Massachusetts Institute of Technology, Cambridge, MA
Laboratory for Information and Decision Systems
Report LIDS-P-2907, July 2013

Incremental Constraint Projection-Proximal Methods for Nonsmooth Convex Optimization

Mengdi Wang
mengdiw@princeton.edu

Dimitri P. Bertsekas*
dimitrib@mit.edu

Abstract

We consider convex optimization problems with structures that are suitable for stochastic sampling. In particular, we focus on problems where the objective function is an expected value or is a sum of a large number of component functions, and the constraint set is the intersection of a large number of simpler sets. We propose an algorithmic framework for projection-proximal methods using random subgradient/function updates and random constraint updates, which contain as special cases several known algorithms as well as 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 towards feasibility and progress towards optimality. Moreover, we consider a number of typical sampling/randomization schemes for the subgradients/component functions and the constraints, and analyze their performance using our unified convergence framework.

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

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 $\Re^n$. Moreover, we allow the cost 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 method has the form

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

*Mengdi Wang is with the Department of Operations Research and Financial Engineering, Princeton University. Dimitri Bertsekas is with the Department of Electrical Engineering and Computer Science, and the Laboratory for Information and Decision Systems (LIDS), M.I.T. Work supported by the Air Force Grant FA9550-10-1-0412.

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 $g'(y-x) \leq f(y) - f(x)$ 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], Prop. 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 classical methods are often inefficient and difficult to use, especially when the constraint set X is complicated [cf. Eq. (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. 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 these methods, by operating with a single set X_i at each iteration, as is done in random and cyclic projection methods that are widely used to solve the feasibility problem of finding some point in X.

Another difficulty arises when f is either the sum of a large number of component functions or is an expected value, i.e., $f(x) = \mathbf{E}\big\{f_v(x)\big\}$ [cf. Eq. (3)]. 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 additional 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 $f_{v_k}(x)$.

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. In particular, we will combine the *incremental constraint projection algorithm*

$$x_{k+1} = \Pi_{w_k}\big[x_k - \alpha_k g(x_k, v_k)\big], \tag{4}$$

and the *incremental constraint proximal algorithm*

$$x_{k+1} = \text{argmin}_{x \in X_{w_k}} \left[f_{v_k}(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right] = \Pi_{w_k}\big[x_k - \alpha_k g(x_{k+1}, v_k)\big], \tag{5}$$

where Π_{w_k} denotes the Euclidean projection onto a set X_{w_k}, $\{w_k\}$ is a sequence of random variables taking values in $\{1, \ldots, m\}$, and $\{v_k\}$ is a sequence of random variables generated by some probabilistic process. An interesting special case is when X is a polyhedral set, i.e., the intersection of a finite number of halfspaces. Then these algorithms involve successive projections onto or minimizations over halfspaces, which are easy to implement and computationally inexpensive. Another interesting special case is when f is an expected value and its value or subgradient can only be obtained through sampling. The proposed algorithms are well suited for problems of such type, and have an "online" focus that uses small storage and rapid updates.

The purpose of this paper is to present a unified analytical framework for the convergence of algorithms (4), (5), and various combinations and extensions. In particular, we focus on the class of incremental algorithms that involve random optimality updates and random feasibility updates, of the form

$$z_k = x_k - \alpha_k g(\bar{x}_k, v_k), \qquad x_{k+1} = z_k - \beta_k \left(z_k - \Pi_{w_k} z_k\right), \tag{6}$$

where $\bar{x}_k$ is a random variable "close" to x_k such as

$$\bar{x}_k = x_k, \qquad \text{or} \qquad \bar{x}_k = x_{k+1}, \tag{7}$$

and $\{\beta_k\}$ is a sequence of positive scalars. We refer to Eqs. (6)-(7) as the *incremental constraint projection-proximal method*. In the case where $\bar{x}_k = x_k$, the kth iteration of algorithm (6) is a subgradient projection

step and takes the form of Eq. (4). In the other case where $\bar{x}_k = x_{k+1}$, the corresponding iteration is a proximal step and takes the form of Eq. (5). Thus our algorithm (6)-(7) is a mixed version of the incremental projection algorithm (4) and the incremental proximal algorithm (5). An interesting case is when f has the form

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

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 (6)-(7) can adaptively choose between a projection step and a proximal step, based on the current sample component function.

Our algorithm (6)-(7) 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. This is an important insight that helps to understand the convergence mechanism. We will propose a unified analytical framework, which involves an intricate interplay between the progress of feasibility updates and the progress of optimality updates, and their associated stepsizes β_k and α_k. In particular, we will provide a coupled convergence theorem which requires that the algorithm operates 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, as we will demonstrate with both analytical and experimental results.

Another important aspect of our analysis relates to the source of the samples v_k and w_k. For example, a common situation arises from applications involving 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 $g(\bar{x}_k, v_k)$ 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.

The proposed algorithmic framework (6) 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 on feasibility, incremental/stochastic gradient, stochastic approximation, and projection-proximal methods, which we will now summarize.

The feasibility update of algorithm (6) is strongly related to known methods for feasibility problems. In particular, when $f(x) = 0$, $g(\bar{x}_k, v_k) = 0$ and $\beta_k = 1$ for all k, we obtain a successive projection algorithm for finding some $x \in X = \cap_{i=1}^m X_i$. For the case where m is a large number and each X_i is a closed convex set with a simple form, incremental methods that make successive projections on the component sets X_i have a long history, starting with von Neumann [vN50], and followed by many other authors: Halperin [Hal62], Gubin et al. [GPR67], Tseng [Tse90], Bauschke et al. [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 (6), especially when f is given as an expected value [cf. Eq. (3)], is closely related to stochastic approximation methods. In the case where $X = X_{w_k}$ for all k and f is given as an expected value, our method becomes a stochastic approximation method for optimization problems, which has been well known in the literature. In particular, we make the typical assumptions $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$ on $\{\alpha_k\}$ in order to establish convergence (see e.g., the textbooks by Bertsekas and Tsitsiklis [BT89], by Kushner and Yin [KY03], and by Borkar [Bor08]). Moreover, similar to several sources on convergence analysis of stochastic algorithms, we use a supermartingale convergence theorem.

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 a form of randomized selection of constraint sets, which can be viewed as a special case of algorithm (6) with $\bar{x}_k = x_k$. It also discusses interesting special cases, where for example the sets X_i are specified by convex inequality constraints. 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] (also discussed in the context of a survey of incremental optimization methods in [Ber12]). 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 (6) 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]. This work also points out that a special case of incremental constraint projections on sets X_i can be implemented via the proximal iterations. It is less general than the current work in that it does not fully consider the randomization of constraints, and it requires the objective function to be Lipchitz continuous.

Another related methodology is the sample average approximation method (SAA); see Shapiro et al. [SDR09], Kleywegt et al. [KSHdM02], Nemirovski et al. [NJLS09]. It solves a sequence of approximate optimization problems that are obtained based on samples of f, and generates approximate solutions that converge to an optimal solution at a rate determined by the central limit theorem. Let us also mention the robust stochastic approximation method proposed by [NJLS09]. It is a modified stochastic approximation method that can use a fixed stepsize instead of a diminishing one. Both these methods focus on optimization problems in which f involves expected values. However, they do not consider constraint sampling as we focus on in this paper. In contrast, our incremental projection method requires a diminishing stepsize due to the uncertainty in processing constraints.

Recently, the idea of an incremental method with constraint randomization has been extended to solution of strongly monotone 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 to use incremental constraint projection, analyzes the two time-scale convergence process, compares the convergence rates of various sampling schemes, and establishes a substantial advantage for random order over cyclic order of constraint selection. 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 current work proposes a more general framework that applies to convex (possibly nondifferentiable) optimization, and is based on the new coupled convergence theorem, which enhances our understanding of the two time-scale process and provides a modular architecture for analyzing new algorithms.

The rest of the paper is organized as follows. Section 2 summarizes our basic assumptions and a few preliminary results. Section 3 proves the coupled convergence theorem, which assuming a feasibility improvement condition and an optimality improvement condition, establishes the almost sure convergence of the randomized algorithm (6). 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 component functions such that the optimality improvement condition is satisfied. Section 6 collects various sets of conditions under which the almost sure convergence of the random incremental algorithms can be achieved. Section 7 discusses the rate of convergence of these algorithms and presents a computational example.

Our notation is summarized as follows. 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 $y_k = O(z_k)$ if there exists a constant $c > 0$ such that $y_k \leq cz_k$ for each k, and write $y_k = \Theta(z_k)$ if there exist constants $c_1 > c_2 > 0$ such that $c_2 z_k \leq y_k \leq c_1 z_k$ 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 "$\stackrel{a.s.}{\longrightarrow}$" means "converges almost surely to," while the abbreviation "i.i.d." means "independent identically distributed."

2 Assumptions and Preliminaries

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

$$x_{k+1} = \Pi\big[x_k - \alpha_k \tilde{\nabla} f(x_k)\big], \tag{8}$$

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

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \big\|\Pi\big[x_k - \alpha_k \tilde{\nabla} f(x_k)\big] - x^*\big\|^2 \\ &\leq \big\|\big(x_k - \alpha_k \tilde{\nabla} f(x_k)\big) - x^*\big\|^2 \\ &= \|x_k - x^*\|^2 - 2\alpha_k \tilde{\nabla} f(x_k)'(x_k - x^*) + \alpha_k^2 \big\|\tilde{\nabla} f(x_k)\big\|^2 \\ &\leq \|x_k - x^*\|^2 - 2\alpha_k \big(f(x_k) - f^*\big) + \alpha_k^2 L^2, \end{aligned} \tag{9}$$

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 Eq. (9) 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 incremental constraint projection-proximal algorithm, restated for convenience here,

$$z_k=x_k-\alpha_k g(\bar{x}_k,v_k),\qquad x_{k+1}=z_k-\beta_k\left(z_k-\Pi_{w_k}z_k\right),\quad \text{with}\ \ \bar{x}_k=x_k\ \text{or}\ \bar{x}_k=x_{k+1},\tag{10}$$

differs from the classical method (8) in a fundamental way: the iterates $\{x_k\}$ generated by the algorithm (10) 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 analog of the fundamental bound (9) now includes the distance of x_k from X, which need not decrease at each iteration. We will show that the incremental projection algorithm guarantees 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 algorithm (10), 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 the algorithm (10) with i.i.d. random projection and $\bar{x}_k=x_k$. Similar to the classical projection method (8), our line of analysis starts with a bound of the iteration error that has the form

$$\|x_{k+1}-x^*\|^2\le\|x_k-x^*\|^2-2\alpha_k\tilde{\nabla}f(x_k)'(x_k-x^*)+e(x_k,\alpha_k,\beta_k,w_k,v_k),\tag{11}$$

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 Eq. (11) and then take conditional expectation on both sides. From this we will obtain that the iteration error is "stochastically decreasing" in the following sense

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

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

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

Finally, based on the preceding two inequalities and through a series of intermediate results, we will end up using the following supermartingale convergence theorem due to Robbins and Siegmund [RS71] to prove an extension, a two-coupled-sequence supermartingale convergence lemma, and then complete the convergence proof of our algorithm.

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

$$\mathbf{E}\,[\xi_{k+1} \mid \mathcal{G}_k] \leq (1+\eta_k)\xi_k - u_k + \mu_k, \qquad \text{for all } \; k \geq 0 \;\; w.p.1,$$

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 $\tilde{\nabla} f(x) \in \partial f(x)$,

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

(b)

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

(c)

$$\mathbf{E}\left[\big\|g(x, v_k)\big\|^2 \;\big|\; \mathcal{F}_k\right] \leq L^2\big(\|x\|^2 + 1\big), \qquad \forall\; x \in \Re^n, \quad w.p.1. \tag{12}$$

Assumption 1 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.

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 to be either 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 1 *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 $x, y \in \Re^n$,

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

Proof. (a) We have

$$\begin{aligned}\|x - \beta(x - \Pi_S x) - y\|^2 &= \|x - y\|^2 + \beta^2\|x - \Pi_S x\|^2 - 2\beta(x-y)'(x - \Pi_S x)\\ &\le \|x - y\|^2 + \beta^2\|x - \Pi_S x\|^2 - 2\beta(x - \Pi_S x)'(x - \Pi_S x)\\ &= \|x - y\|^2 - \beta(2-\beta)\|x - \Pi_S x\|^2,\end{aligned}$$

where the inequality follows from $(y - \Pi_S x)'(x - \Pi_S x) \le 0$, the characteristic property of projection.
(b) We have

$$y - \Pi_S y = (x - \Pi_S x) + (y - x) - (\Pi_S y - \Pi_S x).$$

By using the triangle inequality and the nonexpansiveness of Π_S we obtain

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

Finally we complete the proof by using the inequality $(a+b)^2 \le 2a^2 + 2b^2$ for $a, b \in \Re$. ■

The second lemma gives a decomposition of the iteration error [cf. Eq. (11)], which will serve as the starting point of our analysis.

Lemma 2 *For any* $\epsilon > 0$ *and* $y \in X$, *the sequence* $\{x_k\}$ *generated by iteration* (10) *is such that*

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

Proof. From Lemma 1(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. Eq. (10)], we obtain

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

where the last inequality uses the fact $2a'b \leq \epsilon\|a\|^2 + (1/\epsilon)\|b\|^2$ 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 3 *Let Assumptions 1 and 2 hold, let x^* be a given optimal solution of problem (1), and let $\{x_k\}$ be generated by iteration (10). Then for all $k \geq 0$, with probability 1,*

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

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

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

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

Proof. We will prove parts (c) and (d) first, and prove parts (a) and (b) later.
(c,d) By using the basic inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for $a, b \in \Re^n$ and then applying Assumption 1, we have

$$\begin{aligned}\mathbf{E}\left[\|g(\bar{x}_k, v_k)\|^2 \mid \mathcal{F}_k\right] &\leq 2\mathbf{E}\left[\|g(x_k, v_k)\|^2 \mid \mathcal{F}_k\right] + 2\mathbf{E}\left[\|g(\bar{x}_k, v_k) - g(x_k, v_k)\|^2 \mid \mathcal{F}_k\right] \\ &\leq O\big(\|x_k - x^*\|^2 + 1\big) + O\left(\mathbf{E}\left[\|\bar{x}_k - x_k\|^2 \mid \mathcal{F}_k\right]\right).\end{aligned} \tag{13}$$

Since $\bar{x}_k \in \{x_k, x_{k+1}\}$ and $X \subset X_{w_k}$, we use the definition (10) of the algorithm and obtain

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

so that

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

Note that from Lemma 1(b) we have

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

Then it follows from the preceding two relations that

$$\|\bar{x}_k - x_k\|^2 \leq \|x_{k+1} - x_k\|^2 \leq O(\alpha_k^2)\|g(\bar{x}_k, v_k)\|^2 + O(\beta_k^2)\,\mathrm{d}^2(x_k). \tag{14}$$

By taking expectation on both sides of Eq. (14) and applying Eq. (13), we obtain

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

and by rearranging terms in the preceding inequality, we obtain part (d). Finally, we apply part (d) to Eq. (13) and obtain

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

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

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

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

The next lemma is an extension of Lemma 3. 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 4 *Let Assumptions 1 and 2 hold, let x^* be a given optimal solution of problem (1), let $\{x_k\}$ be generated by iteration (10), and let N be a given positive integer. Then for all $k \geq 0$, with probability 1:*

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

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

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

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

Proof. (a) The case where $N = 1$ has been given in Lemma 3(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 3(a) and the fact $\alpha_{k+1} \leq \alpha_k$. 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 3(b). In the case where $N = 2$, we have

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

where the first equality uses iterated expectation, the second inequality uses Lemma 3(b), and third inequality use Lemma 3(a),(b) and the fact $\alpha_{k+1} \leq \alpha_k$. In the case where $N > 2$, the result follows by applying the preceding argument inductively.

(c) Follows by applying Lemma 3(c) and part (a):

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

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

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

where the first inequality applies Lemma 3(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

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

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

Lemma 5 *Let Assumptions 1 and 2 hold, let x^* be a given optimal solution of problem (1), let $\{x_k\}$ be generated by iteration (10), and let N be a given positive integer. Then for all $k \ge 0$, with probability 1:*

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

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

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

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

$$f(x_k)-f(x_{k+N}) \le -\tilde{\nabla}f(x_k)'(x_{k+N}-x_k) \le \left\|\tilde{\nabla}f(x_k)\right\|\|x_{k+N}-x_k\| \le \frac{\alpha_k}{2}\|\tilde{\nabla}f(x_k)\|^2+\frac{2}{\alpha_k}\|x_{k+N}-x_k\|^2.$$

Taking expectation on both sides, using Assumption 1 and using Lemma 4(d), we obtain

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

(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

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

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 generic algorithm that alternates between an iteration of random optimality update and an iteration of random feasibility update, i.e.,

$$z_k = x_k - \alpha_k g(\bar{x}_k, v_k), \qquad x_{k+1} = z_k - \beta_k \left(z_k - \Pi_{w_k} z_k\right), \quad \text{with } \bar{x}_k = x_k, \text{ or } \bar{x}_k = x_{k+1} \tag{15}$$

[cf. Eqs (6), (10)], without specifying details regarding how the random variables w_k and v_k are generated. We will 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. This is a key result of the paper and is stated as follows.

Proposition 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 (15). Assume that there exist positive integers M, N such that:*

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

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

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

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

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 proposition we provide some discussion. Let us first note that in the preceding proposition, x^* is an arbitrary but fixed optimal solution, and that the $O(\cdot)$ 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 Prop. 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 (15), 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.

In what follows, we will prove a preliminary result that is important 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 Prop. 1.

Lemma 6 (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*

$$\mathbf{E}\left[\xi_{t+1} \mid \mathcal{G}_k\right] \leq (1+\eta_t)\xi_t - u_t + c\theta_t\zeta_t + \mu_t,$$

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

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 $\left\{\mathbf{E}\left[\xi_t\right]\right\}$ and $\left\{\mathbf{E}\left[\zeta_t\right]\right\}$ 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

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

It follows from the definition of J_t that

$$\mathbf{E}\left[J_{t+1} \mid \mathcal{G}_k\right] \leq (1+\eta_t+c\epsilon_t)J_t - (u_t + c\bar{u}_t) + (\mu_t + c\nu_t) \leq (1+\eta_t+c\epsilon_t)J_t + (\mu_t + c\nu_t). \tag{16}$$

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 Theorem (Theorem 1) applies to Eq. (16). 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 $\xi_t \leq J_t$ and $\zeta_t \leq \frac{1}{c}J_t$. 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

$$\sum_{t=0}^{\infty} \epsilon_t\xi_t \leq \left(\sum_{t=0}^{\infty} \epsilon_t\right)\left(\sup_{t\geq 0} \xi_t\right) < \infty, \qquad w.p.1. \tag{17}$$

From Eq. (17), we see that the Supermartingale Convergence Theorem 1 also applies to the given inequality

$$\mathbf{E}\left[\zeta_{t+1} \mid \mathcal{G}_k\right] \leq (1-\theta_t)\zeta_t - \bar{u}_t + \epsilon_t\xi_t + \nu_t \leq (1-\theta_t)\zeta_t + \epsilon_t\xi_t + \nu_t. \tag{18}$$

Therefore ζ_t converges almost surely to a random variable, and

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

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 Eq. (16) and obtain

$$\mathbf{E}\left[J_{t+1}\right] \leq (1 + \eta_t + c\epsilon_t)\mathbf{E}\left[J_t\right] + (\mu_t + c\nu_t). \tag{19}$$

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 theorem 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 Eq. (18), we obtain

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

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

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

Therefore

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

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 6 to prove the results of Prop. 1. However, two issues remain to be addressed. First, the two improvement conditions of Prop. 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 6. 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.

Our proof consists of four steps, and its main idea is to construct a meta-cycle of $M \times N$ iterations, where the t-th 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.

Proof of the Coupled Convergence Theorem (Prop. 1).

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

$$\begin{aligned}\mathbf{E}\left[\|x_{(t+1)MN} - x^*\|^2 \mid \mathcal{F}_{tMN}\right] \leq \|x_{tMN} - x^*\|^2 - 2\sum_{\ell=tM}^{(t+1)M-1}\left(\sum_{k=\ell N}^{(\ell+1)N-1}\alpha_k\right)\Big(\mathbf{E}\left[f(x_{\ell N}) \mid \mathcal{F}_{tMN}\right] - f^*\Big) \\ + \sum_{\ell=tM}^{(t+1)M-1} O(\alpha_{\ell N}^2)\Big(\mathbf{E}\left[\|x_{\ell N} - x^*\|^2 \mid \mathcal{F}_{tMN}\right] + 1\Big) \\ + \sum_{\ell=tM}^{(t+1)M-1} O(\beta_{\ell N}^2)\mathbf{E}\left[\,\mathrm{d}^2(x_{\ell N}) \mid \mathcal{F}_{tMN}\right], \qquad w.p.1.\end{aligned} \tag{20}$$

From Lemma 4(a) and the nonincreasing property of $\{\alpha_k\}$ we obtain the bound

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

From Lemma 4(b) and the nonincreasing property of $\{\beta_k\}$ we obtain the bound

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

By using Lemma 5(c) we further obtain

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

We apply the preceding bounds to Eq. (20), and remove redundant scalars in the big $O(\cdot)$ terms, yielding

$$\begin{aligned}\mathbf{E}\left[\|x_{(t+1)MN} - x^*\|^2 \mid \mathcal{F}_{tMN}\right] \leq \|x_{tMN} - x^*\|^2 - 2\left(\sum_{k=tMN}^{(t+1)MN-1}\alpha_k\right)\left(f(\Pi x_{tMN}) - f^*\right) \\ + O\left(\frac{\alpha_{tMN}^2}{\beta_{tMN}}\right)\left(\|x_{tMN} - x^*\|^2 + 1\right) + O\left(\beta_{tMN}\right)\mathrm{d}^2(x_{tMN}),\end{aligned} \tag{21}$$

for all $t \geq 0$, with probability 1. Note that the term $f(\Pi x_k) - f^*$ is nonnegative. This will allow us to treat Eq. (21) as one of the conditions of Lemma 6.

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

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

with probability 1. Then by using Lemma 4(a) to bound the terms $\mathbf{E}\left[\|x_{\ell M} - x^*\|^2 \mid \mathcal{F}_{tMN}\right]$, we obtain

$$\mathbf{E}\left[\,\mathrm{d}^2(x_{(t+1)MN}) \mid \mathcal{F}_{tMN}\right] \leq \big(1 - \Theta(\beta_{tMN})\big)\,\mathrm{d}^2(x_{tMN}) + O\left(\sum_{k=tMN}^{(t+1)MN-1}\frac{\alpha_k^2}{\beta_k}\right)\left(\|x_{tMN} - x^*\|^2 + 1\right), \tag{22}$$

with probability 1.

Step 3 (Apply the Coupled Supermartingale Convergence Lemma) Let $\epsilon_t = O\left(\sum_{k=tMN}^{(t+1)MN-1} \frac{\alpha_k^2}{\beta_k}\right)$, so we have

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

Therefore the coupleds supermartingale convergence lemma (cf. Lemma 6) applies to inequalities (21) and (22). It follows that $\|x_{tMN} - x^*\|^2$ and $\mathrm{d}^2(x_{tMN})$ converge almost surely,

$$\sum_{t=0}^{\infty} \Theta(\beta_{tMN})\, \mathrm{d}^2(x_{tMN}) < \infty, \qquad w.p.1, \tag{23}$$

and

$$\sum_{t=0}^{\infty} \left(\sum_{k=tMN}^{(t+1)MN-1} \alpha_k \right) \big(f(\Pi x_{tMN}) - f^*\big) < \infty, \qquad w.p.1. \tag{24}$$

Moreover, from the last part of Lemma 6, it follows that the sequence $\{\mathbf{E}\left[\|x_{tMN} - x^*\|^2\right]\}$ is bounded, and we have

$$\sum_{t=0}^{\infty} \Theta(\beta_{tMN}^2) \mathbf{E}\left[\, \mathrm{d}^2(x_{tMN})\right] < \infty. \tag{25}$$

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 (23) 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 (24) together with the assumption $\sum_{k=0}^{\infty} \alpha_k = \infty$ implies that

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

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 4(c), and the boundedness of $\left\{\mathbf{E}\left[\|x_{tMN} - x^*\|^2\right]\right\}$ (as shown in Step 3), we obtain

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

Similarly, by using the Markov inequality, Lemma 4(b), and Eq. (25), we obtain

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

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

$$\begin{aligned}
\max_{tMN \leq k \leq (t+1)MN-1} \|x_k - x_{tMN}\| &\leq \sum_{\ell=tMN}^{(t+1)MN-1} \|x_\ell - x_{\ell+1}\| \quad \text{(from the triangle inequality)} \\
&\leq \sum_{\ell=tMN}^{(t+1)MN-1} O\big(\alpha_\ell \|g(\bar{x}_\ell, v_\ell)\| + \beta_\ell \, \mathrm{d}(x_\ell)\big) \quad \text{(from Eq. (14))} \\
&\xrightarrow{a.s.} 0.
\end{aligned}$$

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 $k \to \infty$. ■

4 Sampling Schemes for Constraints

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

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

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 has been 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 $x \in \Re^n$*

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

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. 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, and indicates that this condition is a mild restriction in practice.

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*

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

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

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

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 (Prop. 1).

Proposition 2 *Let Assumptions 1, 2, 3 and 4 hold, and let x^* be a given optimal solution of problem* (1)*. Then the random projection algorithm* (15) *generates a sequence $\{x_k\}$ such that*

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

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

Proof. Let ϵ be a positive scalar. By applying Lemma 2 with $y = \Pi x_k$, we have

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

By using the following bound which is obtained from Lemma 1(b):

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

we further obtain

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

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 3(c) and Eq. (26), we obtain

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

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

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.,*

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

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

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

Then by using an analysis similar to that of Prop. 2, we obtain the following result.

Proposition 3 *Let Assumptions 1, 2, 3 and 5 hold, and let x^* be a given optimal solution of problem (1). Then the random projection algorithm (15) generates a sequence $\{x_k\}$ such that*

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

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

Proof. The proof is almost identical to that of Prop. 2, except that we use Eq. (27) in place of Eq. (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 t-th cycle, i.e.,*

$$\{X_{w_k}\}, \qquad \text{where } k = tm, tm+1, \ldots, (t+1)m - 1,$$

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 towards feasibility over a number of iterations.

Lemma 7 *Let Assumptions 1, 2, and 3 hold, and let $\{x_k\}$ be generated by algorithm (15). 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:*

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

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

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

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

$$\begin{aligned}
\frac{1}{\sqrt{\eta}}\,\mathrm{d}(x_k) &\leq \|x_k - \Pi_{w_{k^*}} x_k\| \\
&\leq \|x_k - \Pi_{w_{k^*}} z_{k^*}\| \quad \text{(by the definition of } \Pi_{w_{k^*}} x_k \text{ and the fact } \Pi_{w_{k^*}} z_{k^*} \in X_{w_{k^*}}) \\
&= \left\| x_k - \frac{1}{\beta_{k^*}} x_{k^*+1} + \frac{1-\beta_{k^*}}{\beta_{k^*}} z_{k^*} \right\| \quad \text{(by } x_{k^*+1} = z_{k^*} - \beta_{k^*}(z_{k^*} - \Pi_{w_{k^*}} z_{k^*}),\ \text{cf. Eq.(15))} \\
&= \left\| \sum_{\ell=k}^{k^*-1} \frac{\beta_\ell - 1}{\beta_\ell} (z_\ell - x_{\ell+1}) + \sum_{\ell=k}^{k^*} \frac{1}{\beta_\ell} (z_\ell - x_{\ell+1}) - \sum_{\ell=k}^{k^*} (z_\ell - x_\ell) \right\| \\
&\leq \sum_{\ell=k}^{k^*-1} \left| \frac{\beta_\ell - 1}{\beta_\ell} \right| \|z_\ell - x_{\ell+1}\| + \sum_{\ell=k}^{k^*} \frac{1}{\beta_\ell} \|z_\ell - x_{\ell+1}\| + \sum_{\ell=k}^{k^*} \|z_\ell - x_\ell\| \\
&\leq \sum_{\ell=k}^{k+M-2} \left| \frac{\beta_\ell - 1}{\beta_\ell} \right| \|z_\ell - x_{\ell+1}\| + \sum_{\ell=k}^{k+M-1} \frac{1}{\beta_\ell} \|z_\ell - x_{\ell+1}\| + \sum_{\ell=k}^{k+M-1} \|z_\ell - x_\ell\| \\
&\leq \sum_{\ell=k}^{k+M-1} \frac{2}{\beta_\ell} \|z_\ell - x_{\ell+1}\| + \sum_{\ell=k}^{k+M-1} \|z_\ell - x_\ell\| \quad \text{(since } \beta_\ell \in (0,2)) \\
&= 2 \sum_{\ell=k}^{k+M-1} \|z_\ell - \Pi_{w_\ell} z_\ell\| + \sum_{\ell=k}^{k+M-1} \alpha_\ell \big\|g(\bar{x}_\ell, v_\ell)\big\| \quad \text{(by the definition of algorithm (15))} \\
&\leq \sqrt{2M} \left(4 \sum_{\ell=k}^{k+M-1} \|z_\ell - \Pi_{w_\ell} z_\ell\|^2 + \sum_{\ell=k}^{k+M-1} \alpha_\ell^2 \big\|g(\bar{x}_\ell, v_\ell)\big\|^2 \right)^{1/2},
\end{aligned}$$

where the last step follows from the generic inequality $\left(\sum_{i=1}^M a_i + \sum_{i=1}^M b_i\right)^2 \leq 2M \left(\sum_{i=1}^M a_i^2 + \sum_{i=1}^M b_i^2\right)$ 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 4 *Let Assumptions 1, 2, 3 and 6 hold, and let* x^* *be a given optimal solution of problem*

(1). *Then the random projection algorithm* (15) *generates a sequence* $\{x_k\}$ *such that*

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

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

Proof. Let $\epsilon > 0$ be a scalar. By applying Lemma 2 with $y = \Pi x_k$, we have

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

By applying the preceding relation inductively, we obtain

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

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

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

We apply Lemma 7 with $M = 2m$ (since according to Assumption 6, starting with any k, any particular index will be visited in at most 2 cycles of samples), and obtain

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

Let $\epsilon \ll \frac{1}{m\eta}O(\beta_k)$. Taking conditional expectation on both sides and applying Lemma 3(c), we have

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

for all $k \geq 0$ 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 Prop. 4, we obtain the following result.

Proposition 5 *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 the random projection algorithm (15). Then there exists a positive integer M such that*

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

for all $k \geq 0$, 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 $\varepsilon > 0$

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

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

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

We let M be a sufficiently large integer, such that

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

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

By using this fact together with Lemma 7, we obtain

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

It follows that

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

By rewriting the preceding relation and applying Lemma 4(a), we obtain

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

The rest of the proof follows a line of analysis like the one of Prop. 4, with $2m$ replaced with M. Similar to Eq. (29), we have

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

Taking expectation on both sides, we obtain

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

where the second relation uses Eq. (31) and Lemma 4(c), and the third relation holds by letting $\epsilon \leq \Theta\left(\frac{\beta_k}{M\eta}\right)$. ■

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 (Prop. 1), i.e.,

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

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

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

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

Proposition 6 *Let Assumptions 1, 2, 3 and 8 hold, and let x^* be a given optimal solution of problem*

(1). *Then the random projection algorithm* (15) *generates a sequence* $\{x_k\}$ *such that*

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

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

Proof. By applying Lemma 2 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 3(c) yields

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

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

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

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 3(c),(d). Finally, we apply the preceding relation to Eq. (34) and obtain

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

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 $N > 0$,

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

and the stepsizes $\{a_k\}$ are constant within each cycle, i.e,

$$\alpha_{tN} = \alpha_{tN+1} = \cdots = \alpha_{(t+1)N-1}, \qquad \forall\, t \geq 0.$$

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 7 *Let Assumptions 1, 2, 3 and 9 hold, and let x^* be a given optimal solution of problem (1). Then the random projection algorithm (15) generates a sequence $\{x_k\}$ such that*

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

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

Proof. Following the line of analysis of Prop. 6 and applying Eq. (33) repeatedly, we obtain

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

By taking conditional expectation on both sides and by applying Lemma 4(c), we further obtain

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

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

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

Combining the preceding two relations, we obtain

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

By taking expectation on both sides, we further obtain

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

where the second inequality uses the basic fact $2a'b \leq \|a\|^2 + \|b\|^2$ for $a, b \in \Re^n$, and the last inequality uses Assumption 1 and Lemma 4(a),(d). Then from Assumption 9 we have

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

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

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

6 Almost Sure Convergence of Incremental 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 theorem (Prop. 1) are satisfied. Now we will combine the preceding results and apply the coupled convergence theorem. The following theorem collects various combinations of conditions under which our algorithm converges almost surely.

Proposition 8 (Almost Sure Convergence) *Let Assumptions 1, 2, 3 hold, and consider the incremental constraint projection-proximal algorithm* (15)*. 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 any one 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 the algorithm (15) *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)*.*

Proof. The proof is obtained by combining Props. 2, 3, 4, 5 and Props. 6, 7, in conjunction with Prop. 1. ■

7 Discussion and Computational Results

An important issue related to the proposed incremental algorithms is the rate of convergence. As noted earlier, the convergence of these algorithms involve two improvement processes with two different corresponding stepsizes. This coupling greatly complicates the convergence rate analysis. Moreover, the convergence rate also relates to the sampling schemes, properties of the objective function, properties of the constraints, choices of the stepsizes, etc, which complicates the analysis further. 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]. These results involve the strong convexity constant, and suggest an advantage of random sampling over cyclic sampling. Reference [WB12] and the thesis [Wan13] also provide computational results.

For minimization of general convex functions, theoretical analysis of convergence rate is not currently available, except in special cases which involves no constraint sampling, no stochastic subgradients, and/or more restrictive assumptions (see [NB00], [NB01], [Ber11], [Ned11], [WB12]). By extrapolating known results from the strongly convex case, we conjecture that the algorithm with random sampling has better worst-case performance than the one with cyclic sampling. The likely reason is that random sampling may break an unfavorable order of component functions/constraints that may slow down the convergence. Moreover, we conjecture that by sampling adaptively, e.g. choosing the most distant constraint, the algorithm achieves a better convergence rate than by sampling non-adaptively. Our subsequent computational results support the preceding conjectures.

We tested our algorithms on a regression problem involving ℓ_1 regularization, nonnegativity constraints, and basis function approximation. The problem is

$$\begin{aligned} &\min_x \|A\Phi x - b\|^2 + \lambda\|x\|_1 \\ &\text{s.t. } \Phi x \geq 0, \end{aligned} \tag{37}$$

where A is an 1000×1000 matrix, b is a vector in $\Re^{1000}$, Φ is an 1000×20 matrix of basis functions/features, and λ is a positive regularization parameter. This problem has a convex nondifferentiable cost function and a set intersection constraint. The gradient of the quadratic term $\|A\Phi x - b\|^2$ can be written as

$$\Phi' A' A \Phi x - \Phi' A' b = \left(\sum_{i=1}^{1000} \sum_{j=1}^{1000} \sum_{q=1}^{1000} a_{qi} a_{qj} \phi_i \phi_j' \right) x - \left(\sum_{i=1}^{1000} \sum_{j=1}^{1000} \phi_i a_{ji} b_j \right),$$

where a_{ij} and b_i are the corresponding entries of A and b, and ϕ_i' is the ith row of Φ. The constraint $\Phi x \geq 0$ can be viewed as an intersection of halfspaces defined by $\phi_i' x \geq 0$. Applying algorithm (15) to this problem, we obtain

$$\begin{aligned} z_k &= x_k - \alpha_k \left(a_{q_k i_k} a_{q_k j_k} \phi_{i_k} \phi_{j_k}' x_k - \phi_{i_k} a_{j_k i_k} b_{j_k} + \lambda \, s(x_k) \right) \\ x_{k+1} &= z_k - \beta_k \frac{\max\{0, \phi_{w_k}' z_k\}}{\|\phi_{w_k}\|^2} \phi_{w_k}, \end{aligned} \tag{38}$$

where $s(x)$ is a subgradient of $\|x\|_1$ [i.e., $s_i(x) = 1$ if $x_i > 0$, $s_i(x) = -1$ if $x_i \leq 0$], $v_k = (i_k, j_k, q_k)$ and w_k are generated by the component function and constraint sampling schemes, respectively. Note that each iteration of this algorithm is inexpensive and involves only low-order calculation. In the experiments, the columns of Φ were chosen to be sine functions of different frequencies, and the entries of A, b were independently generated according to a uniform distribution in $[-1, 1]$. We tested algorithm (38) with various sampling schemes for $\{w_k, v_k\}$, and we plot the associated trajectories of $\{\|x_k - x^*\|\}$ and $\{\, \mathrm{d}(x_k)\}$ in Figs. 1 and 2.

Figure 1 compares algorithms that use different constraint sampling schemes. It can be seen that the most distant projection scheme clearly outperforms the others. However, this comes at a price - choosing the most distant set incurs a computation overhead on the order of m. We note that the random sampling scheme performs slightly better than the deterministic cyclic sampling scheme, and that the Markov sampling scheme may perform the worst depending on the mixing rate of the Markov chain. Figure 2 compares the samplings schemes for the subgradients/component functions, and suggests that random sampling outperforms cyclic sampling. According to our analysis, the two improvement processes $\{\|x_k - x^*\|\}$ and $\{\, \mathrm{d}(x_k)\}$ are coupled together. Although $\{\, \mathrm{d}(x_k)\}$ has a better modulus of contraction, it is not clear from Figs. 1-2 that it converges faster than $\{\|x_k - x^*\|\}$ does. We have also tested the algorithms using different parameters and observed similar results. Theoretical analysis supporting these results is an interesting subject for future research.

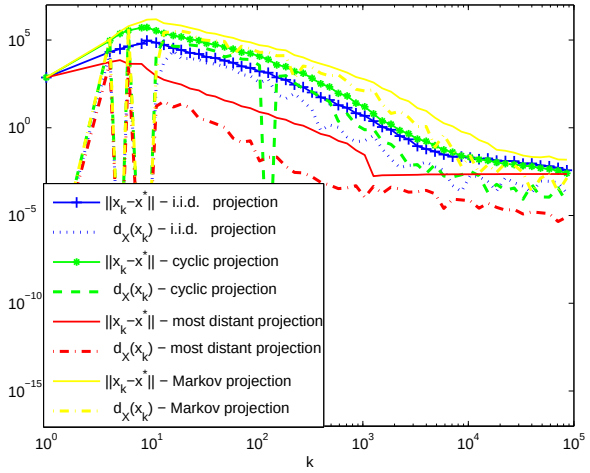

Figure 1: Comparison of constraint sampling schemes. In all cases, we use the exact subgradient of f without sampling, and we take $\alpha_k = 1/k$, $\beta_k = 1$, $\lambda = 0.001$. In the first case (blue), the constraints are sampled independently according to a uniform distribution. In the second case (green), the constraints are sampled according to a deterministic cyclic order. In the third case (red), the constraints are chosen according to the most distant set criterion. In the last case (yellow), the constraints are chosen according to state transitions of a Markov chain, in which the indexes/states stay unchanged with probability 0.1 and move to other states according to a uniform distribution with probability 0.9.

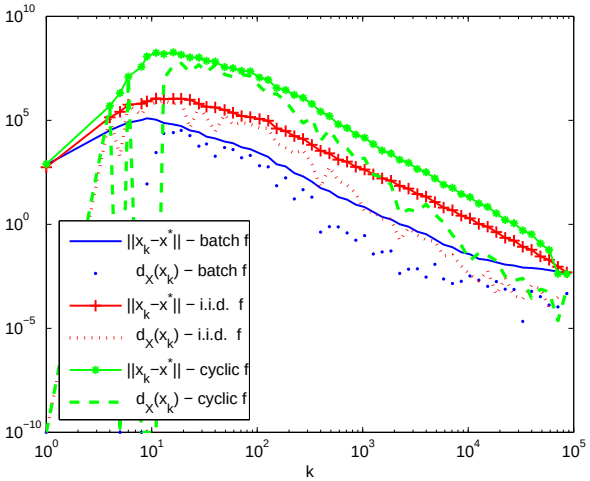

Figure 2: Comparison of various component function sampling schemes. In all cases, we use an i.i.d. uniform constraints sampling scheme, and we take $\alpha_k = 1/k$, $\beta_k = 1$, $\lambda = 0.001$. In the first case (blue), the algorithm uses the exact subgradients of f without sampling. In the second and third cases (red and green), the algorithm chooses samples independently according to a uniform distribution and cyclically according to a fixed order, respectively.

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 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.

An important direction for future research is to develop a convergence rate analysis, incorporate it into the general framework of coupled convergence, and compare the performances of various sampling/randomization schemes for the subgradients and the constraints. 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., [NBB01]). Another extension is to analyze problems with an infinite number of constraints.

References

[Bau96] H. H. Bauschke. Projection algorithms and monotone operators. *Ph.D. thesis, Simon Frazer University, Canada*, 1996.

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

[BBL97] H. Bauschke, J. M. Borwein, and A. S. Lewis. The method of cyclic projections for closed convex sets in Hilbert space. *Contemporary Mathematics*, 204:1–38, 1997.

[Ber11] D. P. Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming, Ser. B*, 129:163–195, 2011.

[Ber12] D. P. Bertsekas. A survey of incremental methods for minimizing a sum $\sum_{i=1}^{m} f_i(x)$, and their applications in inference/machine learning, signal processing, and large-scale and distributed optimization. *Optimization for Machine Learning*, pages 85–119, 2012. An extended version of the survey appeared in Report LIDS-P-2848, MIT, 2010.

[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, MA, 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. Optimization*, 19:1093–1106, 2008.

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

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

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

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

[Hal62] I. Halperin. The product of projection operators. *Acta Scientiarum Mathematicarum*, 23:96–99, 1962.

[KSHdM02] A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.*, 12:479–502, 2002.

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

[LL10] D. Leventhal and A. S. Lewis. Randomized methods for linear constraints: Convergence rates and conditioning. *Mathematics of Operations Research*, 35:641–654, 2010.

[LM08] A. S. Lewis and J. Malick. Alternating projections on manifolds. *Mathematics of Operations Research*, 33:216–234, 2008.

[NB00] A. Nedić and D. P. Bertsekas. Convergence rate of the incremental subgradient algorithm. *Stochastic Optimization: Algorithms and Applications, by S. Uryasev and P. M. Pardalos Eds.*, pages 263–304, 2000.

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

[NBB01] A. Nedić, D. P. Bertsekas, and V. S. Borkar. Distributed asynchronous incremental subgradient methods. *Studies in Computational Mathematics*, 8:381–407, 2001.

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

[Ned11] A. Nedić. Random algorithms for convex minimization problems. *Mathematical Programming, Ser. B*, 129:225–253, 2011.

[NJLS09] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. of Optimization*, 19:1574–1609, 2009.

[RS71] H. Robbins and D. O. Siegmund. A convergence theorem for nonnegative almost supermartingales and some applications. *Optimizing Methods in Statistics*, pages 233–257, 1971.

[SDR09] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, Philadelphia, PA, 2009.

[Tse90] P. Tseng. Successive projection under a quasi-cyclic order. *Lab. for Information and Decision Systems Report LIDS-P-1938, MIT, Cambridge, MA*, 1990.

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

[Wan13] Mengdi Wang. *Stochastic Algorithms for Large-Scale Linear Problems, Variational Inequalities and Convex Optimization*. PhD thesis, MIT, 2013.

[WB12] M. Wang and D. P. Bertsekas. Incremental constraint projection methods for variational inequalities. *Lab. for Information and Decision Systems Report*, LIDS-P-2898, MIT, Cambridge, MA, 2012.