^k\|^2|\mathcal{F}_k]-\tfrac{\lambda}{2}\mathbb{E}[\|x^{k+1}-y^k\|^2|\mathcal{F}_k]-\tfrac{\lambda}{4}\mathbb{E}[\|y^k-x^k\|^2|\mathcal{F}_k]\\
&\quad+\tfrac{\beta_k}{|\mathcal{I}|}\textstyle\sum_{i\in\mathcal{I}}(g_i(x^{k-1})-t_k^i)^2+\mu_k\\
&\le(1+a_k)V(x^k,x^*)-\alpha_k(f(\bar{x}^k)-f^*+\|x^k-\bar{x}^k\|^2)+\tfrac{\beta_k}{|\mathcal{I}|}\textstyle\sum_{i\in\mathcal{I}}(g_i(x^{k-1})-t_k^i)^2+\mu_k,
\end{aligned}$$

with probability 1, which is indeed (C.5). In the second inequality above $-\frac{\gamma_k}{2C_{eb}}+\alpha_kL_f+3\eta\gamma_k\le-\alpha_k$ (Assumption 6) is used.

We now prove (C.6). For each $i\in\mathcal{I}$, as an analogy of Lemma 4.5, we have

$$\begin{aligned}
&\mathbb{E}[(g_i(x^k)-t_{k+1}^i)^2|\mathcal{F}_k]\\
&\le(1-\beta_k)(g_i(x^{k-1})-t_k^i)^2+\tfrac{12\beta_k^2}{\lambda N_k}(C_g^2+\hat{C}_g^2)V(x^k,x^*)+\tfrac{6\mathcal{V}_g(x^*)\beta_k^2}{N_k}+\tfrac{C_g^2}{\beta_k}\|x^k-x^{k-1}\|^2.
\end{aligned}$$

Summing the above inequality over $i\in\mathcal{I}$ and dividing by $|\mathcal{I}|$, we get (C.6). □

Now we are ready to prove Theorem 5.1 by using the above lemmas.

Proof of Theorem 5.1. By using (C.5) and (C.6) and Assumption 6, the proof follows the same line of analysis as in Theorem 4.7. □