SIAM J. Optim.
Vol. 26, No. 1, pp. 681–717

STOCHASTIC FIRST-ORDER METHODS WITH RANDOM CONSTRAINT PROJECTION[*]

MENGDI WANG[†] and DIMITRI P. BERTSEKAS[‡]

Abstract. We consider convex optimization problems with structures that are suitable for sequential treatment or online sampling. In particular, we focus on problems where the objective function is an expected value, and the constraint set is the intersection of a large number of simpler sets. We propose an algorithmic framework for stochastic first-order methods using random projection/proximal updates and random constraint updates, which contain as special cases several known algorithms as well as many new algorithms. To analyze the convergence of these algorithms in a unified manner, we prove a general coupled convergence theorem. It states that the convergence is obtained from an interplay between two coupled processes: progress toward feasibility and progress toward optimality. Under suitable stepsize assumptions, we show that the optimality error decreases at a rate of $\mathcal{O}(1/\sqrt{k})$ and the feasibility error decreases at a rate of $\mathcal{O}(\log k/k)$. We also consider a number of typical sampling processes for generating stochastic first-order information and random constraints, which are common in data-intensive applications, online learning, and simulation optimization. By using the coupled convergence theorem as a modular architecture, we are able to analyze the convergence of stochastic algorithms that use arbitrary combinations of these sampling processes.

Key words. large-scale optimization, subgradient, projection, proximal, stochastic approximation, feasibility, randomized algorithm, online optimization

AMS subject classification. 90C25

DOI. 10.1137/130931278

1. Introduction.

Consider the convex optimization problem

$$\min_{x\in X} \quad f(x), \tag{1}$$

where $f:\Re^n\mapsto\Re$ is a convex function (not necessarily differentiable), and X is a nonempty, closed, and convex set in $\Re^n$. We are interested in problems of this form where the constraint set X is the intersection of a finite number of sets, i.e.,

$$X=\cap_{i=1}^m X_i, \tag{2}$$

with each X_i being a closed and convex subset of $\Re^n$. We also allow the objective function f to be the sum of a large number of component functions, or more generally to be expressed as the expected value

$$f(x)=\mathbf{E}\big[f_v(x)\big], \tag{3}$$

where $f_v:\Re^n\mapsto\Re$ is a function of x involving a random variable v.

Two classical methods for solution of problem (1) are the subgradient projection method (or projection method for short) and the proximal method. The projection

[*] Received by the editors July 31, 2013; accepted for publication (in revised form) December 28, 2015; published electronically March 10, 2016.
http://www.siam.org/journals/siopt/26-1/93127.html

[†] Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08540 (mengdiw@princeton.edu).

[‡] Department of Electrical Engineering and Computer Science and the Laboratory for Information and Decision Systems (LIDS), MIT, Cambridge, MA 02139 (dimitrib@mit.edu). This author's research was supported by Air Force grant FA9550-10-1-0412.

method has the form

$$x_{k+1} = \Pi\big[x_k - \alpha_k \tilde{\nabla} f(x_k)\big],$$

where Π denotes the Euclidean orthogonal projection onto X, $\{\alpha_k\}$ is a sequence of constant or diminishing positive scalars, and $\tilde{\nabla} f(x_k)$ is a subgradient of f at x_k (a vector g is a subgradient of f at x if $g'(y-x) \le f(y) - f(x)$ for any $y \in \Re^n$). The proximal method has the form

$$x_{k+1} = \text{argmin}_{x\in X} \left[f(x) + \frac{1}{2\alpha_k}\|x - x_k\|^2 \right]$$

and can be equivalently written as

$$x_{k+1} = \Pi\big[x_k - \alpha_k \tilde{\nabla} f(x_{k+1})\big]$$

for some subgradient $\tilde{\nabla} f(x_{k+1})$ of f at x_{k+1} (see [Ber11, Proposition 1]). In this way, the proximal method has a form similar to that of the projection method. This enables us to analyze these two methods and their mixed versions with a unified analysis.

In practice, these methods are often difficult to use, especially when the constraint set X is complicated (cf. (2)). At every iteration, the projection method requires the computation of the Euclidean projection, and the proximal method requires solving a constrained minimization, both of which can be time-consuming (or even impossible with limited memory space). In the case where X is the intersection of a large number of simpler sets X_i, it is possible to improve the efficiency of a single iteration for these methods, by operating with a single set X_i at a time.

Another difficulty arises when f either is the sum of a large number of component functions, $f(x) = \sum_i f_i(x)$ or, more generally, is an expected value, i.e., $f(x) = \mathbf{E}\big[f_v(x)\big]$. Then the exact computation of a subgradient $\tilde{\nabla} f(x_k)$ can be either very expensive or impossible due to noise. To address this difficulty, we may use in place of $\tilde{\nabla} f(x_k)$ in the projection method a stochastic sample subgradient $g(x_k, v_k)$. Similarly, we may use in place of $f(x)$ in the proximal method a sample component function $f_{v_k}(x)$.

In this paper, we propose to modify and combine the projection and proximal methods, in order to process the constraints X_i and the component functions $f_v(\cdot)$ sequentially or "online." The purpose of this paper is to present a unified analytical framework for these methods and their extensions. In particular, we focus on the class of methods that interact with a probabilistic process that generates stochastic first-order information and randomized constraints.

Suppose that we have access to a *sampling oracle* ($\mathcal{SO}$) such that

- given a vector $x \in \Re^n$, it returns a random subgradient vector $g(x, v)$ or a random function $f_v(\cdot)$;
- given a vector $z \in \Re^n$, it returns a random projection $\Pi_{w_k} z$.

Here Π_{w_k} denotes the Euclidean projection onto a set X_{w_k}, $\{w_k\}$ is a sequence of random variables taking values in $\{1, \dots, m\}$, and $\{v_k\}$ is a sequence of random variables generated by some probabilistic process. The sampling oracle is a model that is widely used in stochastic and online optimization. In particular, optimization problems arising from statistical estimation and machine learning problems are often driven by large data sets or even streaming data. These applications naturally come with a sampling oracle, i.e., the oracle that accesses the data points. We will state a variety of assumptions on the stochastic sampling process $\{(v_k, w_k)\}$ in subsequent analysis.

Algorithm 1. Stochastic first-order method with random constraint projection.

```
Input: $x_0, z_0 \in \Re^n$, $\mathcal{SO}$, $\{\alpha_k\}$, $\{\beta_k\}$.
for $k = 0, 1, 2, \ldots$ do
    Query the $\mathcal{SO}$ at $x_k$ and obtain either $g(x_k, v_k)$ or $f_{v_k}(\cdot)$.
    Perform either a stochastic gradient descent step,
        $x_{k+1} = \Pi_{w_k}\big[x_k - \alpha_k g(x_k, v_k)\big]$,
    or a stochastic proximal step,
        $x_{k+1} = \mathrm{argmin}_{x \in X_{w_k}} \left[ f_{v_k}(x) + \frac{1}{2\alpha_k}\|x - x_k\|^2 \right]$.
    Query the $\mathcal{SO}$ at $z_k$ and obtain a random projection $\Pi_{w_k} z_k$.
    Update the iterate as
        $x_{k+1} = z_k - \beta_k\left(z_k - \Pi_{w_k} z_k\right)$.
end for
```

We propose an algorithmic framework that involves random optimality updates and random feasibility updates, which are summarized in Algorithm 1. Algorithm 1 updates the iterates by interacting closely with the $\mathcal{SO}$. It is a mixed version of the gradient projection and the proximal methods. In the case where only stochastic gradients are available and $\beta_k = 1$, Algorithm 1 becomes the *random projection algorithm* given by

$$x_{k+1} = \Pi_{w_k}\big[x_k - \alpha_k g(x_k, v_k)\big],$$

and a special case of this algorithm using exact gradient has been considered by Nedić [Ned11]. In the other case where only proximal steps are used and $\beta_k = 1$, the corresponding iteration becomes a *random constraint proximal algorithm*, given by

$$x_{k+1} = \mathrm{argmin}_{x \in X_{w_k}} \left[f_{v_k}(x) + \frac{1}{2\alpha_k}\|x - x_k\|^2 \right] = \Pi_{w_k}\big[x_k - \alpha_k g(x_{k+1}, v_k)\big], \tag{4}$$

which, to the best of our knowledge, has never been considered in the literature. Another interesting case is when f has the form

$$f(x) = \sum_{i=1}^{N} h_i(x) + \sum_{i=1}^{N} \hat{h}_i(x),$$

where h_i are functions whose subgradients are easy to compute, $\hat{h}_i$ are functions that are suitable for the proximal iteration, and a sample component function f_v may belong to either $\{h_i\}$ or $\{\hat{h}_i\}$. In this case, our algorithm can adaptively choose between a projection step and a proximal step, based on the current sample component function.

We notice from (4) that the proximal step is very similar to a gradient update. This allows us to analyze the projection and proximal algorithms together under the same framework. In fact, the iteration of Algorithm 1 is equivalent to

$$z_k = x_k - \alpha_k g(\bar{x}_k, v_k), \qquad x_{k+1} = z_k - \beta_k\left(z_k - \Pi_{w_k} z_k\right), \tag{5}$$

where $\bar{x}_k$ can take either of the following values:

$$\bar{x}_k = x_k, \qquad \text{or} \qquad \bar{x}_k = x_{k+1}. \tag{6}$$

Note that although (5) may involve x_{k+1}, its implementation given by Algorithm 1 does not look ahead into the future. In our analysis, we focus on iteration (5). To make the algorithm as general as possible, we have left open the formal definition of $g(\cdot, v_k)$ and simply regarded it as a noisy evaluation of some subgradient. More details will be specified in section 5 when we revisit the $\mathcal{SO}$ for first-order information.

A major contribution of this work is to propose a unified algorithmic and analytical framework for stochastic first-order methods with constraint randomization. Our Algorithm 1 (which is equivalent to (5)–(6)) can be viewed as alternating between two types of iterations with different objectives: to approach the feasible set and to approach the set of optimal solutions. We will provide a coupled supermartingale convergence lemma (Lemma 7) for the first time, which establishes the convergence of two entangled supermartingales. Then we will provide a coupled convergence theorem (Theorem 1) which requires that the algorithm operate on two different time scales: the convergence to the feasible set, which is controlled by β_k, should have a smaller modulus of contraction than the convergence to the optimal solution, which is controlled by α_k. This coupled improvement mechanism is the key to the almost sure convergence of optimization methods involving randomness in both feasibility and optimality updates. It also provides a modular architecture that can be used to analyze new variants of the algorithms or new assumptions about the $\mathcal{SO}$.

A second contribution of this work is the convergence rate analysis (Theorem 2). Under suitable stepsize assumptions, we prove that the optimality error diminishes to zero at a rate of $\mathcal{O}(1/\sqrt{k})$, while the feasibility error diminishes to zero at a rate of $\mathcal{O}(\log k/k)$. This is consistent with our theory that the convergence has two time scales. More importantly, the convergence rate $\mathcal{O}(1/\sqrt{k})$ is nonimprovable in terms of sample-error complexity. As long as we want to optimize a convex function using noisy first-order information, the error bound $\mathcal{O}(1/\sqrt{k})$ is already optimal with respect to the sample size. This suggests that randomizing the constraint projection *does not* slow down the stochastic convergence (up to a constant).

Another contribution of our analysis relates to the $\mathcal{SO}$ for obtaining the samples v_k and w_k. For example, a common situation arises from applications based on large data sets. Then each component $f(\cdot, v)$ and constraint X_w may relate to a piece of data, so that accessing all of them requires passing through the entire data set. This forces the algorithm to process the components/constraints sequentially, according to either a fixed order or by random sampling. There are also situations in which the component functions or the constraints can be selected adaptively based on the iterates' history. In this work, we will consider several typical cases for generating the random variables w_k and v_k, which we list below and define more precisely later:

- Sampling schemes for constraints X_{w_k}:
  - The samples are nearly independent and all the constraint indexes are visited sufficiently often.
  - The samples are "cyclic," e.g., are generated according to either a deterministic cyclic order or a random permutation of the indexes within a cycle.
  - The samples are selected to be the most distant constraint supersets to the current iterates.
  - The samples are generated according to an irreducible Markov chain with an appropriate invariant distribution.

- Sampling schemes for subgradients $g(\cdot, v_k)$ or component functions f_{v_k}:
  - The samples are conditionally unbiased.
  - The samples are "cyclically obtained" by either a fixed order or random shuffling.

We will consider *all combinations of the preceding sampling schemes* and show that our unified convergence analysis applies to all of them. While it is beyond our scope to identify all possible sampling schemes that may be interesting, one of the goals of the current paper is to propose a unified framework, both algorithmic and analytic, that can be easily adapted to new sampling schemes and algorithms.

Related works. The proposed Algorithm 1 contains as special cases a number of known methods from convex optimization, feasibility, and stochastic approximation. In view of these connections, our analysis uses several ideas from the literature which we will now summarize.

The feasibility update of Algorithm 1 is related to known methods for feasibility problems. In particular, when $f(x) = 0$, $g(\bar{x}_k, v_k) = 0$ and $\beta_k = 1$, we obtain a successive projection algorithm for finding some $x \in X = \cap_{i=1}^m X_i$. Successive projection methods have a long history, starting with von Neumann [vN50], followed by many other authors: Halperin [Hal62], Gubin, Polyak, and Raik [GPR67], Tseng [Tse90], Bauschke, Borwein, and Lewis [BBL97], Deutsch and Hundal [DH06a], [DH06b], [DH08], Cegielski and Suchocka [CS08], Lewis and Malick [LM08], Leventhal and Lewis [LL10], and Nedić [Ned10]. A survey of the work in this area up to 1996 is given by Bauschke [Bau96].

The use of stochastic subgradients in Algorithm 1 is closely related to stochastic approximation methods. In the case where $X = X_{w_k}$ for all k, our method becomes a stochastic approximation method for optimization problems, which has been well known in the literature. Similar to several sources on convergence analysis of stochastic algorithms, we use a supermartingale convergence theorem (see, e.g., the textbooks by Bertsekas and Tsitsiklis [BT89], by Kushner and Yin [KY03], and by Borkar [Bor08]).

Algorithms using random constraint updates for optimization problems of the form (1) were first considered by Nedić [Ned11]. This work proposed a projection method that updates using exact subgradients and randomized selection of constraint sets, which can be viewed as a special case of Algorithm 1 with $\bar{x}_k = x_k$. The work of [Ned11] is less general than the current work in that it does not consider the proximal method, it does not use random samples of subgradients, and it considers only a special case of constraint randomization.

Another closely related work is Bertsekas [Ber11]. It proposed an algorithmic framework that alternates incrementally between subgradient and proximal iterations for minimizing a cost function $f = \sum_{i=1}^m f_i$, the sum of a large but finite number of convex components f_i, over a constraint set X. This can be viewed as a special case of Algorithm 1 with $X_{w_k} = X$. The choice between random and cyclic selection of the components f_i for iteration is a major point of analysis of these methods, similar to earlier works on incremental subgradient methods by Nedić and Bertsekas [NB00], [NB01], [BNO03]. It is less general than the current work in that it does not consider the randomization of constraints, and it requires the objective function to be Lipchitz continuous.

Recently, the idea of constraint randomization has been extended to the solution of variational inequalities, by Wang and Bertsekas in [WB12]. This work is by far the most related to the current work but focuses on a different problem: finding x^* such that $F(x^*)'(x-x^*) \geq 0$ for all $x \in \cap_{i=1}^m X_i$, where $F : \Re^n \mapsto \Re^n$ is a strongly monotone mapping (i.e., $(F(x)-F(y))'(x-y)\geq\sigma\|x-y\|^2$ for some $\sigma>0$ and all $x,y\in\Re^n$). The work of [WB12] modifies the projection method for variational inequalities to use cyclic/random constraint projection and analyzes the convergence performances. This work is related to the present paper in that it addresses a problem that contains the minimization of a differentiable strongly convex function as a special case (whose optimality condition is a strongly monotone variational inequality) and shares some analytical ideas. However, the present paper proposes a substantially more general framework that applies to convex nonsmooth optimization and is based on the new coupled convergence theorem for two time-scale processes, which provides a modular architecture for analyzing new algorithms as well as new sampling schemes.

Outline. Section 2 summarizes our basic assumptions and a few preliminary results. Section 3 proves the coupled convergence theorem and rate of convergence theorem, which, assuming a feasibility improvement condition and an optimality improvement condition, establishes convergence of the general Algorithm 1. Section 4 considers sampling schemes for the constraint sets such that the feasibility improvement condition is satisfied. Section 5 considers sampling schemes for the subgradients or objective functions such that the optimality improvement condition is satisfied. Section 6 collects various sets of conditions under which the convergence of the stochastic algorithms can be achieved. Section 7 gives some numerical results, and section 8 summarizes the current work.

Notation. All vectors in the n-dimensional Euclidean space $\Re^n$ will be viewed as column vectors. For $x\in\Re^n$, we denote by x' its transpose and by $\|x\|$ its Euclidean norm (i.e., $\|x\|=\sqrt{x'x}$). For two sequences of nonnegative scalars $\{y_k\}$ and $\{z_k\}$, we write $y_k=\mathcal{O}(z_k)$ if there exists a constant $c>0$ such that $y_k\leq cz_k$ for each k, and we write $y_k=\Theta(z_k)$ if there exist constants $c_1>c_2>0$ such that $c_2z_k\leq y_k\leq c_1z_k$ for each k. We denote by $\partial f(x)$ the subdifferential (the set of all subgradients) of f at x, denote by X^* the set of optimal solutions for problem (1), and denote by $f^*=\inf_{x\in X}f(x)$ the optimal value. The abbreviation "$\xrightarrow{a.s.}$" means "converges almost surely to," while the abbreviation "i.i.d." means "independent and identically distributed."

2. Assumptions and preliminaries.

To motivate our analysis, we first briefly review the convergence mechanism of the deterministic subgradient projection method

$$x_{k+1}=\Pi\big[x_k-\alpha_k\tilde\nabla f(x_k)\big], \tag{7}$$

where Π denotes the Euclidean orthogonal projection on X. We assume for simplicity that $\|\tilde\nabla f(x)\|\leq L$ for all x and that there exists at least one optimal solution x^* of problem (1). Then we have

$$\begin{aligned}\|x_{k+1}-x^*\|^2&=\big\|\Pi\big[x_k-\alpha_k\tilde\nabla f(x_k)\big]-x^*\big\|^2\\&\leq\big\|\big(x_k-\alpha_k\tilde\nabla f(x_k)\big)-x^*\big\|^2\\&=\|x_k-x^*\|^2-2\alpha_k\tilde\nabla f(x_k)'(x_k-x^*)+\alpha_k^2\big\|\tilde\nabla f(x_k)\big\|^2\\&\leq\|x_k-x^*\|^2-2\alpha_k\big(f(x_k)-f^*\big)+\alpha_k^2L^2,\end{aligned} \tag{8}$$

where the first inequality uses the fact $x^*\in X$ and the nonexpansiveness of the projection, i.e.,

$$\|\Pi x-\Pi y\|\leq\|x-y\|\qquad\forall\;\;x,y\in\Re^n,$$

and the second inequality uses the definition of the subgradient $\tilde\nabla f(x)$, i.e.,

$$\tilde\nabla f(x)'(y-x)\leq f(y)-f(x)\qquad\forall\;\;x,y\in\Re^n.$$

A key fact is that since $x_k \in X$, the value $\big(f(x_k) - f^*\big)$ must be nonnegative. From (8) by taking $k \to \infty$, we have

$$\limsup_{k\to\infty} \|x_{k+1} - x^*\|^2 \leq \|x_0 - x^*\|^2 - 2\sum_{k=0}^{\infty} \alpha_k\big(f(x_k) - f^*\big) + \sum_{k=0}^{\infty} \alpha_k^2 L^2.$$

Assuming that $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, we can use a standard argument to show that $\|x_k - x^*\|$ is convergent for all $x^* \in X^*$ and

$$\sum_{k=0}^{\infty} \alpha_k\big(f(x_k) - f^*\big) < \infty,$$

which implies that $\liminf_{k\to\infty} f(x_k) = f^*$. Finally, by using the continuity of f, we can show that the iterates x_k must converge to some optimal solution of problem (1).

Our proposed Algorithm 1, which is equivalent to

$$z_k = x_k - \alpha_k g(\bar{x}_k, v_k), \quad x_{k+1} = z_k - \beta_k\left(z_k - \Pi_{w_k} z_k\right), \quad \text{with } \bar{x}_k = x_k \text{ or } \bar{x}_k = x_{k+1}, \tag{9}$$

differs from the classical method (7) in a fundamental way: the iterates $\{x_k\}$ generated by the algorithm (9) are not guaranteed to stay in X. Moreover, the projection Π_{w_k} onto a random set X_{w_k} need not decrease the distance between x_k and X at every iteration. As a result, the analogue of the fundamental bound (8) now includes the distance of x_k from X, which need not decrease at each iteration. We will show that $\{x_k\}$ approaches the feasible set X in a stochastic sense as $k \to \infty$. This idea is also implicit in the analyses of [Ned11] and [WB12].

To analyze the stochastic iteration (9), we denote by $\mathcal{F}_k$ the collection of random variables

$$\mathcal{F}_k = \{v_0, \ldots, v_{k-1}, w_0, \ldots, w_{k-1}, z_0, \ldots, z_{k-1}, \bar{x}_0, \ldots, \bar{x}_{k-1}, x_0, \ldots, x_k\}.$$

Moreover, we denote by

$$\mathrm{d}(x) = \|x - \Pi x\|$$

the Euclidean distance of any $x \in \Re^n$ from X.

Let us outline the convergence proof for iteration (9) with i.i.d. random projection and $\bar{x}_k = x_k$. Similar to the classical projection method (7), our line of analysis starts with a bound of the iteration error that has the form

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - 2\alpha_k \tilde{\nabla} f(x_k)'(x_k - x^*) + e(x_k, \alpha_k, \beta_k, w_k, v_k), \tag{10}$$

where $e(x_k, \alpha_k, \beta_k, w_k, v_k)$ is a random variable. Under suitable assumptions, we will bound each term on the right side of (10) and then take conditional expectation on both sides. From this we will obtain that the iteration error is "stochastically decreasing" in the following sense:

$$\begin{aligned}\mathbf{E}\big[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\big] \leq (1 + \epsilon_k)\|x_k - x^*\|^2 - 2\alpha_k\big(f(\Pi x_k) - f(x^*)\big)\\ + \mathcal{O}(\beta_k)\,\mathrm{d}^2(x_k) + \epsilon_k \qquad \text{w.p.1},\end{aligned}$$

where ϵ_k are positive errors such that $\sum_{k=0}^{\infty} \epsilon_k < \infty$. On the other hand, by using properties of random projection, we will obtain that the feasibility error $\mathrm{d}^2(x_k)$ is "stochastically decreasing" at a faster rate, according to

$$\mathbf{E}\left[\,\mathrm{d}^2(x_{k+1}) \mid \mathcal{F}_k\right] \leq \big(1 - \mathcal{O}(\beta_k)\big)\,\mathrm{d}^2(x_k) + \epsilon_k\big(\|x_k - x^*\|^2 + 1\big) \qquad \text{w.p.1}.$$

Finally, based on the preceding two inequalities and through a series of intermediate results, we will end up using the following supermartingale convergence lemma due to Robbins and Siegmund [RS57] to prove an extension and a two-coupled-sequence supermartingale convergence lemma, and then complete the convergence proof of our algorithm.

LEMMA 1. *Let $\{\xi_k\}$, $\{u_k\}$, $\{\eta_k\}$, and $\{\mu_k\}$ be sequences of nonnegative random variables such that*

$$\mathbf{E}\left[\xi_{k+1} \mid \mathcal{G}_k\right] \leq (1+\eta_k)\xi_k - u_k + \mu_k \qquad \forall \quad k \geq 0 \quad w.p.1,$$

where $\mathcal{G}_k$ denotes the collection $\xi_0, \ldots, \xi_k, u_0, \ldots, u_k, \eta_0, \ldots, \eta_k, \mu_0, \ldots, \mu_k$, and

$$\sum_{k=0}^{\infty} \eta_k < \infty, \qquad \sum_{k=0}^{\infty} \mu_k < \infty, \qquad w.p.1.$$

Then the sequence of random variables $\{\xi_k\}$ converges almost surely to a nonnegative random variable, and we have

$$\sum_{k=0}^{\infty} u_k < \infty \qquad w.p.1.$$

This line of analysis is shared with incremental subgradient and proximal methods (see [NB00], [NB01], [Ber11]). However, here the technical details are more intricate because there are two types of iterations, which involve the two different stepsizes α_k and β_k. We will now introduce our assumptions and give a few preliminary results that will be used in the subsequent analysis.

Our first assumption requires that the norm of any subgradient of f be bounded from above by a linear function, which implies that f is bounded by a quadratic function. It also requires that the random samples $g(x, v_k)$ satisfy bounds that involve a multiple of $\|x\|$.

ASSUMPTION 1. *The set of optimal solutions X^* of problem* (1) *is nonempty. Moreover, there exists a constant $L > 0$ such that*

(a) *for any $\tilde{\nabla} f(x) \in \partial f(x)$,*

$$\left\|\tilde{\nabla} f(x)\right\|^2 \leq L^2\big(\|x\|^2 + 1\big) \qquad \forall\ x \in \Re^n,$$

(b)

$$\|g(x, v_k) - g(y, v_k)\| \leq L\big(\|x - y\| + 1\big) \quad \forall\ x, y \in \Re^n, \quad k = 0, 1, 2, \ldots, \quad w.p.1,$$

(c)

$$\mathbf{E}\left[\left\|g(x, v_k)\right\|^2 \mid \mathcal{F}_k\right] \leq L^2\big(\|x\|^2 + 1\big) \quad \forall\ x \in \Re^n \quad w.p.1. \tag{11}$$

Assumption 1 is very general. It contains as special cases a number of conditions that have been frequently assumed in the literature. More specifically, it allows f to be Lipchitz continuous or to have Lipchitz continuous gradient. It also allows f to be nonsmooth and have bounded subgradients. Moreover, it allows f to be a nonsmooth approximation of a smooth function with Lipchitz continuous gradient, e.g., a piecewise linear approximation of a quadratic-like function. These assumptions can be verified easily in many practical applications. Consider, for example, the statistical estimation problems. In these problems, the objective is often the sum of many loss functions, where each loss function is associated with a data point. The loss function is often a negative log-likelihood function. With some knowledge about the likelihood function and distribution of data points, it is usually quite straightforward to verify Assumption 1.

The next assumption includes a standard stepsize condition on α_k, widely used in the literature of stochastic approximation. Moreover, it imposes a certain relationship between the sequences $\{\alpha_k\}$ and $\{\beta_k\}$, which is the key to the coupled convergence process of the proposed algorithm.

Assumption 2. *The stepsize sequences $\{\alpha_k\}$ and $\{\beta_k\}$ are deterministic and nonincreasing and satisfy $\alpha_k \in (0,1)$, $\beta_k \in (0,2)$ for all k, $\lim_{k\to\infty} \beta_k/\beta_{k+1} = 1$, and*

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty, \qquad \sum_{k=0}^{\infty} \beta_k = \infty, \qquad \sum_{k=0}^{\infty} \frac{\alpha_k^2}{\beta_k} < \infty.$$

The condition $\sum_{k=0}^{\infty} \frac{\alpha_k^2}{\beta_k} < \infty$ essentially restricts β_k either to be a constant in $(0,2)$ for all k or to decrease to 0 at a certain rate. Given that $\sum_{k=0}^{\infty} \alpha_k = \infty$, this condition implies that $\liminf_{k\to\infty} \frac{\alpha_k}{\beta_k} = 0$. We will show that as a consequence, the convergence to the feasible set has a better modulus of contraction than the convergence to the optimal solution. This is necessary for the almost sure convergence of the coupled process.

Let us now prove a few preliminary technical lemmas. The first one gives several basic facts regarding projection and has been proved in [WB12, Lemma 1], but we repeat it here for completeness.

Lemma 2. *Let S be a closed convex subset of $\Re^n$, and let Π_S denote orthogonal projection onto S.*

(a) *For all $x \in \Re^n$, $y \in S$, and $\beta > 0$,*

$$\left\|x - \beta(x - \Pi_S x) - y\right\|^2 \le \|x - y\|^2 - \beta(2-\beta)\|x - \Pi_S x\|^2.$$

(b) *For all $x, y \in \Re^n$,*

$$\|y - \Pi_S y\|^2 \le 2\|x - \Pi_S x\|^2 + 8\|x - y\|^2.$$

The second lemma gives a decomposition of the iteration error (cf. (10)), which will serve as the starting point of our analysis.

Lemma 3. *For any $\epsilon > 0$ and $y \in X$, the sequence $\{x_k\}$ generated by Algorithm 1 is such that*

$$\begin{aligned}\|x_{k+1} - y\|^2 &\le \|x_k - y\|^2 - 2\alpha_k g(\bar{x}_k, v_k)'(x_k - y) + \alpha_k^2 \|g(\bar{x}_k, v_k)\|^2 \\ &\quad - \beta_k(2-\beta_k)\|\Pi_{w_k} z_k - z_k\|^2 \\ &\le (1+\epsilon)\|x_k - y\|^2 + (1 + 1/\epsilon)\alpha_k^2 \|g(\bar{x}_k, v_k)\|^2 \\ &\quad - \beta_k(2-\beta_k)\|\Pi_{w_k} z_k - z_k\|^2.\end{aligned}$$

Proof. From Lemma 2(a) and the relations $x_{k+1} = z_k - \beta_k(z_k - \Pi_{w_k} z_k)$, $z_k = x_k - \alpha_k g(\bar{x}_k, v_k)$ (cf. (9)), we obtain

$$\begin{aligned}\|x_{k+1} - y\|^2 &\le \|z_k - y\|^2 - \beta_k(2-\beta_k)\|\Pi_{w_k} z_k - z_k\|^2 \\ &= \|x_k - y - \alpha_k g(\bar{x}_k, v_k)\|^2 - \beta_k(2-\beta_k)\|\Pi_{w_k} z_k - z_k\|^2 \\ &= \|x_k - y\|^2 - 2\alpha_k g(\bar{x}_k, v_k)'(x_k - y) + \alpha_k^2 \|g(\bar{x}_k, v_k)\|^2 \\ &\quad - \beta_k(2-\beta_k)\|\Pi_{w_k} z_k - z_k\|^2 \\ &\le (1+\epsilon)\|x_k - y\|^2 + (1 + 1/\epsilon)\alpha_k^2 \|g(\bar{x}_k, v_k)\|^2 \\ &\quad - \beta_k(2-\beta_k)\|\Pi_{w_k} z_k - z_k\|^2,\end{aligned}$$

where the last inequality uses the fact $2a'b \le \epsilon\|a\|^2 + (1/\epsilon)\|b\|^2$ for any $a, b \in \Re^n$. □

The third lemma gives several basic upper bounds on quantities relating to x_{k+1}, conditioned on the iterates' history up to the kth sample.

LEMMA 4. *Let Assumptions* 1 *and* 2 *hold, let x^* be a given optimal solution of problem* (1), *and let $\{x_k\}$ be generated by Algorithm* 1. *Then for all $k \ge 0$, with probability* 1,

(a) $\mathbf{E}\left[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\right] \le \mathcal{O}\left(\|x_k - x^*\|^2 + \alpha_k^2\right)$,

(b) $\mathbf{E}\left[\,\mathrm{d}^2(x_{k+1}) \mid \mathcal{F}_k\right] \le \mathcal{O}\left(\,\mathrm{d}^2(x_k) + \alpha_k^2\|x_k - x^*\|^2 + \alpha_k^2\right)$,

(c) $\mathbf{E}\left[\|g(\bar{x}_k, v_k)\|^2 \mid \mathcal{F}_k\right] \le O\big(\|x_k - x^*\|^2 + 1\big)$,

(d) $\mathbf{E}\left[\big\|\bar{x}_k - x_k\big\|^2 \mid \mathcal{F}_k\right] \le \mathbf{E}\left[\big\|x_{k+1} - x_k\big\|^2 \mid \mathcal{F}_k\right] \le \mathcal{O}(\alpha_k^2)\big(\|x_k - x^*\|^2 + 1\big) + \mathcal{O}(\beta_k^2)\,\mathrm{d}^2(x_k)$.

Proof. We will prove parts (c) and (d) first and prove parts (a) and (b) later.

(c), (d) By using the basic inequality $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ for $a, b \in \Re^n$ and then applying Assumption 1, we have

$$\begin{aligned}\mathbf{E}\left[\|g(\bar{x}_k, v_k)\|^2 \mid \mathcal{F}_k\right] &\le 2\mathbf{E}\left[\|g(x_k, v_k)\|^2 \mid \mathcal{F}_k\right] + 2\mathbf{E}\left[\|g(\bar{x}_k, v_k) - g(x_k, v_k)\|^2 \mid \mathcal{F}_k\right] \\ &\le O\big(\|x_k - x^*\|^2 + 1\big) + \mathcal{O}\left(\mathbf{E}\left[\|\bar{x}_k - x_k\|^2 \mid \mathcal{F}_k\right]\right).\end{aligned} \tag{12}$$

Since $\bar{x}_k \in \{x_k, x_{k+1}\}$ and $X \subset X_{w_k}$, we use the equivalent form (9) of the algorithm and obtain

$$\|\bar{x}_k - x_k\| \le \|x_{k+1} - x_k\big\| \le \alpha_k\|g(\bar{x}_k, v_k)\| + \beta_k\|z_k - \Pi_{w_k} z_k\| \le \alpha_k\|g(\bar{x}_k, v_k)\| + \beta_k\,\mathrm{d}(z_k)$$

so that

$$\|\bar{x}_k - x_k\|^2 \le \|x_{k+1} - x_k\big\|^2 \le 2\alpha_k^2\|g(\bar{x}_k, v_k)\|^2 + 2\beta_k^2\,\mathrm{d}^2(z_k).$$

Note that from Lemma 2(b) we have

$$\mathrm{d}^2(z_k) \le 2\,\mathrm{d}^2(x_k) + 8\|x_k - z_k\|^2 = 2\,\mathrm{d}^2(x_k) + 8\alpha_k^2\|g(\bar{x}_k, v_k)\|^2.$$

Then it follows from the preceding two relations that

$$\|\bar{x}_k - x_k\|^2 \le \|x_{k+1} - x_k\big\|^2 \le \mathcal{O}(\alpha_k^2)\|g(\bar{x}_k, v_k)\|^2 + \mathcal{O}(\beta_k^2)\,\mathrm{d}^2(x_k). \tag{13}$$

By taking expectation on both sides of (13) and applying (12), we obtain

$$\begin{aligned}\mathbf{E}\left[\|\bar{x}_k - x_k\|^2 \mid \mathcal{F}_k\right] &\le \mathbf{E}\left[\|x_{k+1} - x_k\|^2 \mid \mathcal{F}_k\right] \\ &\le \mathcal{O}(\alpha_k^2)(\|x_k - x^*\|^2 + 1) + \mathcal{O}(\alpha_k^2)\mathbf{E}\left[\|\bar{x}_k - x_k\|^2 \mid \mathcal{F}_k\right] \\ &\quad + \mathcal{O}(\beta_k^2)\,\mathrm{d}^2(x_k),\end{aligned}$$

and by rearranging terms in the preceding inequality, we obtain part (d). Finally, we apply part (d) to (12) and obtain

$$\begin{aligned}\mathbf{E}\left[\|g(\bar{x}_k, v_k)\|^2 \mid \mathcal{F}_k\right] &\le \mathcal{O}(\|x_k - x^*\|^2 + 1) + \mathcal{O}(\alpha_k^2)\big(\|x_k - x^*\|^2 + 1\big) + \mathcal{O}(\beta_k^2)\,\mathrm{d}^2(x_k)\\ &\le \mathcal{O}(\|x_k - x^*\|^2 + 1),\end{aligned}$$

where the second inequality uses the fact $\beta_k \le 2$ and $\mathrm{d}(x_k) \le \|x_k - x^*\|$. Thus we have proved part (c).

(a), (b) Let y be an arbitrary vector in X, and let ϵ be a positive scalar. By using Lemma 3 and part (c), we have

$$\begin{aligned}\mathbf{E}\left[\|x_{k+1} - y\|^2 \mid \mathcal{F}_k\right] &\le (1+\epsilon)\|x_k - y\|^2 + (1 + 1/\epsilon)\alpha_k^2 \mathbf{E}\left[\|g(\bar{x}_k, v_k)\|^2 \mid \mathcal{F}_k\right]\\ &\le (1+\epsilon)\|x_k - y\|^2 + (1 + 1/\epsilon)\alpha_k^2 O\big(\|x_k - x^*\|^2 + 1\big).\end{aligned}$$

By letting $y = x^*$, we obtain (a). By letting $y = \Pi x_k$ and using $\mathrm{d}(x_{k+1}) \le \|x_{k+1} - \Pi x_k\|$, we obtain (b). □

The next lemma is an extension of Lemma 4. It gives the basic upper bounds on quantities relating to x_{k+N}, conditioned on the iterates' history up to the kth samples, with N being a fixed integer.

LEMMA 5. *Let Assumptions* 1 *and* 2 *hold, let x^* be a given optimal solution of problem* (1)*, let $\{x_k\}$ be generated by by Algorithm* 1*, and let N be a given positive integer. Then for all $k \ge 0$, with probability* 1*,*

(a) $\mathbf{E}\left[\|x_{k+N} - x^*\|^2 \mid \mathcal{F}_k\right] \le \mathcal{O}\left(\|x_k - x^*\|^2 + \alpha_k^2\right)$,

(b) $\mathbf{E}\left[\,\mathrm{d}^2(x_{k+N}) \mid \mathcal{F}_k\right] \le \mathcal{O}\left(\,\mathrm{d}^2(x_k) + \alpha_k^2\|x_k - x^*\|^2 + \alpha_k^2\right)$,

(c) $\mathbf{E}\left[\|g(\bar{x}_{k+N}, v_{k+N})\|^2 \mid \mathcal{F}_k\right] \le O\big(\|x_k - x^*\|^2 + 1\big)$,

(d) $\mathbf{E}\left[\|x_{k+N} - x_k\|^2 \mid \mathcal{F}_k\right] \le \mathcal{O}(N^2\alpha_k^2)\big(\|x_k - x^*\|^2 + 1\big) + \mathcal{O}(N^2\beta_k^2)\,\mathrm{d}^2(x_k)$.

Proof. (a) The case where $N = 1$ has been given in Lemma 4(a). In the case where $N = 2$, we have

$$\begin{aligned}\mathbf{E}\left[\|x_{k+2} - x^*\|^2 \mid \mathcal{F}_k\right] &= \mathbf{E}\Big[\mathbf{E}\left[\|x_{k+2} - x^*\|^2 \mid \mathcal{F}_{k+1}\right] \;\Big|\; \mathcal{F}_k\Big]\\ &= \mathbf{E}\left[O\big(\|x_{k+1} - x^*\|^2 + \alpha_{k+1}^2\big) \mid \mathcal{F}_k\right]\\ &= O\big(\|x_k - x^*\|^2 + \alpha_k^2\big),\end{aligned}$$

where the first equality uses iterated expectation, and the second and third inequalities use Lemma 4(a) and the fact $\alpha_{k+1} \le \alpha_k$. In the case where $N > 2$, the result follows by applying the preceding argument inductively.

(b) The case where $N = 1$ has been given in Lemma 4(b). In the case where $N = 2$, we have

$$\begin{aligned}\mathbf{E}\left[\,\mathrm{d}^2(x_{k+2}) \mid \mathcal{F}_k\right] &= \mathbf{E}\Big[\mathbf{E}\left[\,\mathrm{d}^2(x_{k+2}) \mid \mathcal{F}_{k+1}\right] \;\Big|\; \mathcal{F}_k\Big]\\ &\le \mathbf{E}\Big[\mathcal{O}\left(\,\mathrm{d}^2(x_{k+1}) + \alpha_{k+1}^2\|x_{k+1} - x^*\|^2 + \alpha_{k+1}^2\right) \;\Big|\; \mathcal{F}_k\Big]\\ &\le O\Big(\,\mathrm{d}^2(x_k) + \alpha_k^2\|x_k - x^*\|^2 + \alpha_k^2\Big),\end{aligned}$$

where the first equality uses iterated expectation, the second inequality uses Lemma 4(b), and the third inequality uses Lemma 4(a), (b) and the fact $\alpha_{k+1} \le \alpha_k$. In the case where $N > 2$, the result follows by applying the preceding argument inductively.

(c) This follows by applying Lemma 4(c) and part (a):

$$\begin{aligned}\mathbf{E}\left[\big\|g(\bar{x}_{k+N}, v_{k+N})\big\|^2 \mid \mathcal{F}_k\right] &= \mathbf{E}\Big[\mathbf{E}\left[\|g(\bar{x}_{k+N}, v_{k+N})\|^2 \mid \mathcal{F}_{k+N}\right] \;\Big|\; \mathcal{F}_k\Big]\\ &\le \mathbf{E}\left[O\big(\|x_{k+N} - x^*\|^2 + 1\big) \mid \mathcal{F}_k\right]\\ &\le O\big(\|x_k - x^*\|^2 + 1\big).\end{aligned}$$

(d) For any $\ell \ge k$, we have

$$\begin{aligned}\mathbf{E}\left[\big\|x_{\ell+1} - x_\ell\big\|^2 \mid \mathcal{F}_k\right] &= \mathbf{E}\left[\mathbf{E}\left[\big\|x_{\ell+1} - x_\ell\big\|^2 \mid \mathcal{F}_\ell\right] \;\mid \mathcal{F}_k\right]\\ &\le \mathbf{E}\left[\mathcal{O}(\alpha_\ell^2)(\|x_\ell - x^*\|^2 + 1) + \mathcal{O}(\beta_\ell^2)\,\mathrm{d}^2(x_\ell) \mid \mathcal{F}_k\right]\\ &\le \mathcal{O}(\alpha_k^2)\big(\|x_k - x^*\|^2 + 1\big) + \mathcal{O}(\beta_k^2)\,\mathrm{d}^2(x_k),\end{aligned}$$

where the first inequality applies Lemma 4(d) and the second equality uses the fact $\alpha_{k+1} \le \alpha_k$, as well as parts (a), (b) of the current lemma. Then we have

$$\begin{aligned}\mathbf{E}\left[\big\|x_{k+N} - x_k\big\|^2 \mid \mathcal{F}_k\right] &\le N \sum_{\ell=k}^{k+N-1} \mathbf{E}\left[\big\|x_{\ell+1} - x_\ell\big\|^2 \mid \mathcal{F}_k\right]\\ &\le \mathcal{O}(N^2\alpha_k^2)\big(\|x_k - x^*\|^2 + 1\big) + \mathcal{O}(N^2\beta_k^2)\,\mathrm{d}^2(x_k)\end{aligned}$$

for all $k \ge 0$, with probability 1. ∎

LEMMA 6. *Let Assumptions* 1 *and* 2 *hold, let* x^* *be a given optimal solution of problem* (1), *let* $\{x_k\}$ *be generated by Algorithm* 1, *and let* N *be a given positive integer. Then for all* $k \ge 0$ *with probability* 1,

(a) $\mathbf{E}\left[f(x_k) - f(x_{k+N}) \mid \mathcal{F}_k\right] \le \mathcal{O}(\alpha_k)\big(\|x_k - x^*\|^2 + 1\big) + \mathcal{O}\left(\frac{\beta_k^2}{\alpha_k}\right)\mathrm{d}^2(x_k)$,

(b) $f(\Pi x_k) - f(x_k) \le \mathcal{O}\left(\frac{\alpha_k}{\beta_k}\right)\big(\|x_k - x^*\|^2 + 1\big) + \mathcal{O}\left(\frac{\beta_k}{\alpha_k}\right)\mathrm{d}^2(x_k)$,

(c) $f(\Pi x_k) - \mathbf{E}\left[f(x_{k+N}) \mid \mathcal{F}_k\right] \le \mathcal{O}\left(\frac{\alpha_k}{\beta_k}\right)\big(\|x_k - x^*\|^2 + 1\big) + \mathcal{O}\left(\frac{\beta_k}{\alpha_k}\right)\mathrm{d}^2(x_k)$.

Proof. (a) By using the definition of subgradients, we have

$$\begin{aligned}f(x_k) - f(x_{k+N}) &\le -\tilde{\nabla} f(x_k)'(x_{k+N} - x_k) \le \big\|\tilde{\nabla} f(x_k)\big\|\|x_{k+N} - x_k\|\\ &\le \frac{\alpha_k}{2}\|\tilde{\nabla} f(x_k)\|^2 + \frac{2}{\alpha_k}\|x_{k+N} - x_k\|^2.\end{aligned}$$

Taking expectation on both sides, using Assumption 1, and using Lemma 5(d), we obtain

$$\begin{aligned}\mathbf{E}\left[f(x_k) - f(x_{k+N}) \mid \mathcal{F}_k\right] &\le \frac{\alpha_k}{2}\|\tilde{\nabla} f(x_k)\|^2 + \frac{2}{\alpha_k}\mathbf{E}\left[\|x_{k+N} - x_k\|^2 \mid \mathcal{F}_k\right]\\ &\le \mathcal{O}(\alpha_k)\big(\|x_k - x^*\|^2 + 1\big) + \mathcal{O}\left(\frac{\beta_k^2}{\alpha_k}\right)\mathrm{d}^2(x_k).\end{aligned}$$

(b) Similar to part (a), we use the definition of subgradients to obtain

$$f(\Pi x_k) - f(x_k) \le -\tilde{\nabla} f(\Pi x_k)(x_k - \Pi x_k) \le \frac{\alpha_k}{2\beta_k}\left\|\tilde{\nabla} f(\Pi x_k)\right\|^2 + \frac{2\beta_k}{\alpha_k}\|x_k - \Pi x_k\|^2.$$

Also from Assumption 1, we have

$$\|\tilde{\nabla} f(\Pi x_k)\|^2 \leq L(\|\Pi x_k\|^2+1) \leq \mathcal{O}(\|\Pi x_k - x^*\|^2+1) \leq O\big(\|x_k - x^*\|^2+1\big),$$

while

$$\|x_k - \Pi x_k\| = \mathrm{d}(x_k).$$

We combine the preceding three relations and obtain (b).

(c) We sum the relations of (a) and (b) and obtain (c). □

3. The coupled convergence theorem.

In this section, we focus on the general Algorithm 1 that alternates between an iteration of random optimality update and an iteration of random feasibility update, i.e.,

$$z_k = x_k - \alpha_k g(\bar{x}_k, v_k), \qquad x_{k+1} = z_k - \beta_k\left(z_k - \Pi_{w_k} z_k\right) \quad \text{with } \bar{x}_k = x_k \text{ or } \bar{x}_k = x_{k+1}$$

(cf. (5), (9)), without specifying details regarding how the random variables w_k and v_k are generated. We show that as long as both iterations make sufficient improvement "on average," the generic algorithm consisting of their combination is convergent to an optimal solution. We also show that by using appropriate stepsizes, the optimality error and the feasibility error decrease to zero at a rate of $\mathcal{O}(1/\sqrt{k})$ and $\mathcal{O}(\log k/k)$, respectively. The first key result of the paper is stated as follows.

Theorem 1 (coupled convergence theorem). *Let Assumptions* 1 *and* 2 *hold, let* x^* *be a given optimal solution of problem* (1)*, and let* $\{x_k\}$ *be a sequence of random variables generated by Algorithm* 1*. Assume that there exist positive integers* M, N *such that*

(i) *with probability* 1 *for all* $k = 0, N, 2N, \ldots$,

$$\mathbf{E}\left[\|x_{k+N} - x^*\|^2 \mid \mathcal{F}_k\right] \leq \|x_k - x^*\|^2 - 2\left(\sum_{\ell=k}^{k+N-1} \alpha_\ell\right)\big(f(x_k) - f^*\big) + \mathcal{O}(\alpha_k^2)\big(\|x_k - x^*\|^2+1\big) + \mathcal{O}(\beta_k^2)\,\mathrm{d}^2(x_k);$$

(ii) *with probability* 1 *for all* $k \geq 0$,

$$\mathbf{E}\left[\,\mathrm{d}^2(x_{k+M}) \mid \mathcal{F}_k\right] \leq \big(1 - \Theta(\beta_k)\big)\,\mathrm{d}^2(x_k) + \mathcal{O}\left(\frac{\alpha_k^2}{\beta_k}\right)\left(\|x_k - x^*\|^2+1\right).$$

Then the sequence $\{x_k\}$ *converges almost surely to a random point in the set of optimal solutions of the convex optimization problem* (1)*.*

Before proving the theorem we provide some discussion. Let us first note that in the preceding proposition, x^* is an arbitrary but fixed optimal solution and that the $\mathcal{O}(\cdot)$ and $\Theta(\cdot)$ terms in the conditions (i) and (ii) may depend on x^*, as well as M and N. We refer to condition (i) as the *optimality improvement condition* and refer to condition (ii) as the *feasibility improvement condition.* According to the statement of Theorem 1, the recursions for optimality improvement and feasibility improvement are allowed to be coupled with each other, in the sense that either recursion involves iterates of the other one. This coupling is unavoidable due to the design of Algorithm 1, which by itself is a combination of two types of iterations. Despite being closely coupled, the two recursions are not necessarily coordinated with each other, in the sense that their cycles' lengths M and N may not be equal. This makes the proof more challenging.

We prove an important preliminary result for our purpose: the coupled supermartingale convergence lemma. It states that by combining the two improvement processes appropriately, a supermartingale convergence argument applies and both processes can be shown to be convergent. Moreover for the case where $M = 1$ and $N = 1$, the lemma yields "easily" the convergence proof of Theorem 1. To the best knowledge of the authors, there has been no coupled supermartingale convergence result in prior literature.

Lemma 7 (coupled supermartingale convergence lemma). *Let $\{\xi_t\}$, $\{\zeta_t\}$, $\{u_t\}$, $\{\bar{u}_t\}$, $\{\eta_t\}$, $\{\theta_t\}$, $\{\epsilon_t\}$, $\{\mu_t\}$, and $\{\nu_t\}$ be sequences of nonnegative random variables such that*

$$\mathbf{E}\left[\xi_{t+1} \mid \mathcal{G}_k\right] \le (1+\eta_t)\xi_t - u_t + c\theta_t\zeta_t + \mu_t,$$
$$\mathbf{E}\left[\zeta_{t+1} \mid \mathcal{G}_k\right] \le (1-\theta_t)\zeta_t - \bar{u}_t + \epsilon_t\xi_t + \nu_t,$$

where $\mathcal{G}_k$ denotes the collection $\xi_0,\ldots,\xi_t,\zeta_0,\ldots,\zeta_t,u_0,\ldots,u_t,\bar{u}_0,\ldots,\bar{u}_t,\eta_0,\ldots,\eta_t$, $\theta_0,\ldots,\theta_t,\epsilon_0,\ldots,\epsilon_t$, $\mu_0,\ldots,\mu_t$, $\nu_0,\ldots,\nu_t$, and c is a positive scalar. Also, assume that

$$\sum_{t=0}^{\infty}\eta_t < \infty, \qquad \sum_{t=0}^{\infty}\epsilon_t < \infty, \qquad \sum_{t=0}^{\infty}\mu_t < \infty, \qquad \sum_{t=0}^{\infty}\nu_t < \infty, \qquad w.p.1.$$

Then ξ_t and ζ_t converge almost surely to nonnegative random variables, and we have

$$\sum_{t=0}^{\infty}u_t < \infty, \qquad \sum_{t=0}^{\infty}\bar{u}_t < \infty, \qquad \sum_{t=0}^{\infty}\theta_t\zeta_t < \infty, \qquad w.p.1.$$

Moreover, if η_t, ϵ_t, μ_t, and ν_t are deterministic scalars, the sequences $\big\{\mathbf{E}\left[\xi_t\right]\big\}$ and $\big\{\mathbf{E}\left[\zeta_t\right]\big\}$ are bounded, and $\sum_{t=0}^{\infty}\mathbf{E}\left[\theta_t\zeta_t\right] < \infty$.

Proof. We define J_t to be the random variable

$$J_t = \xi_t + c\zeta_t.$$

By combining the given inequalities, we obtain

$$\begin{aligned}\mathbf{E}\left[J_{t+1} \mid \mathcal{G}_k\right] &= \mathbf{E}\left[\xi_{t+1} \mid \mathcal{G}_k\right] + c\cdot\mathbf{E}\left[\zeta_{t+1} \mid \mathcal{G}_k\right]\\ &\le (1+\eta_t+c\epsilon_t)\xi_t + c\zeta_t - (u_t + c\bar{u}_t) + (\mu_t + c\nu_t)\\ &\le (1+\eta_t+c\epsilon_t)(\xi_t + c\zeta_t) - (u_t + c\bar{u}_t) + (\mu_t + c\nu_t).\end{aligned}$$

It follows from the definition of J_t that

$$\text{(14)}\quad \mathbf{E}\left[J_{t+1} \mid \mathcal{G}_k\right] \le (1+\eta_t+c\epsilon_t)J_t - (u_t+c\bar{u}_t) + (\mu_t+c\nu_t) \le (1+\eta_t+c\epsilon_t)J_t + (\mu_t+c\nu_t).$$

Since $\sum_{t=0}^{\infty}\eta_t < \infty$, $\sum_{t=0}^{\infty}\epsilon_t < \infty$, $\sum_{t=0}^{\infty}\mu_t < \infty$, and $\sum_{t=0}^{\infty}\nu_t < \infty$ with probability 1, the supermartingale convergence lemma (Lemma 1) applies to (14). Therefore J_t converges almost surely to a nonnegative random variable, and

$$\sum_{t=0}^{\infty}u_t < \infty, \qquad \sum_{t=0}^{\infty}\bar{u}_t < \infty, \qquad w.p.1.$$

Since J_t converges almost surely, the sequence $\{J_t\}$ must be bounded with probability 1. Moreover, from the definition of J_t we have $\xi_t \le J_t$ and $\zeta_t \le \frac{1}{c}J_t$. Thus the sequences $\{\xi_t\}$ and $\{\zeta_t\}$ are also bounded with probability 1.

By using the relation $\sum_{t=0}^{\infty}\epsilon_t < \infty$ and the almost sure boundedness of $\{\xi_t\}$, we obtain

$$\sum_{t=0}^{\infty}\epsilon_t\xi_t \le \left(\sum_{t=0}^{\infty}\epsilon_t\right)\left(\sup_{t\ge 0}\xi_t\right) < \infty \qquad \text{w.p.1.} \tag{15}$$

From (15), we see that the supermartingale convergence lemma, Lemma 1, also applies to the given inequality

$$\mathbf{E}\left[\zeta_{t+1} \mid \mathcal{G}_k\right] \le (1-\theta_t)\zeta_t - \bar{u}_t + \epsilon_t\xi_t + \nu_t \le (1-\theta_t)\zeta_t + \epsilon_t\xi_t + \nu_t. \tag{16}$$

Therefore ζ_t converges almost surely to a random variable, and

$$\sum_{t=0}^{\infty}\theta_t\zeta_t < \infty \qquad \text{w.p.1.}$$

Since both $J_t = \xi_t + c\zeta_t$ and ζ_t are almost surely convergent, the random variable ξ_t must also converge almost surely to a random variable.

Finally, let us assume that η_t, ϵ_t, μ_t, and ν_t are deterministic scalars. We take expectation on both sides of (14) and obtain

$$\mathbf{E}\left[J_{t+1}\right] \le (1+\eta_t+c\epsilon_t)\mathbf{E}\left[J_t\right] + (\mu_t + c\nu_t). \tag{17}$$

Since the scalars $\eta_t, \epsilon_t, \mu_t$, and ν_t are summable, we obtain that the sequence $\{\mathbf{E}\left[J_t\right]\}$ is bounded (the supermartingale convergence lemma applies and shows that $\mathbf{E}\left[J_t\right]$ converges). This further implies that the sequences $\{\mathbf{E}\left[\xi_t\right]\}$ and $\{\mathbf{E}\left[\zeta_t\right]\}$ are bounded.

By taking expectation on both sides of (16), we obtain

$$\mathbf{E}\left[\zeta_{t+1}\right] \le \mathbf{E}\left[\zeta_t\right] - \mathbf{E}\left[\theta_t\zeta_t\right] + \left(\epsilon_t\mathbf{E}\left[\xi_t\right] + \nu_t\right).$$

By applying the preceding relation inductively and by taking the limit as $k\to\infty$, we have

$$0 \le \lim_{k\to\infty}\mathbf{E}\left[\zeta_{t+1}\right] \le \mathbf{E}\left[\zeta_0\right] - \sum_{t=0}^{\infty}\mathbf{E}\left[\theta_t\zeta_t\right] + \sum_{t=0}^{\infty}\left(\epsilon_t\mathbf{E}\left[\xi_t\right] + \nu_t\right).$$

Therefore

$$\sum_{t=0}^{\infty}\mathbf{E}\left[\theta_t\zeta_t\right] \le \mathbf{E}\left[\zeta_0\right] + \sum_{t=0}^{\infty}\left(\epsilon_t\mathbf{E}\left[\xi_t\right] + \nu_t\right) \le \mathbf{E}\left[\zeta_0\right] + \left(\sum_{t=0}^{\infty}\epsilon_t\right)\sup_{t\ge 0}\left(\mathbf{E}\left[\xi_t\right]\right) + \left(\sum_{t=0}^{\infty}\nu_t\right) < \infty,$$

where the last relation uses the boundedness of $\left\{\mathbf{E}\left[\xi_t\right]\right\}$. □

We are tempted to directly apply the coupled supermartingale convergence lemma, Lemma 7, to prove the results of Theorem 1. However, two issues remain to be addressed. First, the two improvement conditions of Theorem 1 are not fully coordinated with each other. In particular, their cycle lengths, M and N, may be different. Second, even if we let $M = 1$ and $N = 1$, we still cannot apply Lemma 7. The reason is that the optimality improvement condition (i) involves the subtraction of the term $(f(x_k) - f^*)$, which can be either nonnegative or negative. The following proof addresses these issues.

Proof of the coupled convergence theorem, Theorem 1. Our proof consists of four steps, and its main idea is to construct a meta-cycle of $M \times N$ iterations, where the tth cycle of iterations maps from x_{tMN} to $x_{(t+1)MN}$. The purpose is to ensure that both feasibility iterations and optimality iterations make reasonable progress within each meta-cycle, which will be shown in the first and second steps of the proof. The third step is to apply the preceding coupled supermartingale convergence lemma and show that the end points of the meta-cycles, $\{X_{tMN}\}$, form a subsequence that converges almost surely to an optimal solution. Finally, the fourth step is to argue that the maximum deviation of the iterates within a cycle decreases to 0 almost surely. From this we will show that the entire sequence $\{x_k\}$ converges almost surely to a random point in the set of optimal solutions.

Step 1 (derive the optimality improvement from x_{tMN} to $x_{(t+1)MN}$). We apply condition (i) repeatedly to obtain for any $t > 0$ that

$$\begin{aligned} &\mathbf{E}\left[\|x_{(t+1)MN} - x^*\|^2 \mid \mathcal{F}_{tMN}\right] \\ &\quad \le \|x_{tMN} - x^*\|^2 - 2 \sum_{\ell=tM}^{(t+1)M-1} \left(\sum_{k=\ell N}^{(\ell+1)N-1} \alpha_k \right) \Big(\mathbf{E}\left[f(x_{\ell N}) \mid \mathcal{F}_{tMN}\right] - f^* \Big) \\ &\qquad + \sum_{\ell=tM}^{(t+1)M-1} \mathcal{O}(\alpha_{\ell N}^2) \Big(\mathbf{E}\left[\|x_{\ell N} - x^*\|^2 \mid \mathcal{F}_{tMN}\right] + 1 \Big) \\ &\qquad + \sum_{\ell=tM}^{(t+1)M-1} \mathcal{O}(\beta_{\ell N}^2) \mathbf{E}\left[\mathrm{d}^2(x_{\ell N}) \mid \mathcal{F}_{tMN}\right] \qquad \text{w.p.1.} \end{aligned} \tag{18}$$

From Lemma 5(a) and the nonincreasing property of $\{\alpha_k\}$ we obtain the bound

$$\sum_{\ell=tM}^{(t+1)M-1} \mathcal{O}(\alpha_{\ell N}^2) \Big(\mathbf{E}\left[\|x_{\ell N} - x^*\|^2 \mid \mathcal{F}_{tMN}\right] + 1 \Big) \le \mathcal{O}(\alpha_{tMN}^2) \big(\|x_{tMN} - x^*\|^2 + 1 \big).$$

From Lemma 5(b) and the nonincreasing property of $\{\beta_k\}$ we obtain the bound

$$\begin{aligned} &\sum_{\ell=tM}^{(t+1)M-1} \mathcal{O}(\beta_{\ell N}^2) \mathbf{E}\left[\mathrm{d}^2(x_{\ell N}) \mid \mathcal{F}_{tMN}\right] \\ &\quad \le \mathcal{O}(\beta_{tMN}^2)\, \mathrm{d}^2(x_{tMN}) + \mathcal{O}(\alpha_{tMN}^2) \big(\|x_{tMN} - x^*\|^2 + 1 \big). \end{aligned}$$

By using Lemma 6(c) we further obtain

$$\begin{aligned} &- \Big(\mathbf{E}\left[f(x_{\ell N}) \mid \mathcal{F}_{tMN}\right] - f^* \Big) \\ &\le - \Big(f(\Pi x_{tMN}) - f^* \Big) + \Big(\mathbf{E}\left[f(\Pi x_{tMN}) - f(x_{\ell N}) \mid \mathcal{F}_{tMN}\right] \Big) \\ &\le - \Big(f(\Pi x_{tMN}) - f^* \Big) + \mathcal{O}\left(\frac{\alpha_{tMN}}{\beta_{tMN}} \right) \big(\|x_{tMN} - x^*\|^2 + 1 \big) \\ &\quad + \mathcal{O}\left(\frac{\beta_{tMN}}{\alpha_{tMN}} \right) \mathrm{d}^2(x_{tMN}). \end{aligned}$$

We apply the preceding bounds to (18) and remove redundant scalars in the big $\mathcal{O}(\cdot)$ terms, yielding

$$\begin{aligned} &\mathbf{E}\left[\|x_{(t+1)MN} - x^*\|^2 \mid \mathcal{F}_{tMN}\right] \\ &\le \|x_{tMN} - x^*\|^2 - 2\left(\sum_{k=tMN}^{(t+1)MN-1} \alpha_k\right)\left(f(\Pi x_{tMN}) - f^*\right) \\ &\quad + \mathcal{O}\left(\frac{\alpha_{tMN}^2}{\beta_{tMN}}\right)\left(\|x_{tMN} - x^*\|^2 + 1\right) + \mathcal{O}\left(\beta_{tMN}\right) \mathrm{d}^2(x_{tMN}) \end{aligned} \tag{19}$$

for all $t \ge 0$ with probability 1. Note that the term $f(\Pi x_k) - f^*$ is nonnegative. This will allow us to treat (19) as one of the conditions of Lemma 7.

Step 2 (derive the feasibility improvement from x_{tMN} to $x_{(t+1)MN}$). We apply condition (ii) repeatedly to obtain for any $t \ge 0$ that

$$\begin{aligned} \mathbf{E}\left[\mathrm{d}^2(x_{(t+1)MN}) \mid \mathcal{F}_{tMN}\right] &\le \left(\prod_{\ell=tN}^{(t+1)N-1}\left(1 - \Theta(\beta_{\ell M})\right)\right) \mathrm{d}^2(x_{tMN}) \\ &+ \sum_{\ell=tN}^{(t+1)N-1} \mathcal{O}\left(\frac{\alpha_{\ell M}^2}{\beta_{\ell M}}\right)\left(\mathbf{E}\left[\|x_{\ell M} - x^*\|^2 \mid \mathcal{F}_{tMN}\right] + 1\right) \end{aligned}$$

with probability 1. Then by using Lemma 5(a) to bound the terms $\mathbf{E}[\|x_{\ell M} - x^*\|^2 \mid \mathcal{F}_{tMN}]$, we obtain

$$\begin{aligned} &\mathbf{E}\left[\mathrm{d}^2(x_{(t+1)MN}) \mid \mathcal{F}_{tMN}\right] \\ &\le \left(1 - \Theta(\beta_{tMN})\right)\mathrm{d}^2(x_{tMN}) + \mathcal{O}\left(\sum_{k=tMN}^{(t+1)MN-1} \frac{\alpha_k^2}{\beta_k}\right)\left(\|x_{tMN} - x^*\|^2 + 1\right) \end{aligned} \tag{20}$$

with probability 1.

Step 3 (apply the coupled supermartingale convergence lemma). Let $\epsilon_t = \mathcal{O}(\sum_{k=tMN}^{(t+1)MN-1} \frac{\alpha_k^2}{\beta_k})$, so we have

$$\sum_{t=0}^{\infty} \epsilon_t = \sum_{k=0}^{\infty} \mathcal{O}\left(\frac{\alpha_k^2}{\beta_k}\right) < \infty.$$

Therefore the coupled supermartingale convergence lemma (cf. Lemma 7) applies to inequalities (19) and (20). It follows that $\|x_{tMN} - x^*\|^2$ and $\mathrm{d}^2(x_{tMN})$ converge almost surely,

$$\sum_{t=0}^{\infty} \Theta(\beta_{tMN})\,\mathrm{d}^2(x_{tMN}) < \infty \qquad \text{w.p.1} \tag{21}$$

and

$$\sum_{t=0}^{\infty}\left(\sum_{k=tMN}^{(t+1)MN-1}\alpha_k\right)\big(f(\Pi x_{tMN})-f^*\big)<\infty \qquad \text{w.p.1.} \tag{22}$$

Moreover, from the last part of Lemma 7, it follows that the sequence $\{\mathbf{E}\left[\|x_{tMN}-x^*\|^2\right]\}$ is bounded, and we have

$$\sum_{t=0}^{\infty}\Theta(\beta_{tMN})\mathbf{E}\left[\,\mathrm{d}^2(x_{tMN})\right]<\infty. \tag{23}$$

Since β_k is nonincreasing, we have

$$\sum_{t=0}^{\infty}\Theta(\beta_{tMN})\geq\sum_{t=0}^{\infty}\frac{1}{MN}\left(\sum_{k=tMN}^{(t+1)MN-1}\Theta(\beta_k)\right)=\frac{1}{MN}\sum_{k=0}^{\infty}\beta_k=\infty.$$

This together with the almost sure convergence of $\mathrm{d}^2(x_{tMN})$ and relation (21) implies that

$$\mathrm{d}^2(x_{tMN})\xrightarrow{a.s.}0 \qquad \text{as} \quad t\to\infty,$$

(if $\mathrm{d}^2(x_{tMN})$ converges to a positive scalar, then $\Theta(\beta_{tMN})\,\mathrm{d}^2(x_{tMN})$ would no longer be summable). Following a similar analysis, the relation (22) together with the assumption $\sum_{k=0}^{\infty}\alpha_k=\infty$ implies that

$$\liminf_{t\to\infty}f(\Pi x_{tMN})=f^* \qquad \text{w.p.1.}$$

Now let us consider an arbitrary sample trajectory of the stochastic process $\{(w_k,v_k)\}$, such that the associated sequence $\{\|x_{tMN}-x^*\|\}$ is convergent and is thus bounded, $\mathrm{d}^2(x_{tMN})\to 0$, and $\liminf_{t\to\infty}f(\Pi x_{tMN})=f^*$. These relations together with the continuity of f further imply that the sequence $\{x_{tMN}\}$ must have a limit point $\bar{x}\in X^*$. Also, since $\|x_{tMN}-x^*\|^2$ is convergent for arbitrary $x^*\in X^*$, the sequence $\|x_{tMN}-\bar{x}\|^2$ is convergent and has a limit point 0. If follows that $\|x_{tMN}-\bar{x}\|^2\to 0$, so that $x_{tMN}\to\bar{x}$. Note that the set of all such sample trajectories has a probability measure equal to 1. Therefore the sequence of random variables $\{x_{tMN}\}$ is convergent almost surely to a random point in X^* as $t\to\infty$.

Step 4 (prove that the entire sequence $\{x_k\}$ converges). Let $\epsilon>0$ be arbitrary. By using the Markov inequality, Lemma 5(c), and the boundedness of $\{\mathbf{E}\left[\|x_{tMN}-x^*\|^2\right]\}$ (as shown in Step 3), we obtain

$$\begin{aligned}\sum_{k=0}^{\infty}\mathbf{P}\big(\alpha_k\|g(\bar{x}_k,v_k)\|\geq\epsilon\big)&\leq\sum_{k=0}^{\infty}\frac{\alpha_k^2\mathbf{E}\left[\|g(\bar{x}_k,v_k)\|^2\right]}{\epsilon^2}\\&<\sum_{t=0}^{\infty}\frac{\alpha_{tMN}^2\mathbf{E}\left[\mathcal{O}(\|x_{tMN}-x^*\|^2+1)\right]}{\epsilon^2}<\infty.\end{aligned}$$

Similarly, by using the Markov inequality, Lemma 5(b), and (23), we obtain

$$\begin{aligned}\sum_{k=0}^{\infty}\mathbf{P}\big(\beta_k\,\mathrm{d}(x_k)\geq\epsilon\big)&\leq\sum_{k=0}^{\infty}\frac{\beta_k^2\mathbf{E}\left[\,\mathrm{d}^2(x_k)\right]}{\epsilon^2}\\&\leq\sum_{t=0}^{\infty}\frac{\beta_{tMN}^2\mathbf{E}\left[\mathcal{O}\left(\,\mathrm{d}^2(x_{tMN})+\alpha_{tMN}^2(\|x_{tMN}-x^*\|^2+1)\right)\right]}{\epsilon^2}<\infty.\end{aligned}$$

Applying the Borel–Cantelli lemma to the preceding two inequalities and taking ϵ arbitrarily small, we obtain

$$\alpha_k \|g(\bar{x}_k, v_k)\| \xrightarrow{a.s.} 0, \qquad \beta_k \,\mathrm{d}(x_k) \xrightarrow{a.s.} 0 \qquad \text{as} \;\; k \to \infty.$$

For any integer $t \geq 0$ we have

$$\begin{aligned} &\max_{tMN \leq k \leq (t+1)MN-1} \|x_k - x_{tMN}\| \\ &\leq \sum_{\ell = tMN}^{(t+1)MN-1} \|x_\ell - x_{\ell+1}\| \quad \text{(from the triangle inequality)} \\ &\leq \sum_{\ell = tMN}^{(t+1)MN-1} O\big(\alpha_\ell \|g(\bar{x}_\ell, v_\ell)\| + \beta_\ell \,\mathrm{d}(x_\ell)\big) \quad \text{(from (13))} \\ &\xrightarrow{a.s.} 0. \end{aligned}$$

Therefore the maximum deviation within a cycle of length MN decreases to 0 almost surely. To conclude, we have shown that x_k converges almost surely to a random point in X^* as $k \to \infty$. □

Rate of convergence is an important issue related to the proposed two-timescale stochastic algorithms. As noted earlier, the convergence of these algorithms involves two improvement processes with two different corresponding stepsizes. This coupling greatly complicates the convergence rate analysis. In the special case of minimizing a strongly convex and differentiable function, the proposed algorithm is a special case of an algorithm for strongly monotone variational inequalities given in [WB12]. For this algorithm, convergence rates and finite-sample error bounds have been derived in [WB12]. For minimization of general convex functions, convergence rate analysis is not available in the existing literature, except in special cases which involve no constraint sampling and/or more restrictive assumptions (see [NB00], [NB01], [Ber11], [Ned11], [WB12]).

In what follows, we provide a unified convergence rate analysis for Algorithm 1. Due to the coupling nature, the analysis differs substantially from conventional analysis of the stochastic gradient method (which either focuses on unconstrained problems or uses exact constraint projection every iteration). We emphasize that there are two types of errors: feasibility error and optimality error. They need to be analyzed separatedly. As predicted by the coupling convergence proof of Theorem 1, the two errors decrease to zero at different rates. This is verified in the next theorem.

THEOREM 2 (rate of coupled convergence). *Let Assumption* 1 *and conditions* (i), (ii) *of Theorem* 1 *hold, and let* $\alpha_k = \Theta(1/\sqrt{k})$, $\beta_k = \Theta(1)$. *If in addition the iterates* $\{x_k\}$ *are bounded in a sufficiently large ball, we have*

$$\mathbf{E}\left[f\left(\frac{1}{k}\sum_{t=1}^{k} \Pi_X x_t\right)\right] = f^* + \mathcal{O}\left(\frac{1}{\sqrt{k}}\right), \qquad \mathbf{E}\left[\mathrm{d}^2\left(\frac{1}{k}\sum_{t=1}^{k} x_t\right)\right] = \mathcal{O}\left(\frac{\log(k+1)}{k}\right).$$

Proof. Let $\alpha_k = 1/\sqrt{k}$ and $\beta_k = 1$ for simplicity. Our analysis follows from the analysis of Theorem 1. We take a weighted sum of (19) and (20). For some $c_1, c_2 > 0$, we obtain

$$
\begin{aligned}
&\mathbf{E}\left[\|x_{(t+1)MN}-x^*\|^2+c_1\,\mathrm{d}^2(x_{(t+1)MN})\right]\\
&\quad\le \mathbf{E}\left[\|x_{tMN}-x^*\|^2+c_1\,\mathrm{d}^2(x_{tMN})\right]+\mathcal{O}\left(\sum_{k=tMN}^{(t+1)MN-1}\frac{\alpha_k^2}{\beta_k}\right)\left(\mathbf{E}\left[\|x_{tMN}-x^*\|^2\right]+1\right)\\
&\qquad-2\left(\sum_{k=tMN}^{(t+1)MN-1}\alpha_k\right)\mathbf{E}\left[f(\Pi x_{tMN})-f^*\right]-c_2\beta_{tMN}\mathbf{E}\left[\,\mathrm{d}^2(x_{tMN})\right].
\end{aligned}\tag{24}
$$

By using the stepsize assumptions, we have $\sum_{k=tMN}^{(t+1)MN-1}\alpha_k=\Theta(1/\sqrt{k})$, $\sum_{k=tMN}^{(t+1)MN-1}\frac{\alpha_k^2}{\beta_k}=\Theta(1/k)$, and $\beta_{tMN}=1$. By using the bounded iterates assumption, there exists $R>0$ such that $\mathbf{E}\left[\|x_{tMN}-x^*\|^2\right]\le R$ and $\mathbf{E}\left[\,\mathrm{d}^2(x_{tMN})\right]\le \mathbf{E}\left[\|x_{tMN}-x^*\|^2\right]\le R$ for all t. We let

$$J_t=\mathbf{E}\left[\|x_{tMN}-x^*\|^2+c_1\,\mathrm{d}^2(x_{tMN})\right],$$

so we have $J_t\le D\equiv(1+c_1)R$ for all t. Then it follows from (24) that

$$J_{t+1}\le J_t+\mathcal{O}\left(\frac{1}{t}\right)\cdot(D+1)-\Theta\left(\frac{1}{\sqrt{t}}\right)\mathbf{E}\left[f(\Pi x_{tMN})-f^*\right]-c_2\mathbf{E}\left[\,\mathrm{d}^2(x_{tMN})\right].\tag{25}$$

First we analyze the feasibility error. By rearranging the terms in (25) and taking the sum over $t=1,\ldots,T$, we have

$$
\begin{aligned}
c_2\sum_{t=1}^{T}\mathbf{E}\left[\,\mathrm{d}^2(x_{tMN})\right]&\le\sum_{t=1}^{T}\left(J_t-J_{t+1}+\mathcal{O}\left(\frac{1}{t}\right)(D+1)\right)\\
&\le J_1-J_{T+1}+\sum_{t=1}^{T}\mathcal{O}\left(\frac{1}{t}\right)(D+1)\\
&\le D+\mathcal{O}\left(\sum_{t=1}^{T}\frac{1}{t}\right)(D+1)\\
&=\mathcal{O}\left(\log(T+1)\right)\cdot(D+1).
\end{aligned}
$$

By using Lemma 5(b), we have for some $c_3>0$ that

$$
\begin{aligned}
\sum_{t=1}^{k}\mathbf{E}\left[\,\mathrm{d}^2\left(x_t\right)\right]&\le c_3\sum_{t=1}^{k}\left(\mathbf{E}\left[\,\mathrm{d}^2\left(x_{\lfloor t/MN\rfloor MN}\right)\right]+\alpha_{\lfloor t/MN\rfloor MN}^2\right)\\
&\le\mathcal{O}(\log(\lfloor k/MN\rfloor+1))=\mathcal{O}(\log(k+1)).
\end{aligned}
$$

By using the convexity of the squared distance function, we have for all $k=1,2,\ldots$ that $\mathbf{E}\left[\,\mathrm{d}^2\left(\frac{1}{k}\sum_{t=1}^{k}x_t\right)\right]\le\frac{1}{k}\sum_{t=1}^{k}\mathbf{E}\left[\,\mathrm{d}^2\left(x_t\right)\right]=\mathcal{O}\left(\frac{\log(k+1)}{k}\right)$.

Next we analyze the optimality error. By using a similar analysis, we can show from (25) that for some $c_4, c_5 > 0$,

$$\begin{aligned}
c_4 \sum_{t=1}^{T} \mathbf{E}\left[f(\Pi x_{tMN}) - f^*\right] + c_5 \sum_{t=1}^{T} \sqrt{t}\mathbf{E}\left[\,\mathrm{d}^2(x_{tMN})\right] \\
&\le \sum_{t=1}^{T} \sqrt{t}\left(J_t - J_{t+1} + \mathcal{O}\left(\frac{1}{t}\right)(D+1)\right) \\
&\le J_1 + \sum_{t=2}^{T} (\sqrt{t+1} - \sqrt{t}) J_t + \sum_{t=0}^{T} \mathcal{O}\left(\frac{1}{\sqrt{t}}\right)(D+1) \\
&\le D + \sum_{t=2}^{T} (\sqrt{t+1} - \sqrt{t}) D + \mathcal{O}\left(\sum_{t=0}^{T} \frac{1}{\sqrt{t}}\right) \cdot (D+1) \\
&\le \sqrt{T} D + \mathcal{O}\left(\sum_{t=2}^{T} \frac{1}{\sqrt{t}}\right) \cdot (D+1) \\
&= \mathcal{O}\left(\sqrt{T}\right) \cdot (D+1).
\end{aligned}$$

By the convexity of f and basic norm inequalities, we have for all $tMN < k \le (t+1)MN$ and all t that

$$\begin{aligned}
&\mathbf{E}\left[f(\Pi x_k) - f^*\right] - \mathbf{E}\left[f(\Pi x_{tMN}) - f^*\right] \le \mathbf{E}\left[\tilde{\nabla} f(\Pi x_{tMN})'(\Pi x_k - \Pi x_{tMN})\right] \\
&\le \sigma \alpha_k \mathbf{E}\left[\|\tilde{\nabla} f(\Pi x_{tMN})\|^2\right] + \frac{1}{\alpha_k \sigma} \mathbf{E}\left[\|\Pi x_k - \Pi x_{tMN}\|^2\right] \\
&\le \mathcal{O}\left(\sigma \alpha_k\right) L^2 \mathbf{E}\left[\|x_{tMN}\|^2 + 1\right] + \mathcal{O}\left(\frac{\alpha_{tMN}^2}{\alpha_k \sigma}\right) \\
&\qquad \mathbf{E}\left[\|\Pi x_{tMN} - x^*\|^2 + 1\right] + \mathcal{O}\left(\frac{\beta_{tMN}^2}{\alpha_k \sigma}\right) \mathbf{E}\left[\,\mathrm{d}^2(x_{tMN})\right] \\
&\le \mathcal{O}\left(\frac{\sigma + 1/\sigma}{\sqrt{k}} + \frac{\sqrt{k}}{\sigma} \mathbf{E}\left[\,\mathrm{d}^2(x_{tMN})\right]\right) \\
&\le \mathcal{O}\left(\frac{1}{\sqrt{k}}\right) + \frac{c_5/c_4\sqrt{k}}{MN} \mathbf{E}\left[\,\mathrm{d}^2(x_{tMN})\right],
\end{aligned}$$

where the second inequality holds for any $\sigma > 0$, the third inequality uses Assumption 1(a) and Lemma 5(d), the fourth inequality uses the boundedness of iterates, and the last inequality holds for σ sufficiently large. It follows from the preceding inequalities that

$$\begin{aligned}
&\sum_{t=1}^{k} \mathbf{E}\left[f\left(\Pi_X x_t\right)\right] \\
&\quad \le \sum_{t=1}^{k} \left(\mathbf{E}\left[f\left(\Pi_X x_{\lfloor t/MN \rfloor MN}\right)\right] + \frac{c_5/c_4}{\sqrt{t}} \mathbf{E}\left[\,\mathrm{d}^2(x_{\lfloor t/MN \rfloor MN})\right] + \mathcal{O}(1/\sqrt{t})\right) \\
&\quad = k \cdot f^* + \mathcal{O}\left(\sqrt{\lfloor k/MN \rfloor} + \sqrt{k}\right) \\
&\quad = k \cdot f^* + \mathcal{O}(\sqrt{k}).
\end{aligned}$$

Finally, by the convexity of f we have $\mathbf{E}\left[f(\frac{1}{k}\sum_{t=1}^{k}\Pi_X x_t)\right] \leq \frac{1}{k}\sum_{t=1}^{k}\mathbf{E}\left[f\left(\Pi_X x_t\right)\right] \leq f^* + \mathcal{O}(1/\sqrt{k})$. ◻

In Theorem 2, we have picked a particular choice of stepsizes, which are known to be optimal for stochastic gradient methods (up to a constant, suggested in [SZ12]). For simplicity of analysis, we have also assumed that all iterates are bounded by a large ball. This can be verified, for example, if all constraint supersets X_i are bounded by a large ball. If the assumption cannot be verified, one may add an auxiliary constraint (e.g., a very large ball that contains at least one optimal solution) and to project onto it every iteration to keep the iterates bounded. We also note that the big $\mathcal{O}$ in the convergence rates involves some constants. These constants relate to the sampling schemes, variance of samples, properties of the objective function, properties of the constraints, choices of the stepsizes, etc. Computing the constants exactly for all possible cases is beyond the scope of the current paper. When analyzing a particular algorithm, one may follow a similar line of analysis to obtain customized convergence rates.

Theorem 2 says that the optimality error and feasibility error decrease on the order of $\mathcal{O}(1/\sqrt{k})$ and $\mathcal{O}(\log k/k)$, respectively. Indeed, the feasibility error decreases at a faster rate than the optimality error, so that the coupling convergence occurs. Let us compare the convergence rate given in Theorem 2 with some known convergence rate results:

- For finding a feasible point in the intersection of many sets, the random projection method converges at a linear rate (under a linear regularity condition regarding the sets). In contrast, the optimality error of Algorithm 1 decreases at a rate of $\mathcal{O}(\log k/k)$. The reason is that the coupling with the optimality improvement process slows down the convergence of the feasibility error.
- For minimization of convex functions, stochastic first-order methods (without random projection) have a convergence rate $\mathcal{O}(1/\sqrt{k})$ which is optimal with respect to the sample size k (see [ABRW12] for an information-theoretical lower bound). Our Algorithm 1, which uses both noisy first-order information and random projection, achieves a convergence rate on the same order $\mathcal{O}(1/\sqrt{k})$. This implies that our convergence rate is nonimprovable with respect to k, as long as noisy first-order information is used.

Surprisingly, using random projection in place of exact projection does not deteriorate the optimization convergence rate (up to a constant). This is critical evidence that supports the usage of random projection.

In this section, we have presented and proved the coupled convergence Theorems 1 and 2 that establish the convergence and the $\mathcal{O}(1/\sqrt{k})$ convergence rate of the general Algorithm 1. As a by-product of the analysis, we have proved an enhanced version of the supermartingle convergence lemma, namely, the coupled supermartingale convergence Lemma 7. It can be used to analyze two coupled sequences of random variables that satisfy two improvement inequalities, respectively, at different time scales.

Theorems 1 and 2 do not specify how the random samples of first-order information or constraints are generated. Instead, they establish the convergence and rate of convergence of the general algorithm assuming two conditions: a feasibility improvement condition and an optimality improvement condition. More precisely, they require that the feasibility update part of the algorithm makes sufficient progress toward feasibility within a number of steps and that the optimality update part makes sufficient progress toward optimality within a number of steps. On one hand, the two improvement processes are coupled together because the algorithm alternates between the two types of updates. On the other hand, the two improvement processes do not need to synchronize or coordinate with each other, as long as either one meets the sufficient improvement condition.

In what follows, we will consider a number of typical assumptions on the $\mathcal{SO}$ that generate stochastic first-order information and random constraint projections. We will show that all of them satisfy the optimality/feasibility improvement condition required in Theorems 1 and 2. As a result, we may apply Theorems 1 and 2 in conjunction with these sufficient improvement conditions and establish convergence results for a broad variety of stochastic algorithms.

4. Sampling schemes for constraints.

In this section, we focus on sampling schemes used in the $\mathcal{SO}$ for the constraints X_{w_k} that satisfy the feasibility improvement condition required by the coupled convergence theorem, i.e.,

$$\mathbf{E}\left[\, \mathrm{d}^2(x_{k+M}) \mid \mathcal{F}_k\right] \le \big(1-\Theta(\beta_k)\big)\, \mathrm{d}^2(x_k)+\mathcal{O}\left(\frac{\alpha_k^2}{\beta_k}\right)\big(\|x_k-x^*\|^2+1\big) \quad \forall\, k\ge 0 \quad \text{w.p.1},$$

where M is a positive integer. To satisfy the preceding condition, it is necessary that the distance between x_k and X asymptotically decreases as a contraction in a stochastic sense. We will consider several assumptions regarding the incremental projection process $\{\Pi_{w_k}\}$, including nearly independent sampling, most distant sampling, cyclic order sampling, Markov Chain sampling, etc.

Throughout our analysis in this section, we will require that the collection $\{X_i\}_{i=1}^m$ possesses a *linear regularity property*. This property was originally introduced by Bauschke [Bau96] in a more general Hilbert space setting; see also Bauschke and Borwein [BB96, Definition 5.6, p. 40].

Assumption 3 (linear regularity). *There exists a positive scalar η such that for any $x \in \Re^n$*

$$\|x-\Pi x\|^2 \le \eta \max_{\{i=1,\ldots,m\}} \|x-\Pi_{X_i}x\|^2.$$

Recently, the linear regularity property has been studied by Deutsch and Hundal [DH08] in order to establish linear convergence of a cyclic projection method for finding a common point of finitely many convex sets. Intuitively speaking, the linear regularity condition is related to angles between the sets X_i. It requires that the sets X_i behave like linear sets where they intersect with one another. This property is automatically satisfied when X is a polyhedral set. The discussions in [Bau96] and [DH08] identify several other situations where the linear regularity condition holds. As indicated by these references, the linear regularity condition is a mild restriction (in practice, it is rare to find examples that do not satisfy this condition.)

4.1. Nearly independent sample constraints.

We start with the easy case where the sample constraints are generated "nearly independently." In this case, it is necessary that each constraint is always sampled with sufficient probability, regardless of the sample history. This is formulated as the following assumption.

Assumption 4. *The random variables w_k, $k=0,1,\ldots$, are such that*

$$\inf_{k\ge 0} \mathbf{P}(X_{w_k}=X_i \mid \mathcal{F}_k) \ge \frac{\rho}{m}, \qquad i=1,\ldots,m,$$

with probability 1, *where $\rho\in(0,1]$ is some scalar.*

Under Assumptions 3 and 4, we claim that the expression

$$\mathbf{E}\big[\|x-\Pi_{w_k}x\|^2 \mid \mathcal{F}_k\big],$$

which may be viewed as the "average progress" of random projection at the kth iteration, is bounded from below by a multiple of the distance between x and X. Indeed, by Assumption 4, we have for any $j=1,\ldots,m$,

$$\mathbf{E}\big[\|x-\Pi_{w_k}x\|^2 \mid \mathcal{F}_k\big] = \sum_{i=1}^{m} \mathbf{P}\left(w_k = i \mid \mathcal{F}_k\right)\|x-\Pi_i x\|^2 \geq \frac{\rho}{m}\|x-\Pi_j x\|^2.$$

By maximizing the right-hand side of this relation over j and by using Assumption 3, we obtain

$$\text{(26)} \quad \mathbf{E}\big[\|x-\Pi_{w_k}x\|^2 \mid \mathcal{F}_k\big] \geq \frac{\rho}{m}\max_{1\leq j\leq m}\|x-\Pi_j x\|^2 \geq \frac{\rho}{m\eta}\|x-\Pi x\|^2 = \frac{\rho}{m\eta}\,\mathrm{d}^2(x)$$

for all $x \in \Re^n$ and $k \geq 0$, with probability 1. This indicates that the average feasibility progress of the nearly independent constraint sampling method is comparable to the feasibility error, i.e., the distance from x_k to X.

Now we are ready to show that the nearly independent constraint sampling scheme satisfies the feasibility improvement condition of the coupled convergence theorem (Theorem 1).

Proposition 1. *Let Assumptions* 1, 2, 3*, and* 4 *hold, and let x^* be a given optimal solution of problem* (1)*. Then Algorithm* 1 *generates a sequence $\{x_k\}$ such that*

$$\mathbf{E}\left[\,\mathrm{d}^2(x_{k+1}) \mid \mathcal{F}_k\right] \leq \left(1-\frac{\rho}{m\eta}\Theta(\beta_k)\right)\mathrm{d}^2(x_k) + \mathcal{O}\left(\frac{m\alpha_k^2}{\beta_k}\right)\left(\|x_k-x^*\|^2+1\right)$$

for all $k \geq 0$ with probability 1.

Proof. Let ϵ be a positive scalar. By applying Lemma 3 with $y=\Pi x_k$, we have

$$\begin{aligned}\mathrm{d}^2(x_{k+1}) \leq \|x_{k+1}-\Pi x_k\|^2 \leq (1+\epsilon)\|x_k-\Pi x_k\|^2 \\ + (1+1/\epsilon)\alpha_k^2\|g(\bar{x}_k,v_k)\|^2 - \beta_k(2-\beta_k)\|z_k-\Pi_{w_k}z_k\|^2.\end{aligned}$$

By using the bound

$$\|x_k-\Pi_{w_k}x_k\|^2 \leq 2\|z_k-\Pi_{w_k}z_k\|^2 + 8\|x_k-z_k\|^2 = 2\|z_k-\Pi_{w_k}z_k\|^2 + 8\alpha_k^2\|g(\bar{x}_k,v_k)\|^2,$$

which is obtained from Lemma 2(b), we further obtain

$$\begin{aligned}\mathrm{d}^2(x_{k+1}) &\leq (1+\epsilon)\|x_k-\Pi x_k\|^2 + \big(1+1/\epsilon+4\beta_k(2-\beta_k)\big)\alpha_k^2\|g(\bar{x}_k,v_k)\|^2 \\ &\quad - \frac{\beta_k(2-\beta_k)}{2}\|x_k-\Pi_{w_k}x_k\|^2 \\ &\leq (1+\epsilon)\,\mathrm{d}^2(x_k) + (5+1/\epsilon)\alpha_k^2\|g(\bar{x}_k,v_k)\|^2 - \Theta(\beta_k)\|x_k-\Pi_{w_k}x_k\|^2,\end{aligned}$$

where the second relation uses the facts $\|x_k-\Pi x_k\|^2 = \mathrm{d}^2(x_k)$ and $\Theta(\beta_k) \leq \beta_k(2-\beta_k) \leq 1$. Taking conditional expectation of both sides and applying Lemma 4(c) and (26), we obtain

$$\begin{aligned}\mathbf{E}\left[\,\mathrm{d}^2(x_{k+1}) \mid \mathcal{F}_k\right] &\leq (1+\epsilon)\,\mathrm{d}^2(x_k) + \mathcal{O}(1+1/\epsilon)\alpha_k^2\big(\|x_k-x^*\|^2+1\big) \\ &\quad - \frac{\rho}{m\eta}\Theta(\beta_k)\,\mathrm{d}^2(x_k) \\ &\leq \left(1-\frac{\rho}{m\eta}\Theta(\beta_k)\right)\mathrm{d}^2(x_k) + \mathcal{O}(m\alpha_k^2/\beta_k)\big(\|x_k-x^*\|^2+1\big),\end{aligned}$$

where the second relation is obtained by letting $\epsilon \ll \Theta(\beta_k)$. ∎

We remark that the condition in Assumption 4 can be weakened to

$$\inf_{k\geq 0} \mathbf{P}(X_{w_k} = X_i \mid \mathcal{F}_k) \geq \frac{\rho_k}{m}, \qquad i = 1, \ldots, m,$$

where ρ_k is driven to 0 at a suitable rate. Then we would obtain results analogous to the current one, where the modulus of contraction for the feasibility improvement becomes $1-\Theta(\frac{\rho_k \beta_k}{m\eta})$. To achieve the overall convergence, we would need an additional stepsize assumption that

$$\sum_{k=0}^{\infty} \rho_k \beta_k = \infty.$$

In this way, we can adapt the entire analysis of Theorem 1 to apply to the case where the constraint sampling distribution slowly varies.

4.2. Most distant sample constraint. Next we consider the case where we select the constraint superset that is the most distant from the current iterate. This yields an adaptive algorithm that selects the projection based on the iterates' history.

ASSUMPTION 5. *The random variable w_k is the index of the most distant constraint superset, i.e.,*

$$w_k = \mathrm{argmax}_{i=1,\ldots,m} \|x_k - \Pi_{X_i} x_k\|, \qquad k = 0, 1, \ldots.$$

By using Assumption 5 together with Assumption 3, we see that

$$\mathbf{E}\big[\|x_k - \Pi_{w_k} x_k\|^2 \mid \mathcal{F}_k\big] = \max_{i=1,\ldots,m} \|x_k - \Pi_i x_k\| \geq \frac{1}{\eta}\,\mathrm{d}^2(x_k) \qquad \forall\ k \geq 0 \quad \text{w.p.1.} \tag{27}$$

Then by using an analysis similar to that of Proposition 1, we obtain the following result.

PROPOSITION 2. *Let Assumptions* 1, 2, 3, *and* 5 *hold, and let x^* be a given optimal solution of problem* (1). *Then Algorithm* 1 *generates a sequence $\{x_k\}$ such that*

$$\mathbf{E}\left[\mathrm{d}^2(x_{k+1}) \mid \mathcal{F}_k\right] \leq \left(1 - \Theta\left(\frac{\beta_k}{\eta}\right)\right) \mathrm{d}^2(x_k) + \mathcal{O}\left(\frac{\alpha_k^2}{\beta_k}\right)\left(\|x_k - x^*\|^2 + 1\right)$$

for all $k \geq 0$, with probability 1.

Proof. The proof is almost identical to that of Proposition 1, except that we use (27) in place of (26). □

4.3. Sample constraints according to a cyclic order. Now let us consider the case where the constraint supersets $\{X_{w_k}\}$ are sampled in a cyclic manner, either by random shuffling or according to a deterministic cyclic order.

ASSUMPTION 6. *With probability* 1, *for all $t \geq 0$, the sequence of constraint sets of the tth cycle, i.e.,*

$$\{X_{w_k}\}, \qquad \textit{where } k = k, k+1, \ldots, k+M-1,$$

is a permutation of $\{X_1, \ldots, X_m\}$.

Under Assumption 6, it is no longer true that the distance from x_k to the feasible set is "stochastically decreased" at every iteration. However, all the constraint sets will be visited at least once within a cycle of m iterations. This suggests that the distance to the feasible set is improved on average every m iterations. We first prove a lemma regarding the progress toward feasibility over a number of iterations.

LEMMA 8. *Let Assumptions* 1, 2, *and* 3 *hold, and let* $\{x_k\}$ *be generated by Algorithm* 1. *Assume that for given integers* $k > 0$ *and* $M > 0$, *any particular index in* $\{1, \ldots, m\}$ *will be visited at least once by the random variables* $\{w_k, \ldots, w_{k+M-1}\}$. *Then*

$$\frac{1}{2M\eta}\, \mathrm{d}^2(x_k) \le 4 \sum_{\ell=k}^{k+M-1} \|z_\ell - \Pi_{w_\ell} z_\ell\|^2 + \sum_{\ell=k}^{k+M-1} \alpha_\ell^2 \big\|g(\bar{x}_\ell, v_\ell)\big\|^2.$$

Proof. Let $k^* \in \{k, \ldots, k+M-1\}$ be the index that attains the maximum in the linear regularity assumption for x_k (cf. Assumption 3), so that

$$\mathrm{d}^2(x_k) \le \eta \max_{i=1,\ldots,m} \|x_k - \Pi_{X_i} x_k\|^2 = \eta \|x_k - \Pi_{w_{k^*}} x_k\|^2.$$

Such k^* always exists, because it is assumed that any particular index will be visited by the sequence $\{w_k, \ldots, w_{k+M-1}\}$. We have

$$\begin{aligned}
\frac{1}{\sqrt{\eta}}\, \mathrm{d}(x_k) &\le \|x_k - \Pi_{w_{k^*}} x_k\| \\
&\le \|x_k - \Pi_{w_{k^*}} z_{k^*}\| \\
&\quad \text{(by the definition of } \Pi_{w_{k^*}} x_k \text{ and the fact } \Pi_{w_{k^*}} z_{k^*} \in X_{w_{k^*}}) \\
&= \left\| x_k - \frac{1}{\beta_{k^*}} x_{k^*+1} + \frac{1-\beta_{k^*}}{\beta_{k^*}} z_{k^*} \right\| \\
&\quad \text{(by } x_{k^*+1} = z_{k^*} - \beta_{k^*}(z_{k^*} - \Pi_{w_{k^*}} z_{k^*});\ \text{cf. (3))} \\
&= \left\| \sum_{\ell=k}^{k^*-1} \frac{\beta_\ell - 1}{\beta_\ell} (z_\ell - x_{\ell+1}) + \sum_{\ell=k}^{k^*} \frac{1}{\beta_\ell} (z_\ell - x_{\ell+1}) - \sum_{\ell=k}^{k^*} (z_\ell - x_\ell) \right\| \\
&\le \sum_{\ell=k}^{k^*-1} \left| \frac{\beta_\ell - 1}{\beta_\ell} \right| \|z_\ell - x_{\ell+1}\| + \sum_{\ell=k}^{k^*} \frac{1}{\beta_\ell} \|z_\ell - x_{\ell+1}\| + \sum_{\ell=k}^{k^*} \|z_\ell - x_\ell\| \\
&\le \sum_{\ell=k}^{k+M-2} \left| \frac{\beta_\ell - 1}{\beta_\ell} \right| \|z_\ell - x_{\ell+1}\| + \sum_{\ell=k}^{k+M-1} \frac{1}{\beta_\ell} \|z_\ell - x_{\ell+1}\| + \sum_{\ell=k}^{k+M-1} \|z_\ell - x_\ell\| \\
&\le \sum_{\ell=k}^{k+M-1} \frac{2}{\beta_\ell} \|z_\ell - x_{\ell+1}\| + \sum_{\ell=k}^{k+M-1} \|z_\ell - x_\ell\| \quad \text{(since } \beta_\ell \in (0,2)) \\
&= 2 \sum_{\ell=k}^{k+M-1} \|z_\ell - \Pi_{w_\ell} z_\ell\| + \sum_{\ell=k}^{k+M-1} \alpha_\ell \big\|g(\bar{x}_\ell, v_\ell)\big\| \\
&\quad \text{(by the definition of algorithm (3))} \\
&\le \sqrt{2M} \left(4 \sum_{\ell=k}^{k+M-1} \|z_\ell - \Pi_{w_\ell} z_\ell\|^2 + \sum_{\ell=k}^{k+M-1} \alpha_\ell^2 \big\|g(\bar{x}_\ell, v_\ell)\big\|^2 \right)^{1/2},
\end{aligned}$$

where the last step follows from the generic inequality $(\sum_{i=1}^M a_i + \sum_{i=1}^M b_i)^2 \le 2M(\sum_{i=1}^M a_i^2 + \sum_{i=1}^M b_i^2)$ for real numbers a_i, b_i. By rewriting the preceding relation we complete the proof. □

Now we are ready to prove that the feasibility improvement condition holds for the cyclic order constraint sampling scheme.

Proposition 3. *Let Assumptions* 1, 2, 3, *and* 6 *hold, and let* x^* *be a given optimal solution of problem* (1). *Then Algorithm* 1 *generates a sequence* $\{x_k\}$ *such that*

$$\mathbf{E}\left[\,\mathrm{d}^2(x_{k+2m}) \mid \mathcal{F}_k\right] \le \left(1 - \Theta\left(\frac{\beta_k}{m\eta}\right)\right)\mathrm{d}^2(x_k) + \mathcal{O}\left(\frac{m^2\alpha_k^2}{\beta_k}\right)\left(\|x_k - x^*\|^2 + 1\right) \tag{28}$$

for all $k \ge 0$ *with probability* 1.

Proof. Let $\epsilon > 0$ be a scalar. By applying Lemma 3 with $y = \Pi x_k$, we have

$$\begin{aligned}\mathrm{d}^2(x_{k+1}) &\le \|x_{k+1} - \Pi x_k\|^2 \\ &\le (1+\epsilon)\,\mathrm{d}^2(x_k) + (1+1/\epsilon)\alpha_k^2\big\|g(\bar{x}_k, v_k)\big\|^2 - \beta_k(2-\beta_k)\|z_k - \Pi_{w_k} z_k\|^2.\end{aligned}$$

By applying the preceding relation inductively, we obtain

$$\begin{aligned}\mathrm{d}^2(x_{k+2m}) &\le (1+\epsilon)^{2m}\left(\mathrm{d}^2(x_k) + (1+1/\epsilon)\sum_{\ell=k}^{k+M-1}\alpha_\ell^2\big\|g(\bar{x}_\ell, v_\ell)\big\|^2\right) \\ &\quad - \sum_{\ell=k}^{k+M-1}\beta_\ell(2-\beta_\ell)\|z_\ell - \Pi_{w_\ell} z_\ell\|^2 \\ &\le \big(1+\mathcal{O}(\epsilon)\big)\,\mathrm{d}^2(x_k) + \mathcal{O}(1+1/\epsilon)\sum_{\ell=k}^{k+M-1}\alpha_\ell^2\big\|g(\bar{x}_\ell, v_\ell)\big\|^2 \\ &\quad - \Theta(\beta_k)\sum_{\ell=k}^{k+M-1}\|z_\ell - \Pi_{w_\ell} z_\ell\|^2,\end{aligned} \tag{29}$$

where the second inequality uses the facts that β_k is nonincreasing and that $\beta_k/\beta_{k+1} \to 1$ to obtain

$$\min_{\ell=k,\ldots,k+2m-1}\beta_\ell(2-\beta_\ell) \ge \Theta(\beta_k).$$

We apply Lemma 8 with $M = 2m$ (since according to Assumption 6, starting with any k, any particular index will be visited in at most two cycles of samples) and obtain

$$\mathrm{d}^2(x_{k+2m}) \le (1+\mathcal{O}(\epsilon))\,\mathrm{d}^2(x_k) + \mathcal{O}(1+1/\epsilon)\sum_{\ell=k}^{k+M-1}\alpha_\ell^2\|g(\bar{x}_\ell, v_\ell)\|^2 - \frac{\Theta(\beta_k)}{m\eta}\,\mathrm{d}^2(x_k).$$

Let $\epsilon \ll \frac{1}{m\eta}\mathcal{O}(\beta_k)$. Taking conditional expectation on both sides and applying Lemma 4(c), we have

$$\mathbf{E}\left[\,\mathrm{d}^2(x_{k+2m}) \mid \mathcal{F}_k\right] \le \left(1 - \frac{\Theta(\beta_k)}{m\eta}\right)\mathrm{d}^2(x_k) + \mathcal{O}\left(\frac{m^2\alpha_k^2}{\beta_k}\right)\left(\|x_k - x^*\|^2 + 1\right)$$

for all $k \ge 0$ with probability 1. □

4.4. Sample constraints according to a Markov chain.

Finally, we consider the case where the sample constraints X_{w_k} are generated through state transitions of a Markov chain. To ensure that all constraints are sampled adequately, we assume the following.

Assumption 7. *The sequence* $\{w_k\}$ *is generated by an irreducible and aperiodic Markov chain with states* $1, \ldots, m$.

By using an analysis analogous to that of Proposition 3, we obtain the following result.

PROPOSITION 4. *Let Assumptions* 1, 2, 3, *and* 7 *hold, let x^* be a given optimal solution of problem* (1), *and let the sequence $\{x_k\}$ be generated by Algorithm* 1. *Then there exists a positive integer M such that*

$$\mathbf{E}\left[\mathrm{d}^2(x_{k+M}) \mid \mathcal{F}_k\right] \le \left(1 - \Theta\left(\frac{\beta_k}{M\eta}\right)\right)\mathrm{d}^2(x_k) + \mathcal{O}\left(\frac{M^2\alpha_k^2}{\beta_k}\right)\left(\|x_k - x^*\|^2 + 1\right) \tag{30}$$

for all $k \ge 0$ with probability 1.

Proof. According to Assumption 7, the Markov chain is irreducible and aperiodic. Therefore its invariant distribution, denoted by $\xi \in \Re^m$, satisfies for some $\varepsilon > 0$

$$\min_{i=1,\dots,m} \xi_i > \varepsilon,$$

and moreover, there exist scalars $\rho \in (0,1)$ and $c > 0$ such that

$$\left|\mathbf{P}(w_{k+\ell} = X_i \mid \mathcal{F}_k) - \xi_i\right| \le c \cdot \rho^\ell, \qquad i = 1,\dots,m, \ \ \forall\, k \ge 0,\ \ell \ge 0, \qquad \text{w.p.1.}$$

We let M be a sufficiently large integer, such that

$$\min_{i=1,\dots,m} \mathbf{P}(w_{k+M-1} = X_i \mid \mathcal{F}_k) \ge \min_{i=1,\dots,m} \xi_i - c\rho^M \ge \Theta(\varepsilon) > 0 \qquad \forall\, k \ge 0 \quad \text{w.p.1.}$$

This implies that, starting with any w_k, there is a positive probability $\Theta(\varepsilon)$ to reach any particular index in $\{1,\dots,m\}$ in the next M samples.

By using this fact together with Lemma 8, we obtain

$$\mathbf{P}\left(\frac{1}{2M\eta}\,\mathrm{d}^2(x_k) \le 4\sum_{\ell=k}^{k+M-1} \|z_\ell - \Pi_{w_\ell} z_\ell\|^2 + \sum_{\ell=k}^{k+M-1} \alpha_\ell^2 \|g(\bar{x}_\ell, v_\ell)\|^2 \ \middle|\ \mathcal{F}_k\right) \ge \Theta(\varepsilon).$$

It follows that

$$\begin{aligned}&\mathbf{E}\left[4\sum_{\ell=k}^{k+M-1} \|z_\ell - \Pi_{w_\ell} z_\ell\|^2 + \sum_{\ell=k}^{k+M-1} \alpha_\ell^2 \|g(\bar{x}_\ell, v_\ell)\|^2 \ \middle|\ \mathcal{F}_k\right] \\ &\quad \ge \Theta(\varepsilon) \cdot \frac{1}{2M\eta}\,\mathrm{d}^2(x_k) + (1 - \Theta(\varepsilon)) \cdot 0.\end{aligned}$$

By rewriting the preceding relation and applying Lemma 5(a), we obtain

$$\mathbf{E}\left[\sum_{\ell=k}^{k+M-1} \|z_\ell - \Pi_{w_\ell} z_\ell\|^2 \ \middle|\ \mathcal{F}_k\right] \ge \frac{\Theta(\varepsilon)}{8M\eta}\,\mathrm{d}^2(x_k) - \mathcal{O}(\alpha_k^2)\big(\|x_k - x^*\|^2 + 1\big). \tag{31}$$

The rest of the proof follows a line of analysis like the one of Proposition 3, with $2m$ replaced with M. Similar to (29), we have

$$\begin{aligned}\mathrm{d}^2(x_{k+M}) &\le (1 + \mathcal{O}(\epsilon))\,\mathrm{d}^2(x_k) + \mathcal{O}(1 + 1/\epsilon) \sum_{\ell=k}^{k+M-1} \alpha_\ell^2 \big\|g(\bar{x}_\ell, v_\ell)\big\|^2 \\ &\quad - \Theta(\beta_k) \sum_{\ell=k}^{k+M-1} \|z_\ell - \Pi_{w_\ell} z_\ell\|^2.\end{aligned}$$

Taking expectation on both sides, we obtain

$$\begin{aligned}\mathbf{E}\left[\, \mathrm{d}^2(x_{k+M}) \mid \mathcal{F}_k\right] &\leq \big(1+\mathcal{O}(\epsilon)\big)\,\mathrm{d}^2(x_k) + \mathcal{O}(1+1/\epsilon)\mathbf{E}\left[\sum_{\ell=k}^{k+M-1} \alpha_\ell^2 \big\|g(\bar{x}_\ell, v_\ell)\big\|^2 \;\Bigg|\; \mathcal{F}_k\right] \\ &\quad - \Theta(\beta_k)\mathbf{E}\left[\sum_{\ell=k}^{k+M-1} \|z_\ell - \Pi_{w_\ell} z_\ell\|^2 \;\Bigg|\; \mathcal{F}_k\right] \\ &\leq (1+\mathcal{O}(\epsilon))\,\mathrm{d}^2(x_k) + \mathcal{O}(1+1/\epsilon)\alpha_k^2\left(\|x_k - x^*\|^2 + 1\right) \\ &\quad - \Theta\left(\frac{\beta_k}{M\eta}\right)\mathrm{d}^2(x_k) \\ &\leq \left(1 - \Theta\left(\frac{\beta_k}{M\eta}\right)\right)\mathrm{d}^2(x_k) + \mathcal{O}\left(\frac{M^2\alpha_k^2}{\beta_k}\right)(\|x_k - x^*\|^2 + 1),\end{aligned}$$

where the second relation uses (31) and Lemma 5(c), and the third relation holds by letting $\epsilon \leq \Theta(\frac{\beta_k}{M\eta})$. □

5. Sampling schemes for subgradients/component functions.

In this section, we focus on sampling schemes for the subgradients/component functions that satisfy the optimality improvement condition required by the coupled convergence theorem (Theorem 1), i.e.,

$$\begin{aligned}\mathbf{E}\left[\|x_{k+N} - x^*\|^2 \mid \mathcal{F}_k\right] &\leq \|x_k - x^*\|^2 - 2\left(\sum_{\ell=k}^{k+N-1}\alpha_\ell\right)\big(f(x_k) - f^*\big) \\ &\quad + \mathcal{O}(\alpha_k^2)\big(\|x_k - x^*\|^2 + 1\big) + \mathcal{O}(\beta_k^2)\,\mathrm{d}^2(x_k)\end{aligned}$$

with probability 1, where $k = 0, N, 2N, \ldots$, and N is a positive integer.

In what follows, we consider the case of unbiased samples and the case of cyclic samples. Either one of the following subgradient/function sampling schemes can be combined with any one of the constraint sampling schemes in section 4 to yield a convergent incremental algorithm.

5.1. Unbiased sample subgradients/component functions.

We start with the relatively simple case where the sample component functions chosen by the algorithm are conditionally unbiased. We assume the following.

Assumption 8. *Let each $g(x, v_k)$ be the subgradient of a random component function $f_{v_k} : \Re^n \mapsto \Re$ at x,*

$$g(x, v_k) \in \partial f_{v_k}(x) \qquad \forall\, x \in \Re^n,$$

and let the random variables v_k, $k = 0, 1, \ldots$, be such that

$$\mathbf{E}\big[f_{v_k}(x) \mid \mathcal{F}_k\big] = f(x) \qquad \forall\, x \in \Re^n, \quad k \geq 0, \qquad \text{w.p.1.} \tag{32}$$

We use a standard line of argument for gradient descent to obtain the optimality improvement inequality.

Proposition 5. *Let Assumptions* 1, 2, 3*, and* 8 *hold, and let x^* be a given optimal solution of problem* (1)*. Then Algorithm* 1 *generates a sequence $\{x_k\}$ such that*

$$\begin{aligned}\mathbf{E}\big[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\big] &\leq \|x_k - x^*\|^2 - 2\alpha_k\big(f(x_k) - f^*\big) \\ &\quad + \mathcal{O}(\alpha_k^2)\big(\|x_k - x^*\|^2 + 1\big) + \mathcal{O}(\beta_k^2)\,\mathrm{d}^2(x_k)\end{aligned}$$

for all $k \geq 0$ with probability 1*.*

Proof. By applying Lemma 3 with $y = x^*$, we obtain

$$\|x_{k+1} - x^*\|^2 \le \|x_k - x^*\|^2 - 2\alpha_k g(\bar{x}_k, v_k)'(x_k - x^*) + \alpha_k^2 \|g(\bar{x}_k, v_k)\|^2. \tag{33}$$

Taking conditional expectation on both sides and applying Lemma 4(c) yields

$$\begin{aligned}\mathbf{E}\left[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\right] \le {}& \|x_k - x^*\|^2 \\ & - 2\alpha_k \mathbf{E}\left[g(\bar{x}_k, v_k)'(x_k - x^*) \mid \mathcal{F}_k\right] + \alpha_k^2 O\big(\|x_k - x^*\|^2 + 1\big).\end{aligned} \tag{34}$$

According to Assumption 8, since $x_k \in \mathcal{F}_k$, we have

$$\begin{aligned}&\mathbf{E}\big[g(\bar{x}_k, v_k)'(x_k - x^*) \mid \mathcal{F}_k\big] \\ &\quad = \mathbf{E}\big[g(\bar{x}_k, v_k)'(\bar{x}_k - x^*) \mid \mathcal{F}_k\big] + \mathbf{E}\big[g(\bar{x}_k, v_k)'(x_k - \bar{x}_k) \mid \mathcal{F}_k\big] \\ &\quad \ge \mathbf{E}\big[f(\bar{x}_k) - f^* \mid \mathcal{F}_k\big] + \mathbf{E}\big[g(\bar{x}_k, v_k)'(x_k - \bar{x}_k) \mid \mathcal{F}_k\big] \\ &\quad = f(x_k) - f^* + \mathbf{E}\big[f(\bar{x}_k) - f(x_k) \mid \mathcal{F}_k\big] + \mathbf{E}\big[g(\bar{x}_k, v_k)'(x_k - \bar{x}_k) \mid \mathcal{F}_k\big] \\ &\quad \ge f(x_k) - f^* + \mathbf{E}\big[g(x_k, v_k)'(\bar{x}_k - x_k) + g(\bar{x}_k, v_k)'(x_k - \bar{x}_k) \mid \mathcal{F}_k\big] \\ &\quad \ge f(x_k) - f^* - \frac{\alpha_k}{2}\mathbf{E}\big[\|g(x_k, v_k)\|^2 + \|g(\bar{x}_k, v_k)\|^2 \mid \mathcal{F}_k\big] - \frac{1}{\alpha_k}\mathbf{E}\left[\|\bar{x}_k - x_k\|^2 \mid \mathcal{F}_k\right] \\ &\quad \ge f(x_k) - f^* - \alpha_k O\big(\|x_k - x^*\|^2 + 1\big) - \frac{1}{\alpha_k}\left(\alpha_k^2 O\big(\|x_k - x^*\|^2 + 1\big) + \beta_k^2\, \mathrm{d}^2(x_k)\right) \\ &\quad \ge f(x_k) - f^* - \alpha_k O\big(\|x_k - x^*\|^2 + 1\big) - \frac{\beta_k^2}{\alpha_k}\, \mathrm{d}^2(x_k),\end{aligned}$$

where the first and second inequalities use the definition of subgradients, the third inequality uses $2ab \le a^2 + b^2$ for any $a, b \in \Re$, and the fourth inequality uses Assumption 1 and Lemma 4(c), (d). Finally, we apply the preceding relation to (34) and obtain

$$\begin{aligned}&\mathbf{E}\big[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\big] \\ &\quad \le \|x_k - x^*\|^2 - 2\alpha_k(f(x_k) - f^*) + \mathcal{O}(\alpha_k^2)\big(\|x_k - x^*\|^2 + 1\big) + \mathcal{O}(\beta_k^2)\, \mathrm{d}^2(x_k)\end{aligned}$$

for all $k \ge 0$ with probability 1. □

5.2. Cyclic sample subgradients/component functions. Now we consider the analytically more challenging case, where the subgradients are sampled in a cyclic manner. More specifically, we assume that the subgradient samples are associated with a "cyclic" sequence of component functions.

ASSUMPTION 9. *Each $g(x, v_k)$ is the subgradient of function $f_{v_k} : \Re^n \mapsto \Re$ at x, i.e.,*

$$g(x, v_k) \in \partial f_{v_k}(x) \qquad \forall\, x \in \Re^n,$$

the random variables v_k, $k = 0, 1, \ldots$, are such that for some integer $N > 0$,

$$\frac{1}{N}\sum_{\ell=tN}^{(t+1)N-1} \mathbf{E}\big[f_{v_\ell}(x) \mid \mathcal{F}_{tN}\big] = f(x) \qquad \forall\, x \in \Re^n, \qquad t \ge 0, \qquad \text{w.p.1.} \tag{35}$$

In the next proposition, we show that the optimality improvement condition is satisfied when we select the component functions and their subgradients according to a cyclic order, either randomly or deterministically. The proof idea is to consider the total optimality improvement with a cycle of N iterations.

Proposition 6. *Let Assumptions* 1, 2, 3, *and* 9 *hold, and let x^* be a given optimal solution of problem* (1). *Then Algorithm* 1 *generates a sequence* $\{x_k\}$ *such that*

$$\begin{aligned}\mathbf{E}\big[\|x_{k+N}-x^*\|^2 \mid \mathcal{F}_k\big] \le \|x_k-x^*\|^2 - 2\left(\sum_{\ell=k}^{k+N-1}\alpha_\ell\right)\big(f(x_k)-f^*\big)\\ + \mathcal{O}(N\alpha_k^2)\big(\|x_k-x^*\|^2+1\big) + \mathcal{O}(N\beta_k^2)\,\mathrm{d}^2(x_k)\end{aligned}$$

for all $k = 0, N, 2N, \ldots$ *with probability* 1.

Proof. Following the line of analysis of Proposition 5 and applying (33) repeatedly, we obtain

$$\|x_{k+N}-x^*\|^2 \le \|x_k-x^*\|^2 - 2\sum_{\ell=k}^{k+N-1}\alpha_\ell g(\bar{x}_\ell,v_\ell)'(x_\ell-x^*) + \sum_{\ell=k}^{k+N-1}\alpha_\ell^2\|g(\bar{x}_\ell,v_\ell)\|^2.$$

By taking conditional expectation on both sides and by applying Lemma 5(c), we further obtain

$$\begin{aligned}&\mathbf{E}\left[\|x_{k+N}-x^*\|^2 \mid \mathcal{F}_k\right]\\ &\quad\le \|x_k-x^*\|^2 - 2\sum_{\ell=k}^{k+N-1}\alpha_\ell\mathbf{E}\left[g(\bar{x}_\ell,v_\ell)'(x_\ell-x^*) \mid \mathcal{F}_k\right] + \mathcal{O}(N\alpha_k^2)\big(\|x_k-x^*\|^2+1\big)\end{aligned} \tag{36}$$

for all $k = 0, N, 2N, \ldots$ with probability 1.

For $\ell = k, \ldots, k+N-1$, we have

$$g(\bar{x}_\ell,v_\ell)'(x_\ell-x^*) = g(\bar{x}_\ell,v_\ell)'(\bar{x}_\ell-x^*) + g(\bar{x}_\ell,v_\ell)'(x_\ell-\bar{x}_\ell).$$

Since $g(x,v_\ell)\in\partial f_{v_\ell}(x)$ for all x, we apply the definition of subgradients and obtain

$$g(\bar{x}_\ell,v_\ell)'(\bar{x}_\ell-x^*) \ge f_{v_\ell}(\bar{x}_\ell)-f^* \ge f_{v_\ell}(x_k)-f^*+g(x_k,v_\ell)'(\bar{x}_\ell-x_k).$$

Combining the preceding two relations, we obtain

$$g(\bar{x}_\ell,v_\ell)'(x_\ell-x^*) \ge f_{v_\ell}(x_k)-f^*+g(x_k,v_\ell)'(\bar{x}_\ell-x_k)+g(\bar{x}_\ell,v_\ell)'(x_\ell-\bar{x}_\ell).$$

By taking expectation on both sides, we further obtain

$$\begin{aligned}&\mathbf{E}\left[g(\bar{x}_\ell,v_\ell)'(x_\ell-x^*) \mid F_k\right]\\ &\quad\ge \mathbf{E}\left[f_{v_\ell}(x_k) \mid F_k\right]-f^*+\mathbf{E}\left[g(\bar{x}_\ell,v_\ell)'(x_\ell-\bar{x}_\ell)+g(x_k,v_\ell)'(\bar{x}_\ell-x_k) \mid F_k\right]\\ &\quad\ge \mathbf{E}\left[f_{v_\ell}(x_k) \mid F_k\right]-f^*\\ &\qquad -\mathcal{O}(\alpha_\ell)\mathbf{E}\left[\|g(\bar{x}_\ell,v_\ell)\|^2+\|g(x_k,v_\ell)\|^2 \mid F_k\right]-\mathcal{O}(1/\alpha_\ell)\mathbf{E}\left[\|\bar{x}_\ell-x_k\|^2 \mid F_k\right]\\ &\quad\ge \mathbf{E}\left[f_{v_\ell}(x_k) \mid F_k\right]-f^*-\mathcal{O}(\alpha_\ell)\big(\|x_k-x^*\|^2+1\big)-\mathcal{O}\left(\frac{\beta_k^2}{\alpha_\ell}\right)\mathrm{d}^2(x_k),\end{aligned}$$

where the second inequality uses the basic fact $2a'b \le \|a\|^2+\|b\|^2$ for $a,b\in\Re^n$, and the last inequality uses Assumption 1 and Lemma 5(a), (d). Then from Assumption 9 we have

$$
\begin{aligned}
&\sum_{\ell=k}^{k+N-1} \alpha_\ell \mathbf{E}\left[g(\bar{x}_\ell, v_\ell)'(x_\ell - x^*) \mid F_k\right] \\
&\quad \geq \sum_{\ell=k}^{k+N-1} \alpha_\ell \Big(\mathbf{E}\left[f_{v_\ell}(x_k) \mid F_k\right] - f^*\Big) \\
&\qquad - \sum_{\ell=k}^{k+N-1} \mathcal{O}(\alpha_k \alpha_\ell)\big(\|x_k - x^*\|^2 + 1\big) - \mathcal{O}\left(N\beta_k^2\right) \mathrm{d}^2(x_k) \\
&\quad = \sum_{\ell=k}^{k+N-1} \alpha_\ell \big(f(x_k) - f^*\big) - \mathcal{O}(N\alpha_k^2)\big(\|x_k - x^*\|^2 + 1\big) - \mathcal{O}\left(N\beta_k^2\right) \mathrm{d}^2(x_k)
\end{aligned}
$$

with probability 1. Finally, we apply the preceding relation to (36) and complete the proof. ◻

6. Convergence of randomized constraint projection-proximal algorithms.

In sections 4 and 5, we have considered a number of sampling schemes for both the constraints and component functions such that the feasibility and optimality improvement conditions required by the coupled convergence theorems, Theorems 1 and 2, are satisfied. Now we will combine the preceding results and apply the coupled convergence theorems. The following theorem collects various combinations of conditions under which our algorithm converges almost surely.

THEOREM 3 (almost sure convergence and rate of convergence). *Let Assumptions* 1, 2, and 3 *hold, and consider the incremental constraint projection-proximal Algorithm* 1. *Assume that the constraint sampling scheme satisfies any one of the following:*

(i) *The constraints are sampled randomly as in Assumption* 4.
(ii) *The constraints are sampled adaptively according to the most distant set criterion as in Assumption* 5.
(iii) *The constraints are sampled cyclically as in Assumption* 6.
(iv) *The constraints are sampled using a Markov chain as in Assumption* 7.

Assume further that the subgradient/component function sampling scheme satisfies either of the following:

(i) *The component samples are conditionally unbiased as in Assumption* 8.
(ii) *The component samples are unbiased over a cycle as in Assumption* 9.

Then Algorithm 1 *generates a sequence of random variables* $\{x_k\}$ *that converges almost surely to a random point in the set of optimal solutions of the convex optimization problem* (1). *In addition, if the stepsizes satisfy* $\alpha_k = \Theta(1/\sqrt{k})$, $\beta_k = \Theta(1)$ *instead of Assumption* 2, *we have*

$$
\mathbf{E}\left[f\left(\frac{1}{k}\sum_{t=1}^{k} \Pi_X x_t\right)\right] \leq f^* + \mathcal{O}\left(\frac{1}{\sqrt{k}}\right), \qquad \mathbf{E}\left[\mathrm{d}^2\left(\frac{1}{k}\sum_{t=1}^{k} x_t\right)\right] \leq \mathcal{O}\left(\frac{\log k}{k}\right).
$$

Proof. The proof is obtained by combining Propositions 1, 2, 3, and 4 and Propositions 5 and 6, in conjunction with Theorems 1 and 2. ◻

7. Numerical experiments.

In this section, we test Algorithm 1 on randomly generated instances of problem (1) and study its performance in various settings. We let the constraint be a system of M linear equalities, each generated randomly. We let

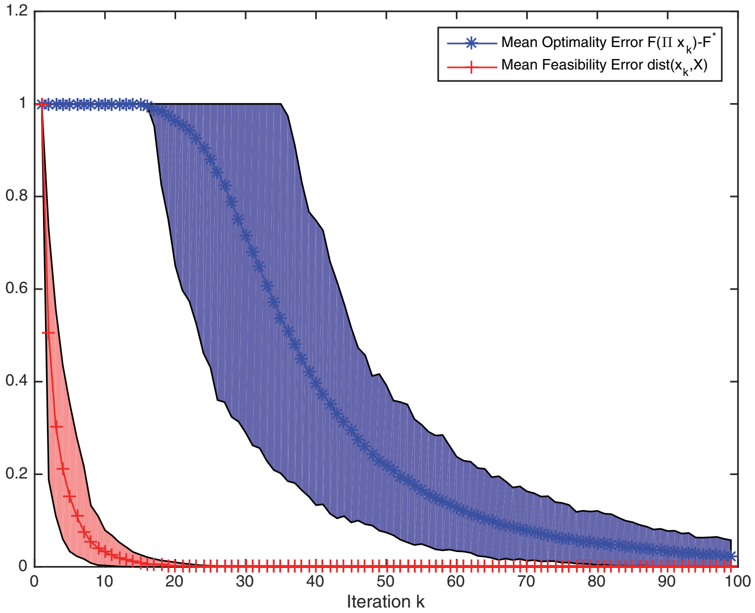

FIG. 1. *Convergence of Algorithm* 1 *($M = N = 50$). We plot the mean and* 90% *confidence intervals of the optimality error and feasibility error, for which* 100 *random trajectories are generated. They have been normalized to start at* 1*. We observe that the optimality error diminishes to* 0 *at a rate slower than the feasibility error. The feasibility error decreases nearly at a geometric rate in the initial iterations.*

the objective be the sum of N quadratic functions, each randomly generated. We have conducted extensive experiments of the random projection method using various sampling schemes and stepsizes. Some representative results are illustrated in Figures 1, 2, and 3. Figure 1 plots the convergence trajectories of the feasibility and optimality errors (which are normalized to start at 1). The mean and 90% confidence intervals of the errors based on 100 trial runs are illustrated in Figure 1. Figure 2 plots the convergence of errors when different gradient sampling schemes are used. Figure 3 plots the convergence of errors when different constraint sampling schemes are used.

The numerical results validate the convergence rates predicted by the theory. Our observations are summarized as follows:

1. The feasibility error converges much faster than the optimality error converges. We have tested various choices of stepsizes. Regardless of the stepsizes, the feasibility error always decreases at a faster rate as long as convergence is ensured. This validates our analysis of the coupled convergence process.
2. When picking stepsizes as suggested in Theorem 2 and 3, we observe that the feasibility error is bounded by $\mathcal{O}(\log k/k)$, while the optimality error seems to be between $\mathcal{O}(1/k)$ and $\mathcal{O}(1/\sqrt{k})$. This validates the rate of convergence predicted in Theorems 2 and 3.
3. When taking β_k to be a constant, the feasibility error decreases at a geometric rate in the first few iterations. After a few projections, the iterates become nearly feasible. This can be understood using the feasible improve-

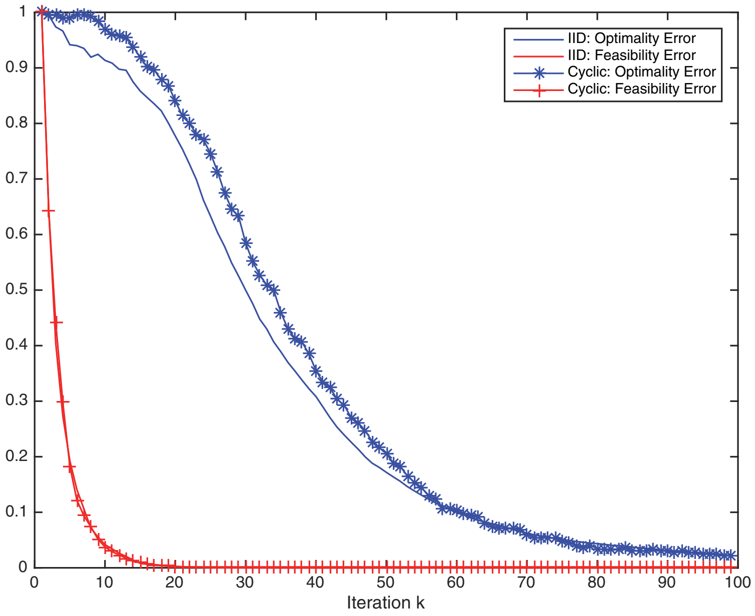

FIG. 2. *Comparison of gradient sampling schemes. We apply Algorithm* 1 *using two gradient sampling schemes: i.i.d. sampling and cyclic sampling. Both experiments use identical stepsizes and initial point, as well as i.i.d. uniform samples of constraints. We have experimented with various parameters and dimensions of the problem. We note that the trajectories are quite sensitive to parameters other than the gradient sampling scheme. In all of our experiments, the independent sampling schemes demonstrate more robust convergence of the optimality error compared to the cyclic sampling schemes. In contrast, the feasibility error is somewhat insensitive to the choice of gradient sampling schemes.*

ment inequality given in Theorem 1, which is almost a contraction when the feasibility error is large.

4. The algorithm with random gradient/constraint sampling has better worst-case performance than the one with cyclic gradient/constraint sampling. The likely reason is that random sampling may break an unfavorable order of component functions/constraints that may slow down the convergence.
5. By sampling constraints adaptively, e.g., choosing the most distant set, the algorithm achieves a substantially better convergence rate than algorithms using other schemes. However, we remark that projecting to the most distant set requires identifying the set, which is time-consuming or even impossible in many practical contexts.
6. The performance of Markov constraint sampling is very sensitive to the mixing rate and invariant distribution of the Markov chain. Indeed, both i.i.d. uniform sampling and cyclic sampling can be viewed as special cases of Markov sampling. The general Algorithm 1 works with any Markov sampling scheme that is recurrent. In particular applications, one may design a customized Markov chain Monte Carlo method as the sampling oracle, in order to achieve the best algorithm efficiency.

We note that the convergence properties of the algorithm also depend on other prob-

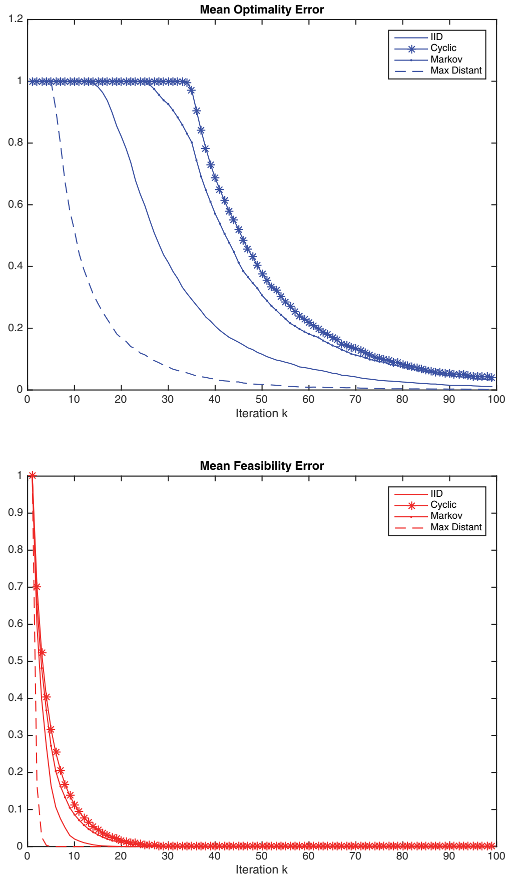

Fig. 3. *Comparison of constraint sampling schemes. We study four constraint sampling schemes: independent uniform sampling, cyclic sampling, Markov sampling, and using the most distant set for projection. In all of our experiments, the most distant projection method outperforms the other three. The i.i.d. uniform sampling scheme is more robust than the cyclic sampling scheme. Both the optimality error and the feasibility error are very sensitive to schemes of constraint sampling.*

lem parameters, e.g., N, M, the condition number, variance of sample gradients, and variance of sample constraints. We also note that the sampling oracle is not to be chosen in many practical contexts, e.g., applications that process streaming data. There could be a large number of possible situations, depending on the application areas. Thus it is beyond the scope of the current paper to provide a case-by-case convergence rate analysis and numerical experiment. A in-depth customized analysis and experiment addressing specific applications would be a direction for future work.

8. Conclusions. In this paper, we have proposed a class of stochastic algorithms, based on subgradient projection and proximal methods, which alternate between random optimality updates and random feasibility updates. We characterized the behavior of these algorithms in terms of two coupled improvement processes: optimality improvement and feasibility improvement. We have provided a unified convergence and rate of convergence framework, based on the coupled convergence theorem, which serves as a modular architecture for convergence analysis and can accommodate a broad variety of sampling schemes, such as independent sampling, cyclic sampling, Markov chain sampling, etc. We show that the optimality error decreases on the order of $\mathcal{O}(1/\sqrt{k})$ and the feasibility error decreases on the order of $\mathcal{O}(\log k/k)$. The convergence rate of optimality error is nonimprovable for stochastic first-order methods, implying that using random projection does not deteriorate the convergence rate up to a constant.

For future research, an important direction is to customize the convergence rate analysis to specific applications. It is also interesting to consider modifications of our algorithm involving finite memory and multiple recent samples. Related research on this subject includes asynchronous algorithms using "delayed" subgradients with applications in parallel computing (see, e.g., [ANaB01]). Another extension is to analyze problems with an infinite number of constraints.

REFERENCES

[ABRW12] A. Agarwal, P. Bartlett, P. Ravikumar, and M. Wainwright, *Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization*, IEEE Trans. Inform. Theory, 58 (2012), pp. 3235–3249.

[ANaB01] D. P. Bertsekas, A. Nedic, and V. S. Borkar, *Distributed asynchronous incremental subgradient methods*, Stud. Comput. Math., 8 (2001), pp. 381–407.

[Bau96] H. H. Bauschke, *Projection Algorithms and Monotone Operators*, Ph.D. thesis, Simon Frazer University, Burnaby, BC, Canada, 1996.

[BB96] H. H. Bauschke and J. M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.

[BBL97] H. Bauschke, J. M. Borwein, and A. S. Lewis, *The method of cyclic projections for closed convex sets in hilbert space*, Contemp. Math., 204 (1997), pp. 1–38.

[Ber11] D. P. Bertsekas, *Incremental proximal methods for large scale convex optimization*, Math. Program. Ser. B, 129 (2011), pp. 163–195.

[BNO03] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, Belmont, MA, 2003.

[Bor08] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*, Cambridge University Press, Cambridge, UK, 2008.

[BT89] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, Belmont, MA, 1989.

[CS08] A. Cegielski and A. Suchocka, *Relaxed alternating projection methods*, SIAM J. Optim., 19 (2008), pp. 1093–1106.

[DH06a] F. Deutsch and H. Hundal, *The rate of convergence for the cyclic projections algorithm* I*: Angles between convex sets*, J. Approx. Theory, 142 (2006), pp. 36–55.

[DH06b] F. Deutsch and H. Hundal, *The rate of convergence for the cyclic projections algorithm* II*: Norms of nonlinear operators*, J. Approx. Theory, 142 (2006), pp. 56–82.

[DH08] F. Deutsch and H. Hundal, *The rate of convergence for the cyclic projections algorithm* III*: Regularity of convex sets*, J. Approx. Theory, 155 (2008), pp. 155–184.

[GPR67] L. G. Gubin, B. T. Polyak, and E. V. Raik, *The method of projections for finding the common point of convex sets*, USSR Comput. Math. Math. Phys., 7 (1967), pp. 1211–1228.

[Hal62] I. Halperin, *The product of projection operators*, Acta Sci. Math., 23 (1962), pp. 96–99.

[KY03] H. J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, New York, 2003.

[LL10] D. Leventhal and A. S. Lewis, *Randomized methods for linear constraints: Convergence rates and conditioning*, *Math. Oper. Res.*, 35 (2010), pp. 641–654.

[LM08] A. S. Lewis and J. Malick, *Alternating projections on manifolds*, Math. Oper. Res., 33 (2008), pp. 216–234.

[NB00] A. Nedić and D. P. Bertsekas, *Convergence rate of incremental subgradient algorithms*, in Stochastic Optimization: Algorithms and Applications, S. Uryasev and P. M. Pardalos, eds., Appl. Optim. 54, Springer, New York, 2000, pp. 263–304.

[NB01] A. Nedić and D. P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM J. Optim., 12 (2001), pp. 109–138.

[Ned10] A. Nedić, *Random projection algorithms for convex set intersection problems*, in Proceedings of the 49th IEEE Conference on Decision and Control, Atlanta, GA, 2010, pp. 7655–7660.

[Ned11] A. Nedić, *Random algorithms for convex minimization problems*, Math. Program. Ser. B, 129 (2011), pp. 225–253.

[RS57] H. Robbins and D. Siegmund, *A convergence theorem for nonnegative almost supermartingales and some applications*, in Optimizing Methods in Statistics, J. S. Rostagi, ed., Academic Press, New York, pp. 233–257.

[SZ12] O. Shamir and T. Zhang, *Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes*, in Proceedings of the 30th International Conference on Machine Learning, 2013, pp. 71–79.

[Tse90] P. Tseng, *Successive Projection Under a Quasi-Cyclic Order*, Report LIDS-P-1938, MIT, Cambridge, MA, 1990.

[vN50] J. von Neumann, *Functional Operators*, Princeton University Press, Princeton, NJ, 1950.

[WB12] M. Wang and D. P. Bertsekas, *Incremental constraint projection methods for variational inequalities*, Math. Program., 150 (2015), pp. 321–363.