Incremental constraint projection methods for variational inequalities

Citation: Wang, Mengdi, and Dimitri P. Bertsekas. "Incremental Constraint Projection Methods for Variational Inequalities." Math. Program. 150, no. 2 (May 17, 2014): 321–363.

As Published: http://dx.doi.org/10.1007/s10107-014-0769-x

Publisher: Springer-Verlag

Persistent URL: http://hdl.handle.net/1721.1/99757

Version: Original manuscript: author's manuscript prior to formal peer review

Massachusetts Institute of Technology, Cambridge, MA
Laboratory for Information and Decision Systems
Report LIDS-P-2898, December 2012

Incremental Constraint Projection Methods for Variational Inequalities

Mengdi Wang
mdwang@mit.edu

Dimitri P. Bertsekas*
dimitrib@mit.edu

Abstract

We consider the solution of strongly monotone variational inequalities of the form $F(x^*)'(x-x^*) \geq 0$, for all $x \in X$. We focus on special structures that lend themselves to sampling, such as when X is the intersection of a large number of sets, and/or F is an expected value or is the sum of a large number of component functions. We propose new methods that combine elements of incremental constraint projection and stochastic gradient. We analyze the convergence and the rate of convergence of these methods with various types of sampling schemes, and we establish a substantial rate of convergence advantage for random sampling over cyclic sampling.

Key words: random projection, alternate/cyclic projection, variational inequalities, stochastic gradient, incremental method, sampling, stochastic approximation.

1 Introduction

Variational inequalities (VI) is a general class of problems, which under appropriate assumptions, include as special cases several fundamental problems in applied mathematics and operations research, such as convex differentiable optimization, solution of systems of equations and their approximation by Galerkin approximation or aggregation, saddle point problems, and equilibrium problems. They take the form

$$F(x^*)'(x-x^*) \geq 0, \qquad \forall\, x \in X, \tag{1}$$

where $F : \Re^n \mapsto \Re^n$ is a mapping, and X is a closed and convex set in $\Re^n$. For extensive background on VI, we refer to the books by Kinderlehrer and Stampacchia [KiS80], by Patriksson [Pat99], and by Facchinei and Pang [FaP03]. These books contain theoretical analysis as well as a wide range of algorithms and applications.

We are interested in a VI of the form (1) in which the constraint set X is the intersection of many sets, i.e.,

$$X = \cap_{i \in M} X_i,$$

with each X_i being a closed and convex subset of $\Re^n$, and M being the set of constraint indexes. Moreover we allow the function F to have the form of an expected value, or a sum of a large number of component functions. We assume throughout that the mapping F is strongly monotone, so the VI has a unique solution x^* (see e.g., [FaP03]). We will later introduce additional assumptions, including the condition that F is Lipschitz continuous.

The classical projection method for solution of a VI (and also for convex optimization when F is the gradient of a convex function) has the form

$$x_{k+1} = \Pi\big[x_k - \alpha_k F(x_k)\big], \tag{2}$$

*Mengdi Wang and Dimitri Bertsekas are 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, and $\{\alpha_k\}$ is a sequence of constant or diminishing positive scalars. The projection exists and is unique since X is closed and convex. It is well known that if F is strongly monotone, this method converges to the unique solution x^* if all α_k lie within a sufficiently small interval $(0, \bar{\alpha})$ and $\sum_{k=0}^{\infty} \alpha_k = \infty$, as first shown by Sibony [Sib70].

A major difficulty when using this method in practice is the computation of the projection at each iteration, which can be time-consuming. In the case where the constraint set X is the intersection of a large number of simpler sets X_i, it is possible to exploit this structure and improve the method, by projecting onto a single set X_i at each iteration, in the spirit of random and cyclic projection methods that are widely used to solve the feasibility problem of finding some point in X. This suggests the following modification of the algorithm (2):

$$x_{k+1} = \Pi_{w_k}\big[x_k - \alpha_k F(x_k)\big], \tag{3}$$

where we denote by Π_{w_k} the Euclidean orthogonal projection onto X_{w_k}, and $\{w_k\}$ is a sequence of random variables taking values in M. An interesting special case is when X is a polyhedral set, i.e., the intersection of a finite number of halfspaces. Then the algorithm involves successive projections onto halfspaces, which are easy to implement and computationally inexpensive.

A second important difficulty arises when F is either the sum of a large number of component functions, or more generally can be expressed as the expected value

$$F(x) = \mathbf{E}\big[f(x, v)\big], \tag{4}$$

where f is some function of x and a random variable v. Then the exact computation of $F(x_k)$ can be either very time-consuming or impossible due to some noise. To address this additional difficulty, we may use in place of $F(x_k)$ in Eq. (3) a stochastic sample $f(x_k, v_k)$. This motivates the *incremental constraint projection algorithm*

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

where $\{v_k\}$ and $\{w_k\}$ are sequences of random variables generated by some probabilistic process, and $\{\alpha_k\}$ and $\{\beta_k\}$ are sequences of positive scalar stepsizes. For convergence to x^*, we will assume that α_k is diminishing and β_k is constant or slowly diminishing (precise conditions will be given later).

The purpose of this paper is to analyze the convergence and rate of convergence properties of the algorithm (5). We focus primarily on the case where the number of constraint sets is finite, i.e, $M = \{1, \ldots, m\}$, where m is a positive integer. However, a large portion of our analysis can be adapted to allow an infinite number of constrain sets. To the best of our knowledge the algorithm and its analysis are new: there seems to be no prior literature on projection methods for VI that involve feasibility updates using projection on component supersets X_i of the constraint set X.

The convergence mechanism of our algorithm involves an intricate interplay between the progress of the constraint projection steps and the function projection steps, and their associated stepsizes β_k and α_k. An important new insight that emerges from our analysis is that the algorithm operates on two different time scales: the convergence to the feasible set, which is controlled by β_k, is faster than the convergence to the optimal solution, which is controlled by α_k. Thus, *asymptotically, the method operates nearly as if the projections are done on the entire set X.* This two-time-scale 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 method of selection of the samples v_k and w_k. We will consider the two cases where:

- The samples v_k and w_k are generated randomly, so that all the indexes are sampled sufficiently often. We refer to this as the *random projection algorithm.*

- The samples v_k and w_k are generated "cyclically," e.g., according to either a deterministic cyclic order or a random permutation of the component indexes within a cycle (a precise definition will be given later). We refer to this as the *cyclic projection algorithm.*

For versions of the algorithm with non-diminishing stepsizes α_k and β_k, we show that $\{x_k\}$ converges to within an appropriate neighborhood of x^*. In addition, we develop convergence rate estimates for the number of iterations needed for the algorithm to converge to within a given error tolerance. Our comparison of the rates of convergence of the random and the cyclic sampling cases indicates an advantage for random sampling. This has also been confirmed by computational experimentation, and is consistent with earlier results on incremental subgradient methods [NeB01], [BNO03], [Ber10].

Our proposed algorithm (5) is related to a number of known methods from convex optimization, feasibility, VIs, and stochastic approximation. In particular, when $\beta_k = 1$ and $F(x_k) = 0$ for all k in iteration (3) we obtain a successive projection algorithm for finding some $x \in X = \cap_{i\in M} X_i$, of the type proposed and analyzed in many sources. In the case where $F(x_k)$ is the gradient at x_k of a strongly convex function f, possibly of the additive form $f = \sum_{i=1}^m f_i$, we obtain special cases of recently proposed algorithms for minimizing f over $x \in X = \cap_{i\in M} X_i$ (see the following discussion). Finally, in the case where $X = X_{w_k}$ for all w_k and F is given as an expected value [cf. Eq. (4)], our method becomes a stochastic approximation method for VIs, which has been well known in the literature.

In view of the connections just noted, our analysis uses several ideas from the literature on projection, feasibility, incremental/stochastic gradient, and stochastic approximation methods, which we will now summarize. The projection method for numerical solution of strongly monotone VIs has been studied extensively (see e.g., Bertsekas and Tsitsiklis [BeT89], and Facchinei and Pang [FaP03] for textbook accounts of its properties and convergence analysis). A survey on deterministic projection-type methods is given by Xiu and Zhang [XiZ03]. Some recent developments have considered a stochastic framework and used a projection-type stochastic approximation method (see for example Gürkan et al. [GOR99], and Jiang and Xu [JiX08]). The recent works by Kannan and Shanbhag [KaS13] and by Kannan et al. [KNS12] have considered an iterative regularization method and an iterative proximal point method for (stochastic) variational inequalities that are not necessarily strongly monotone, where the former uses a diminishing regularization term and an exact constraint projection step at each iteration, and the latter applies iterative projection steps towards the proximal problem with changing centers. The papers by Fukushima [Fuk86] and more recently Censor and Gibali [CeG08] have considered methods that utilize outer approximations of X by deterministic projection onto a specially selected halfspace separating X and the iterate. These methods share the motivation of constraint relaxation with the proposed algorithms of the current work. However, the assumptions, algorithmic details, applications and convergence mechanisms of the methods differ fundamentally from each other. Other works in the area include finding common solutions to VIs (see for example Censor et al. [CGRS12a], [CGRS12b]), and general VIs (see the survey by Noor [Noo04] and the citations there).

The feasibility problem of finding a point with certain properties within a set intersection $\cap_{i\in M} X_i$ arises in many contexts. For the case where $M = \{1, \ldots, m\}$ with m being a large number and each of the sets 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 [vNe50], and followed by many other authors Halperin [Hal62], Gubin et al. [GPR67], Tseng [Tse90], Bauschke et al. [BBL97], Lewis and Malick [LeM08], Leventhal and Lewis [LeL10], Cegielski and Suchocka [CeS08], Deutsch and Hundal [DeH06a], [DeH06b], [DeH08], and Nedić [Ned10]. A survey of the work in this area up to 1996 is given by Bauschke [Bau96]. In our analysis we will require that the collection $\{X_i\}$ possesses a *linear regularity property*. This notion has been originally introduced by Bauschke [Bau96] in a more general Hilbert space setting, and finds extensive application in alternating (or cyclic) projection algorithms for solving feasibility problems (see for example [DeH08]).

Two works on incremental and randomized methods for convex optimization, by Bertsekas [Ber11a] and by Nedić [Ned11], are strongly related with ours, in somewhat different ways. The work of [Ned11] focuses on gradient and subgradient projection methods with random feasibility steps for convex optimization. In particular, it considers the minimization of a function f over a constraint of the form $X = X_0 \cap \big\{ \cap_{i\in M} X_i \big\}$, where X_0 and X_i are closed convex sets, M is a possibly infinite index set, and f is assumed convex over X_0. Among the methods proposed by [Ned11] the one most closely related to ours is the one for the case $X_0 = \Re^n$, which is given by

$$z_k = x_k - \alpha_k g_k, \qquad x_{k+1} = z_k - \beta \left(z_k - \Pi_{w_k} z_k \right), \tag{6}$$

where g_k can be any subgradient of f at x_k, w_k is a randomly selected index from M, β is a constant stepsize with $0 < \beta < 2$, and α_k is a diminishing stepsize. The analysis focuses on convergence under conditions that are related to the linear regularity assumption for the constraints, noted earlier, as well as on error bounds for the case where the stepsize α_k is instead taken to be constant. By comparing the method (6) of [Ned11] with our method (5) applied to convex optimization problems, we see that the analysis of [Ned11] is different from ours in that it allows f to be nondifferentiable and not necessarily strongly convex (so the problem may have multiple optimal solutions), but it relies on the convex structure of the objective function. On the other hand, the framework of [Ned11] is less general in that it solves a convex optimization problem rather than a VI, it does not consider the case where the objective function is the sum of components or is an expected value, and it does not consider the use of cyclic order projection. Consequently it does not use stochastic samples of the gradients/subgradients, and does not provide a comparative analysis of the random and cyclic orders of constraint-component selection approaches as we do.

The work of [Ber11a] (earlier discussed in the context of a survey of incremental optimization methods in [Ber10]) proposed an algorithmic framework which 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. Random or 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 [NeB00], [NeB01], [BNO03]. However, X is not assumed to be of the form $\cap_{i\in M} X_i$ as in the work of [Ned11] and in the current work. Instead a special case of incremental constraint projections on sets X_i can be implemented via the proximal iterations. In particular, the case $X = \cap_{i=1}^{m} X_i$ is handled (requiring Lipchitz continuity of each f_i, but not requiring the linear regularity assumption) by eliminating each constraint $x \in X_i$, while adding to f_i a penalty function of the form $\gamma \operatorname{dist}(x, X_i)$, where γ is a sufficiently large penalty parameter. A proximal iteration applied to this penalty function is equivalent to a projection iteration applied to the constraint set X_i. When proximal iterations are incrementally applied to the penalty functions $\gamma \operatorname{dist}(x, X_i)$, and are combined with subgradient iterations for f_i, the resulting method takes the form

$$z_k = x_k - \alpha_k g_{i_k}, \qquad x_{k+1} = z_k - \beta_k \left(z_k - \Pi_{w_k} z_k\right),$$

where i_k and w_k are randomly selected indexes from $\{1, \ldots, m\}$, g_{i_k} is any subgradient of f_{i_k} at x_k, α_k is a constant or a diminishing stepsize, and

$$\beta_k = \min\left\{1, \frac{\alpha_k \gamma}{\operatorname{dist}(z_k; X_{w_k})}\right\}.$$

Here the stepsize β_k must be specified by using z_k and X_{w_k}, and is coupled to α_k. This algorithm allows the components f_i to be nondifferentiable, it introduces proximal iterations, and it does not require the linear regularity assumption. It is less general in that, like the method of [Ned11], it applies to a convex optimization problem rather than a VI, and it requires the objective function to be Lipchitz continuous.

The algorithmic treatment of the uncertainties when F is given as an expected value [cf. Eq. (4)], is strongly related to stochastic approximation methods. 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 Kushner and Yin [KuY03], and by Borkar [Bor08], and the aforementioned papers [GOR99] and [JiX08]). Moreover, similar to many sources on convergence analysis of stochastic algorithms, we use a supermartingale convergence theorem.

The remainder of the paper is organized as follows. Section 2 summarizes our assumptions, proof techniques, and several preliminary results. Section 3 focuses on the algorithm with random projection, and derives convergence results and a constant-stepsize error bound. It also discusses extensions of the convergence analysis to various schemes of constraint superset selection that may involve adaptive sampling and/or allow an infinite number of constraint sets. Section 4 obtains corresponding results for the algorithm with cyclic or randomly permuted order projection, and compares its convergence rate with the one of the random projection algorithm. Section 5 discusses applications of the proposed algorithms and presents some computational experiments.

Our notation in summary is as follows. For $x \in \Re^n$, we denote by x' its transpose, and by $\|x\|$ its Euclidean norm (i.e., $\|x\| = \sqrt{x'x}$). The abbreviation "$\xrightarrow{a.s.}$" means "converges almost surely to," while the abbreviation "i.i.d." means "independent identically distributed." For two sequences $\{y_k\}$ and $\{z_k\}$, we write $y_k = O(z_k)$ if there exists a constant $c > 0$ such that $\|y_k\| \le c\|z_k\|$ for each k. In the case where $\{y_k\}$ and $\{z_k\}$ are sequences of random variables, we write "$y_k = O(z_k)$ $w.p.1$" if there exists a constant $c > 0$ such that $\|y_k\| \le c\|z_k\|$ for each k with probability 1. We denote by $\mathcal{F}_k$ the collection

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

so $\{\mathcal{F}_k\}$ is an increasing sequence.

2 Assumptions and Preliminaries

To motivate our assumptions, we first briefly review the convergence mechanism of the classical projection method

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

where Π denotes projection on X [cf. Eq. (2)]. The solution x^* of the VI (1) is the unique fixed point of the preceding iteration for any $\alpha_k > 0$, i.e.,

$$x^* = \Pi\big[x^* - \alpha_k F(x^*)\big].$$

We assume that F is strongly monotone with a constant $\sigma > 0$ such that

$$\big(F(x) - F(y)\big)'(x - y) \ge \sigma\|x - y\|^2, \qquad \forall\, x, y \in \Re^n,$$

and is Lipschitz continuous with a constant $L > 0$ such that

$$\big\|F(x) - F(y)\big\| \le L\|x - y\|, \qquad \forall\, x, y \in \Re^n.$$

Then iteration (7) is strictly contractive for sufficiently small $\alpha_k > 0$. This can be shown as follows:

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \big\|\Pi\big[x_k - \alpha_k F(x_k)\big] - \Pi\big[x^* - \alpha_k F(x^*)\big]\big\|^2 \\
&\le \big\|\big[x_k - \alpha_k F(x_k)\big] - \big[x^* - \alpha_k F(x^*)\big]\big\|^2 \\
&= \big\|(x_k - x^*) - \alpha_k(F(x_k) - F(x^*))\big\|^2 \\
&= \|x_k - x^*\|^2 - 2\alpha_k\big(F(x_k) - F(x^*)\big)'(x_k - x^*) + \alpha_k^2\big\|F(x_k) - F(x^*)\big\|^2 \\
&\le (1 - 2\sigma\alpha_k + \alpha_k^2 L^2)\|x_k - x^*\|^2,
\end{aligned}$$

where the first inequality uses the nonexpansiveness of the projection (i.e., that $\|\Pi x - \Pi y\| \le \|x - y\|$ for all $x, y \in \Re^n$), and the second inequality uses the strong monotonicity and Lipschitz continuity of F. In the case of a constant stepsize, assuming that $\alpha_k = \alpha \in \big(0, \frac{2\sigma}{L^2}\big)$ for all k, the iteration is strictly contractive and converges linearly to the unique fixed point x^*. In the case of diminishing stepsizes, assuming that $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, the iteration can be shown to converge to x^* by using a stochastic approximation argument (see the subsequent analysis).

Our proposed incremental constraint projection algorithm, restated for convenience here,

$$z_k = x_k - \alpha_k f(x_k, v_k), \qquad x_{k+1} = z_k - \beta_k\left(z_k - \Pi_{w_k} z_k\right), \tag{8}$$

differs from the classical method (7) in two important respects. First, the iterates $\{x_k\}$ generated by the algorithm (8) 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. Instead, the incremental projection process guarantees that $\{x_k\}$ approaches X in a stochastic sense as $k \to \infty$. Second, the stepsize α_k must

be diminishing rather than be a constant α. This is necessary because if α_k were a constant, the solution x^* would not be a fixed point of algorithm (8), even if $f(x,v) = F(x)$ for all x and v. Indeed, as we will show later, the stepsize $\{\alpha_k\}$ must be decreased to 0 at a rate faster than $\{\beta_k\}$ in order that the algorithm converges. Additionally a diminishing stepsize α_k is needed if samples $f(x,v)$ of $F(x)$ are used in place of $F(x)$, even if the projection is on X rather than X_{w_k}. This can be understood in light of the stochastic approximation character of the algorithm in the case where $X = \Re^n$.

Let us outline the convergence proof for the algorithm (8) with random projection. Similar to the classical projection method (7), our line of analysis starts with a bound of the iteration error that has the form

$$\|x_{k+1} - x^*\|^2 \le \|x_k - x^*\|^2 - 2\alpha_k F(x_k)'(x_k - x^*) + e(x_k, \alpha_k, \beta_k, w_k, v_k), \tag{9}$$

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. (9) by using properties of random projection and monotone mappings, and then take conditional expectation on both sides. From this we obtain that the random projection algorithm is "stochastically contractive" in the following sense

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

where σ is the constant of strong monotonicity, and δ_k, ϵ_k are positive errors such that $\sum_{k=0}^{\infty} \delta_k < \infty$ and $\sum_{k=0}^{\infty} \epsilon_k < \infty$. Finally, we will use the following supermartingale convergence theorem result due to Robbins and Siegmund [RoS71] to complete the proof.

Theorem 1 *Let $\{y_k\}$, $\{u_k\}$, $\{a_k\}$ and $\{b_k\}$ be sequences of nonnegative random variables so that*

$$\mathbf{E}\,[y_{k+1} \mid \mathcal{G}_k] \le (1 + a_k)y_k - u_k + b_k, \qquad \text{for all } \ k \ge 0 \ \ w.p.1,$$

where $\mathcal{G}_k$ denotes the collection $y_0, \ldots, y_k, u_0, \ldots, u_k, a_0, \ldots, a_k, b_0, \ldots, b_k$. Also, let $\sum_{k=0}^{\infty} a_k < \infty$ and $\sum_{k=0}^{\infty} b_k < \infty$ with probability 1. Then y_k converges almost surely to a nonnegative random variable, and $\sum_{k=0}^{\infty} u_k < \infty$ with probability 1.

This line of analysis is shared with incremental subgradient and proximal methods (see [NeB00], [NeB01]). 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.

Assumption 1 *The mapping F is strongly monotone with a constant $\sigma > 0$, i.e.,*

$$\big(F(x) - F(y)\big)'(x - y) \ge \sigma\|x - y\|^2, \qquad \forall\ x, y \in \Re^n.$$

The mapping $f(\cdot, v)$ is "stochastically Lipschitz continuous" with a constant $L > 0$, i.e.,

$$\mathbf{E}\big[\big\|f(x, v_k) - f(y, v_k)\big\|^2 \mid \mathcal{F}_k\big] \le L^2\|x - y\|^2, \qquad \forall\ x, y \in \Re^n, \tag{10}$$

with probability 1. Moreover, there exists a constant $B > 0$ such that

$$\big\|F(x^*)\big\| \le B, \qquad \mathbf{E}\big[\big\|f(x^*, v_k)\big\|^2 \mid \mathcal{F}_k\big] \le B^2, \qquad \text{for all } k \ge 0,$$

with probability 1.

The stochastic Lipschitz continuity condition (10) resembles ordinary Lipschitz continuity. If $f(x,v) = F(x)$ for all x and v, the scalar L is equal to the Lipschitz continuity constant of F. If v takes finitely many values, Lipschitz continuity of each $f(\cdot, v)$ implies the stochastic Lipschitz continuity condition.

In order for the distance between x_k and X to decrease "on average," we make several assumptions regarding the constraint sets $\{X_i\}$ and the incremental projection process $\{\Pi_{w_k}\}$. The following assumption is a form of regularity of the collection of constraint sets $\{X_i\}$.

Assumption 2 *There exists a positive scalar η such that for any $x \in \Re^n$*

$$\|x - \Pi x\|^2 \leq \eta \max_{i \in M} \|x - \Pi_{X_i} x\|^2,$$

where M is a finite set of indexes, $M = \{1, \ldots, m\}$.

This assumption is known as *linear regularity*, and was introduced and studied by Bauschke [Bau96] (Definition 4.2.1, p. 53) in the more general setting of a Hilbert space; see also Bauschke and Borwein [BaB96] (Definition 5.6, p. 40). Recently, it has been studied by Deutsch and Hundal [DeH08] for the purpose of establishing linear convergence of a cyclic projection method for finding a common point of finitely many convex sets. This linear regularity condition is automatically satisfied when X is a polyhedral set. The discussion in the preceding references provides several other situations where the linear regularity condition holds, and indicates that this condition is a mild restriction in practice.

Although the linear regularity assumption requires $\{X_i\}$ to be a collection of finitely many sets, it can be relaxed to accommodate an infinite number of constraints for random projection algorithms. Consequently, a substantial part of our subsequent analysis can be adapted to the relaxation of Assumption 2; see the discussion of Section 3.3. However, for cyclic projection algorithms, the number of constraints must be finite.

Assumption 3 *We have $\alpha_k \in (0,1)$, $\beta_k \in (0,2)$ for all k, and*

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

where $\gamma_k = \beta_k(2 - \beta_k)$.

Note that to satisfy the condition

$$\sum_{k=0}^{\infty} \frac{\alpha_k^2}{\gamma_k} = \sum_{k=0}^{\infty} \frac{\alpha_k^2}{\beta_k(2-\beta_k)} < \infty,$$

we may either let β_k be equal to a constant in $(0,2)$ for all k, or let β_k decrease to 0 or increase to 2 at a certain rate. Given that $\sum_{k=0}^{\infty} \alpha_k = \infty$, the preceding condition implies that

$$\frac{\alpha_k}{\beta_k} \to 0 \quad \text{or} \quad \frac{\alpha_k}{2-\beta_k} \to 0.$$

We will show that as a consequence of this, the convergence to the constraint set is faster than the convergence to the optimal solution.

Let us now prove a few preliminary technical lemmas. The first gives several basic facts regarding projection.

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 $\beta > 0$,

$$\big\|x - \beta(x - \Pi_S x) - y\big\|^2 \le \|x - y\|^2 - \beta(2-\beta)\|x - \Pi_S x\|^2. \tag{11}$$

(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 $a, b \in \Re$. ■

From Lemma 1(a) and the relation $x_{k+1} = z_k - \beta_k(z_k - \Pi_{w_k} z_k)$ [cf. Eq. (8)], we obtain for any $y \in X$

$$\|x_{k+1} - y\|^2 \le \|z_k - y\|^2 - \beta_k(2-\beta_k)\|\Pi_{w_k} z_k - z_k\|^2 = \big\|(z_k - x_k) + (x_k - y)\big\|^2 - \gamma_k\|\Pi_{w_k} z_k - z_k\|^2,$$

and finally

$$\|x_{k+1} - y\|^2 \le \|x_k - y\|^2 + 2(z_k - x_k)'(x_k - y) + \|z_k - x_k\|^2 - \gamma_k\|\Pi_{w_k} z_k - z_k\|^2, \qquad \forall\, y \in X. \tag{12}$$

This decomposition of the iteration error will serve as the starting point of our main proof of convergence.

The next lemma derives a lower bound for the term $F(x_k)'(x_k - x^*)$ that arises from the decomposition of the iteration error $\|x_{k+1} - x^*\|^2$ [cf. Eq. (9)]. Estimating this term is complicated by the fact that x_k need not belong to X, so the lemma involves the Euclidean distance of x from X, denoted by

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

Lemma 2 *Under Assumption 1, we have*

$$F(x)'(x - x^*) \ge \sigma\|x - x^*\|^2 - B\,\mathrm{d}(x), \qquad \forall\, x \in \Re^n. \tag{13}$$

Proof. We have

$$F(x)'(x - x^*) = \big(F(x) - F(x^*)\big)'(x - x^*) + F(x^*)'(\Pi x - x^*) + F(x^*)'(x - \Pi x). \tag{14}$$

By using the strong monotonicity of F we obtain

$$\big(F(x) - F(x^*)\big)'(x - x^*) \ge \sigma\|x - x^*\|^2,$$

while from the definition of x^* as the solution of the VI (1),

$$F(x^*)'(\Pi x - x^*) \geq 0.$$

Also, by using the inequality $x'y \geq -\|x\|\|y\|$ and the relation $\|F(x^*)\| \leq B$ (cf. Assumption 1), we have

$$F(x^*)'(x - \Pi x) \geq -\big\|F(x^*)\big\|\|x - \Pi x\| \geq -B\,\mathrm{d}(x).$$

By using the preceding three relations in Eq. (14), the desired inequality follows. ■

The next lemma derives some useful estimates based on the Lipschitz condition of Assumption 1.

Lemma 3 *Under Assumption 1, for any $x \in \Re^n$ and $k \geq 0$, we have*

$$\mathbf{E}\big[\|f(x, v_k)\| \mid \mathcal{F}_k\big] \leq L\|x - x^*\| + B,$$

and

$$\mathbf{E}\big[\|f(x, v_k)\|^2 \mid \mathcal{F}_k\big] \leq 2L^2\|x - x^*\|^2 + 2B^2,$$

with probability 1.

Proof. For any $x \in \Re^n$ and $k \geq 0$, we use the triangle inequality to write

$$\big\|f(x, v_k)\big\| \leq \big\|f(x, v_k) - f(x^*, v_k)\big\| + \big\|f(x^*, v_k)\big\|.$$

By taking expectation in the above relation, and using the Cauchy-Schwarz inequality $\mathbf{E}\big[\|y\|\big] \leq \mathbf{E}\big[\|y\|^2\big]^{1/2}$, and Assumption 1, we have

$$\mathbf{E}\big[\|f(x, v_k)\| \mid \mathcal{F}_k\big] \leq \mathbf{E}\big[\|f(x, v_k) - f(x^*, v_k)\|^2 \mid \mathcal{F}_k\big]^{1/2} + \mathbf{E}\big[\|f(x^*, v_k)\|^2 \mid \mathcal{F}_k\big]^{1/2} \leq L\|x - x^*\| + B,$$

with probability 1. By using the inequality $(a + b)^2 \leq 2a^2 + 2b^2$, for all $a, b \in \Re$, we also have

$$\mathbf{E}\big[\|f(x, v_k)\|^2 \mid \mathcal{F}_k\big] \leq 2\mathbf{E}\big[\|f(x, v_k) - f(x^*, v_k)\|^2 + \|f(x^*, v_k)\|^2 \mid \mathcal{F}_k\big] \leq 2L^2\|x - x^*\|^2 + 2B^2,$$

with probability 1. ■

3 Convergence of Random Projection Algorithms

In this section, we will analyze the algorithm

$$z_k = x_k - \alpha_k f(x_k, v_k), \qquad x_{k+1} = z_k - \beta_k\left(z_k - \Pi_{w_k} z_k\right), \tag{15}$$

for the case where the projections Π_{w_k} are randomly generated. We make the following assumption, which requires that each X_i be sampled sufficiently often, and that the samples $f(x_k, v_k)$ be conditionally unbiased.

Assumption 4 *(a) 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.

(b) The random variables v_k, $k = 0, 1, \ldots$, are such that

$$\mathbf{E}\big[f(x, v_k) \mid \mathcal{F}_k\big] = F(x), \qquad \forall\, x \in \Re^n, \quad k \geq 0, \tag{16}$$

with probability 1.

Assumption 4 requires that the random samples of the constraint sets be "nearly independent," in the sense that each constraint is always sampled with sufficient probability, regardless of the sample history. In the case where w_k are independent identically distributed in $\{1, \ldots, m\}$, we have $\rho = 1$. Thus the constant ρ can be viewed as a metric of the efficiency of the sampling process, and it does not scale with the number of constraints m.

3.1 Almost Sure Convergence

Consider the nonnegative function of x

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

which measures the "average progress" of random projection at the kth iteration. This function can be used as a metric of distance between x and the entire constraint set X, as shown by the following lemma.

Lemma 4 *Under Assumptions 2 and 4, we have*

$$\mathbf{E}\big[\|x - \Pi_{w_k} x\|^2 \mid \mathcal{F}_k\big] \geq \frac{\rho}{m\eta}\, \mathrm{d}^2(x), \qquad \forall\, x \in \Re^n, \quad k \geq 0, \tag{17}$$

with probability 1, where $\rho \in (0, 1]$ is the constant of Assumption 4(a).

Proof. By Assumption 2, the index set M is finite, $M = \{1, \ldots, m\}$. 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 2, we obtain

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

■

We are now ready to present the first of the main results of the paper.

Proposition 1 (Convergence of Random Projection Algorithm) *Let Assumptions 1-4 hold. Then the random projection algorithm* (15) *generates a sequence $\{x_k\}$ that converges almost surely to the unique solution x^* of the VI* (1).

Proof. By applying Eq. (12) [which follows from Lemma 1(a)] with $y = x^*$, we obtain

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + 2(z_k - x_k)'(x_k - x^*) + \|z_k - x_k\|^2 - \gamma_k \|\Pi_{w_k} z_k - z_k\|^2. \tag{18}$$

By using Lemma 1(b), we further have

$$\|\Pi_{w_k} x_k - x_k\|^2 \leq 2\|\Pi_{w_k} z_k - z_k\|^2 + 8\|z_k - x_k\|^2,$$

which combined with Eq. (18) yields

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + 2(z_k - x_k)'(x_k - x^*) + (1 + 4\gamma_k)\|z_k - x_k\|^2 - \frac{\gamma_k}{2} \|\Pi_{w_k} x_k - x_k\|^2. \tag{19}$$

Defining $g_k = f(x_k, v_k) - F(x_k)$, we have

$$\begin{aligned}(z_k - x_k)'(x_k - x^*) &= -\alpha_k f(x_k, v_k)'(x_k - x^*) \\ &= -\alpha_k F(x_k)'(x_k - x^*) - \alpha_k g_k'(x_k - x^*) \\ &\leq -\alpha_k \sigma \|x_k - x^*\|^2 + B\alpha_k \,\mathrm{d}(x_k) - \alpha_k g_k'(x_k - x^*),\end{aligned}$$

where the inequality follows from Lemma 2. We apply the preceding inequality to Eq. (19) and obtain

$$\begin{aligned}\|x_{k+1} - x^*\|^2 \leq (1 - 2\alpha_k\sigma)\|x_k - x^*\|^2 - 2\alpha_k g_k'(x_k - x^*) + (1 + 4\gamma_k)\|z_k - x_k\|^2 \\ + 2B\alpha_k \,\mathrm{d}(x_k) - \frac{\gamma_k}{2}\|\Pi_{w_k} x_k - x_k\|^2.\end{aligned} \tag{20}$$

According to Assumption 4, since $x_k \in \mathcal{F}_k$, we have

$$\mathbf{E}\big[g_k'(x_k - x^*) \mid \mathcal{F}_k\big] = \big(\mathbf{E}\big[f(x_k, v_k) \mid \mathcal{F}_k\big] - F(x_k)\big)'(x_k - x^*) = 0. \tag{21}$$

From Lemma 3, we have

$$\mathbf{E}\big[\|z_k - x_k\|^2 \mid \mathcal{F}_k\big] = \alpha_k^2 \mathbf{E}\big[\|f(x_k, v_k)\|^2 \mid \mathcal{F}_k\big] \leq \alpha_k^2(2L^2\|x_k - x^*\|^2 + 2B^2). \tag{22}$$

From Lemma 4, we have

$$\mathbf{E}\big[\,\|\Pi_{w_k} x_k - x_k\|^2 \mid \mathcal{F}_k\big] \geq \frac{\rho}{m\eta}\,\mathrm{d}^2(x_k). \tag{23}$$

Taking conditional expectation on both sides of Eq. (20) and applying Eqs. (21)-(23), we obtain

$$\begin{aligned}\mathbf{E}\big[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\big] \leq (1 - 2\alpha_k\sigma)\|x_k - x^*\|^2 + 2\alpha_k^2(1 + 4\gamma_k)\left(L^2\|x_k - x^*\|^2 + B^2\right) \\ + 2B\alpha_k \,\mathrm{d}(x_k) - \frac{\rho\gamma_k}{2m\eta}\,\mathrm{d}^2(x_k).\end{aligned}$$

Finally, by writing the last two terms in the right-hand side as

$$2B\alpha_k \,\mathrm{d}(x_k) - \frac{\rho\gamma_k}{2m\eta}\,\mathrm{d}^2(x_k) = -\frac{\rho\gamma_k}{2m\eta}\left(\mathrm{d}(x_k) - 2Bm\eta\rho^{-1}\frac{\alpha_k}{\gamma_k}\right)^2 + 2B^2 m\eta\rho^{-1}\frac{\alpha_k^2}{\gamma_k},$$

and bounding them by $2B^2 m\eta\rho^{-1}\dfrac{\alpha_k^2}{\gamma_k}$, we further obtain

$$\begin{aligned}\mathbf{E}\big[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\big] &\leq \left(1 - 2\alpha_k\sigma + 2L^2(1 + 4\gamma_k)\alpha_k^2\right)\|x_k - x^*\|^2 + 2B^2(1 + 4\gamma_k)\alpha_k^2 + 2B^2 m\eta\rho^{-1}\frac{\alpha_k^2}{\gamma_k} \\ &\leq \left(1 - 2\alpha_k\sigma + O\left(\alpha_k^2\right)\right)\|x_k - x^*\|^2 + O\left(\alpha_k^2 + \frac{\alpha_k^2}{\gamma_k}\right).\end{aligned} \tag{24}$$

From Assumption 3, we have $\sum_{k=0}^{\infty}\alpha_k^2 < \infty$ and $\sum_{k=0}^{\infty}\left(\alpha_k^2 + \frac{\alpha_k^2}{\gamma_k}\right) < \infty$, so the Supermartingale Convergence Theorem 1 applies to Eq. (24). It follows that $\|x_k - x^*\|^2$ converges almost surely to a nonnegative random variable, and that

$$\sum_{k=0}^{\infty} 2\alpha_k\sigma\|x_k - x^*\|^2 < \infty, \qquad w.p.1.$$

The preceding relation implies that $\|x_k - x^*\|^2 \xrightarrow{a.s.} 0$ with probability 1 [if $\|x_k - x^*\|^2$ converged to a nonzero random variable, then $\sum_{k=0}^{\infty} 2\alpha_k\sigma\|x_k - x^*\|^2 = \infty$ with positive probability, thus yielding a contradiction]. To conclude, we have $\|x_k - x^*\|^2 \xrightarrow{a.s.} 0$, or equivalently, $x_k \xrightarrow{a.s.} x^*$. ■

3.2 Convergence Rate and Constant Stepsize Error Bound

Let us now focus on the rate of convergence of the random projection algorithm (15). We will consider a diminishing stepsize α_k, and derive the convergence rates of the iteration error $\|x_k - x^*\|$ and the feasibility error $\mathrm{d}(x_k)$. In particular we will derive an estimate on the number of iterations needed to achieve a specified error tolerance. We will also consider a constant stepsize α_k, and show that in this case the algorithm converges to within a certain error bound. This error bound will be compared later to the corresponding error bound for the cyclic projection algorithm.

Proposition 2 (Random Projection Algorithm: Convergence Rate for Diminishing $\{\alpha_k\}$) *Let Assumptions 1-4 hold, let $\alpha_k \in \left(0, \frac{\sigma}{5L^2}\right)$ for all k, and let $\{x_k\}$ be generated by the random projection algorithm* (15)*. For any positive scalar ϵ, there exists a random variable N such that*

$$\min_{0\le k\le N} \left\{\|x_k - x^*\|^2 - \delta_k\right\} \le \epsilon, \tag{25}$$

with probability 1, where

$$\delta_k = \frac{\alpha_k}{\sigma}\left(L^2\epsilon + B^2 + B^2 m\eta\rho^{-1}\gamma_k^{-1}\right) + O\left(\alpha_k^2 + \alpha_k\gamma_k\right) \le O\left(\frac{\alpha_k}{\gamma_k}\right),$$

and

$$\mathbf{E}\left[\sum_{k=0}^{N-1}\alpha_k\right] \le \frac{\|x_0 - x^*\|^2}{2\sigma\epsilon}. \tag{26}$$

Proof. Given $\epsilon > 0$, we let

$$\delta_k = \frac{\alpha_k}{2\sigma - c_{1,k}\alpha_k}\left(c_{1,k}\epsilon + c_{2,k} + c_{3,k}\gamma_k^{-1}\right),$$

where

$$c_{1,k} = 2L^2(1+4\gamma_k), \qquad c_{2,k} = 2B^2(1+4\gamma_k), \qquad c_{3,k} = 2B^2 m\eta\rho^{-1}.$$

It can be seen that

$$\delta_k \le \frac{\alpha_k}{\sigma}\left(L^2\epsilon + B^2 + B^2 m\eta\rho^{-1}\gamma_k^{-1}\right) + O\left(\alpha_k^2 + \alpha_k\gamma_k\right),$$

where L, B are the constants in Assumption 1, and ρ is the constant in Lemma 4. Note that, since $\alpha_k \in (0, \frac{\sigma^2}{5L^2})$ and $\gamma_k = \beta_k(2-\beta_k) \le 1$, we can verify that $2\sigma - c_{1,k}\alpha_k \ge 2\sigma - 10L^2\alpha_k > 0$ and $\delta_k > 0$ for all k.

Define a new process $\{\hat{x}_k\}$, which is identical to $\{x_k\}$ except that once $\hat{x}_k$ enters the level set

$$L_k = \left\{x \in \Re^n \mid \|x - x^*\|^2 \le \delta_k + \epsilon\right\}$$

the process stays at $\hat{x}_k = x^*$ for all future k. Following the analysis of Prop. 1 [cf. Eq. (24)], we have for all k with probability 1 that

$$\mathbf{E}\big[\|\hat{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_k\big] \le \left(1 - 2\alpha_k\sigma + c_{1,k}\alpha_k^2\right)\|\hat{x}_k - x^*\|^2 + c_{2,k}\alpha_k^2 + c_{3,k}\frac{\alpha_k^2}{\gamma_k}.$$

We write this relation as

$$\mathbf{E}\big[\|\hat{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_k\big] \le \|\hat{x}_k - x^*\|^2 - \xi_k, \tag{27}$$

where we define

$$\xi_k = \begin{cases} \left(2\alpha_k\sigma - c_{1,k}\alpha_k^2\right)\|\hat{x}_k - x^*\|^2 - c_{2,k}\alpha_k^2 - c_{3,k}\dfrac{\alpha_k^2}{\gamma_k} & \text{if } \hat{x}_k \notin L_k, \\ 0 & \text{otherwise.} \end{cases} \tag{28}$$

When $\hat{x}_k \notin L_k$, we can verify by using the definition of δ_k that

$$\xi_k \geq \left(2\alpha_k\sigma - c_{1,k}\alpha_k^2\right)(\delta_k + \epsilon) - \left(c_{2,k}\alpha_k^2 + c_{3,k}\frac{\alpha_k^2}{\gamma_k}\right) = (2\sigma\epsilon)\alpha_k. \tag{29}$$

Note that $\xi_k \geq 0$ for all k. By applying Theorem 1 to Eq. (27), we have

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

If $\hat{x}_k \notin L_k$ for all k, by using Eq. (29) and Assumption 3 we would obtain

$$\sum_{k=0}^{\infty} \xi_k \geq (2\sigma\epsilon)\sum_{k=0}^{\infty} \alpha_k = \infty,$$

with positive probability, yielding a contradiction. Thus $\{\hat{x}_k\}$ enters L_k eventually, with probability 1.

Let N be the smallest integer N such that $\hat{x}_k \in L_k$ for all $k \geq N$ with probability 1, so that Eq. (25) holds. By taking expectation on both sides of Eq. (27), we have

$$\mathbf{E}\left[\|\hat{x}_{k+1} - x^*\|^2\right] \leq \|\hat{x}_0 - x^*\|^2 - \mathbf{E}\left[\sum_{t=0}^{k} \xi_t\right]. \tag{30}$$

By letting $k \to \infty$ and using the monotone convergence theorem, we obtain

$$\|x_0 - x^*\|^2 \geq \mathbf{E}\left[\sum_{k=0}^{\infty} \xi_k\right] = \mathbf{E}\left[\sum_{k=0}^{N-1} \xi_k\right] \geq (2\sigma\epsilon)\mathbf{E}\left[\sum_{k=0}^{N-1} \alpha_k\right].$$

This proves Eq. (26). ■

Equation (26) quantifies the random number of iterations needed to achieve the solution accuracy specified by Eq. (25). If we take α_k and β_k to be constant stepsizes, we obtain the following result with a nearly identical proof to the one of the preceding proposition.

Proposition 3 (Random Projection Algorithm: Error Bound for Constant $\{\alpha_k\}$ and $\{\beta_k\}$) *Let Assumptions 1, 2, and 4 hold, let the stepsizes be constant scalars satisfying*

$$\alpha_k = \alpha \in \left(0, \frac{\sigma}{5L^2}\right), \qquad \beta_k = \beta \in (0,2), \qquad \gamma_k = \gamma = \beta(2-\beta), \qquad \forall\ k \geq 0,$$

and let $\{x_k\}$ be generated by the random projection algorithm (15)*. Then*

$$\liminf_{k\to\infty} \|x_k - x^*\|^2 \leq \delta(\alpha,\gamma) \stackrel{\text{def}}{=} \frac{\alpha\left(1 + 4\gamma + m\eta\rho^{-1}\gamma^{-1}\right)B^2}{\sigma - L^2(1+4\gamma)\alpha} \leq O\left(\frac{m\alpha}{\sigma\gamma}\right),$$

with probability 1. For any positive scalar ϵ, there exists a random variable N such that

$$\min_{0\leq k\leq N} \|x_k - x^*\|^2 \leq \delta(\alpha,\gamma) + \epsilon,$$

with probability 1, and N satisfies

$$\mathbf{E}[N] \leq \frac{\|x_0 - x^*\|^2}{\left(2\sigma - 2L^2(1+4\gamma)\alpha\right)\epsilon\alpha}.$$

Proof. We repeat the analysis of Prop. 2 with α_k replaced with α, β_k replaced with β, γ_k replaced with γ, δ_k replaced with $\delta(\alpha,\gamma)$, and ϵ replaced with $\frac{(2\sigma-2L^2(1+4\gamma)\alpha)}{2\sigma}\epsilon$. Then L_k is replaced by

$$L(\epsilon) = \left\{x \in \Re^n \mid \ \|x - x^*\|^2 \le \delta(\alpha,\gamma) + \epsilon\right\}, \qquad k = 0, 1, \ldots.$$

Similar to Eq. (27), we have

$$\mathbf{E}\big[\|\hat{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_k\big] \le \|\hat{x}_k - x^*\|^2 - \xi_k, \tag{31}$$

where we can verify that

$$\begin{aligned} &\xi_k \ge \big(2\sigma - 2L^2(1+4\gamma)\alpha\big)\epsilon\alpha, && \text{if } \hat{x}_k \notin L(\epsilon), \\ &\xi_k = 0, && \text{if } \hat{x}_k \in L(\epsilon). \end{aligned}$$

Since $\alpha \in \big(0, \frac{\sigma}{5L^2}\big)$ and $\gamma = \beta(2-\beta) < 1$, we have $\xi_k \ge 0$ for all k.

By applying Theorem 1 to Eq. (31) and using a similar analysis as in Prop. 2, we can show that $\xi_k = 0$ for all k sufficiently large with probability 1. This implies that $\{\hat{x}_k\}$ and $\{x_k\}$ both enter $L(\epsilon)$ eventually, and since $\epsilon > 0$ is arbitrary, it further implies that

$$\liminf_{k\to\infty} \|x_k - x^*\|^2 \le \delta(\alpha,\gamma), \qquad w.p.1.$$

Finally, by defining N similar to the proof of Prop. 2, we can prove the desired error bounds involving N. ■

Let us also consider the convergence rate of the distance to the constraint set for our algorithm. By using an analysis similar to that of Prop. 2, we obtain the following result.

Proposition 4 (Random Projection Algorithm: Convergence Rate of $\mathrm{d}(x_k)$) *Let Assumptions 1-4 hold, let $\alpha_k \in \big(0, \frac{\sigma}{5L^2}\big)$ for all k, and let $\{x_k\}$ be generated by the random projection algorithm* (15)*. For any positive scalar ϵ, there exists a random variable N such that*

$$\min_{0\le k\le N} \left\{ \mathrm{d}^2(x_k) - \delta_k \right\} \le \epsilon, \tag{32}$$

where

$$\delta_k = 8B^2 m\eta\rho^{-1}\left(4 + {\gamma_k}^{-1} + 2m\eta\rho^{-1}\gamma_k^{-2}\right)\alpha_k^2 \le O\left(\frac{\alpha_k^2}{\gamma_k^2}\right),$$

with probability 1, and

$$\mathbf{E}\left[\sum_{k=0}^{N-1}\gamma_k\right] \le \left(4m\eta\rho^{-1}\right)\frac{\|x_0 - x^*\|^2}{\epsilon}. \tag{33}$$

Proof. Given $\epsilon > 0$, we let

$$\delta_k = \frac{4m\eta\rho^{-1}}{\gamma_k}\left(c_{2,k}\alpha_k^2 + 2c_{3,k}\frac{\alpha_k^2}{\gamma_k}\right) = 8B^2 m\eta\rho^{-1}\left(4 + {\gamma_k}^{-1} + 2m\eta\rho^{-1}\gamma_k^{-2}\right)\alpha_k^2,$$

where $c_{2,k} = 2B^2(1+4\gamma_k)$ and $c_{3,k} = 2B^2 m\eta\rho^{-1}$.

Define a new process $\{\hat{x}_k\}$ which is identical to $\{x_k\}$ except that once $\hat{x}_k$ enters the level set

$$L_k = \left\{x \in \Re^n \mid \ \|x - x^*\|^2 \le \delta_k + \epsilon\right\}$$

the process stays at $\hat{x}_k = x^*$ for all future k. Following the analysis of Prop. 1 [cf. Eq. (24)], we have

$$\begin{aligned}
\mathbf{E}[\|\hat{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_k] &\leq \left(1 - 2\sigma\alpha_k + c_{1,k}\alpha_k^2\right)\|\hat{x}_k - x^*\|^2 - \frac{\rho\gamma_k}{2m\eta}\left(\mathrm{d}(x_k) - 2Bm\eta\rho^{-1}\frac{\alpha_k}{\gamma_k}\right)^2 + c_{2,k}\alpha_k^2 + c_{3,k}\frac{\alpha_k^2}{\gamma_k} \\
&\leq \|\hat{x}_k - x^*\|^2 - \frac{\rho\gamma_k}{2m\eta}\left(\mathrm{d}(x_k) - 2Bm\eta\rho^{-1}\frac{\alpha_k}{\gamma_k}\right)^2 + c_{2,k}\alpha_k^2 + c_{3,k}\frac{\alpha_k^2}{\gamma_k} \\
&\leq \|\hat{x}_k - x^*\|^2 - \frac{\rho\gamma_k}{2m\eta}\left(\frac{1}{2}\mathrm{d}^2(x_k) - 4B^2m^2\eta^2\rho^{-2}\frac{\alpha_k^2}{\gamma_k^2}\right) + c_{2,k}\alpha_k^2 + c_{3,k}\frac{\alpha_k^2}{\gamma_k} \\
&= \|\hat{x}_k - x^*\|^2 - \frac{\rho\gamma_k}{4m\eta}\mathrm{d}^2(x_k) + c_{2,k}\alpha_k^2 + 2c_{3,k}\frac{\alpha_k^2}{\gamma_k},
\end{aligned}$$

where the first inequality uses the fact $\alpha_k \in (0, \frac{\sigma}{5L^2}) \subset (0, \frac{\sigma}{1+4\gamma_k})$, and the second inequality uses the fact $-(a-b)^2 \leq -\left(\frac{1}{2}a^2 - b^2\right)$ for any $a, b \in \Re$. Equivalently, this is the relation

$$\mathbf{E}\big[\|\hat{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_k\big] \leq \|\hat{x}_k - x^*\|^2 - \xi_k, \tag{34}$$

where we define

$$\xi_k = \begin{cases} \dfrac{\gamma_k\rho}{4m\eta}\mathrm{d}^2(x_k) - c_{2,k}\alpha_k^2 - 2c_{3,k}\dfrac{\alpha_k^2}{\gamma_k} & \text{if } \hat{x}_k \notin L_k, \\ 0 & \text{otherwise.} \end{cases} \tag{35}$$

When $\hat{x}_k \notin L_k$, we can verify by using the definition of δ_k that

$$\xi_k \geq \frac{\gamma_k\rho}{4m\eta}(\delta_k + \epsilon) - \left(c_{2,k}\alpha_k^2 + 2c_{3,k}\frac{\alpha_k^2}{\gamma_k}\right) = \frac{\gamma_k\rho\epsilon}{4m\eta}. \tag{36}$$

Note that $\xi_k \geq 0$ for all k. By applying Theorem 1 to Eq. (34), we have $\sum_{k=0}^{\infty}\xi_k < \infty$ with probability 1. It follows that $\{\hat{x}_k\}$ enters L_k eventually, or equivalently, $\xi_k = 0$ for all k sufficiently large with probability 1.

Let N be the smallest integer N such that $\hat{x}_k \in L_k$ for all $k \geq N$ with probability 1, so that Eq. (32) holds. By taking total expectation of both sides of Eq. (34) and adding over indexes up to k, we have

$$\mathbf{E}\left[\|\hat{x}_{k+1} - x^*\|^2\right] \leq \|\hat{x}_0 - x^*\|^2 - \mathbf{E}\left[\sum_{t=0}^{k}\xi_t\right]. \tag{37}$$

By letting $k \to \infty$ and using the monotone convergence theorem, we obtain

$$\|x_0 - x^*\|^2 \geq \mathbf{E}\left[\sum_{k=0}^{\infty}\xi_k\right] = \mathbf{E}\left[\sum_{k=0}^{N-1}\xi_k\right] \geq \frac{\rho\epsilon}{4m\eta}\mathbf{E}\left[\sum_{k=0}^{N-1}\gamma_k\right],$$

where the last inequality follows from Eq. (36). This proves Eq. (33). ■

A comparison between Prop. 2 and 4 suggests that the random projection algorithm converges to the constraint set at a faster rate than to the optimal solution. In particular, from Prop. 4 it follows that $\mathrm{d}^2(x_k)$ converges to a smaller error bound (on the order of $\frac{\alpha_k^2}{\gamma_k^2}$) in fewer iterations, compared to the convergence of $\|x_k - x^*\|^2$ into an error bound (on the order of $\frac{\alpha_k}{\gamma_k}$) as given by Prop. 2, since γ_k is much larger than α_k. This two time scale phenomenon is consistent with the way we select the stepsizes.

3.3 Extensions

In the presentation of the current section, we have focused on the case where $\{X_i\}$ is a finite collection of sets, possessing the linear regularity property, and each X_i is sampled nearly independently, with the purpose

of drawing a comparison between random and cyclic orders later in Section 4. However, our analysis can be adapted to hold under a more general set of conditions.

A key step of the analysis of the current section is to obtain a lower bound of the progress towards feasibility for the algorithm [cf. Eq. (17) of Lemma 4]. Intuitively, as long as every projection step makes sufficient progress "on average," the convergence of the algorithm follows. In particular, we can replace Assumptions 2 and 4(a) with the following more general condition: *there exists* $c > 0$ *such that for any* $k \geq 0$

$$\mathbf{E}\big[\|x_k - \Pi_{w_k} x_k\|^2 \mid \mathcal{F}_k\big] \geq c\, \mathrm{d}^2(x_k), \qquad w.p.1. \tag{38}$$

Under this condition, as well as Assumptions 1, 3, 4(b), the proof of Prop. 1 goes through. Therefore, under these more general assumptions, the random projection algorithm (15) is still convergent to the unique solution x^* of the VI (1). Moreover, the rate of convergence results of Props. 2-4 can also be adapted accordingly, by replacing the constant $\dfrac{\rho}{m\eta}$, which comes from Eq. (17), with the constant c from Eq. (38).

Condition (38) allows more flexibility in choosing the sample constraint sets $\{X_{w_k}\}$. In particular, we may select the sample constraints adaptively based on the current iterates. For an example, consider the case where the collection $\{X_i\}$ is finite and possesses the linear regularity property (cf. Assumption 2), and let w_k be the index of the most distant constraint set to the current iterate, i.e.,

$$w_k = \mathrm{argmax}_{i \in M} \|x_k - \Pi_i x_k\|.$$

Then by using linear regularity, we obtain

$$\|x_k - \Pi_{w_k} x_k\| = \max_{i \in M} \|x_k - \Pi_i x_k\| = \frac{1}{\sqrt{\eta}}\, \mathrm{d}(x_k), \qquad w.p.1. \tag{39}$$

It follows that condition (38) is satisfied with $c = 1/\eta$, and the associated random projection algorithm is convergent. In fact, this algorithm has a better rate of convergence than the algorithm that uses nearly independent samples of the constraints (cf. Assumption 4). More specifically, by using projection to the most distant constraint set, we can remove the factor m in the error bounds of Props. 2-4. In particular, in an analog of Prop. 3 the error constant $\delta(\alpha, \gamma)$ is of the form $O\left(\dfrac{\alpha}{\sigma\gamma}\right)$ instead of $O\left(\dfrac{m\alpha}{\sigma\gamma}\right)$, while in an analog of Prop. 4, Eq. (33) takes the form

$$\mathbf{E}\left[\sum_{k=0}^{N-1} \gamma_k\right] \leq \frac{4\eta\|x_0 - x^*\|^2}{\epsilon},$$

indicating a much faster attainment of feasibility. However, this approach, although having a superior convergence rate, is often impractical because finding the most distant constraint set index can be expensive. Instead, an index of a "nearly" most distant constraint set may either be deterministically computed or stochastically obtained by sampling (e.g., according to an importance sampling distribution related to the iterates' history). The structure and properties of such constraint selection rules are interesting subjects for future research.

More generally, condition (38) extends to the case where $\{X_i\}_{i \in M}$ is a collection of infinitely (even uncountably) many sets, which applies to a broader range of contexts. Since any closed convex set X is the intersection of all the halfspaces containing it, the idea of random superset projection can be extended to problems with arbitrary convex constraint. By appropriately selecting the halfspaces, we may obtain a bound of the form (38) and establish the convergence of the associated algorithm. As an example, at each iteration we may select a halfspace X_{w_k} that properly separates from X a neighborhood of the current iterate x_k. This type of analysis is related to the works by [Fuk86] and [CeG08], and is another interesting subject for future research.

4 Convergence of Cyclic Projection Algorithms

An alternative to random projection, in the case where $M = \{1, \ldots, m\}$, is to cyclically select the constraint set X_{w_k} from the collection $\{X_i\}_{i=1}^m$ according to either a deterministic order or a randomly permuted order. Each cycle consists of m iterations. To be more general, we allow the samples $f(x_k, v_k)$ to be selected in a cyclic manner as well. The algorithm takes the same form as Eq. (15), i.e.,

$$z_k = x_k - \alpha_k f(x_k, v_k), \qquad x_{k+1} = z_k - \beta_k \left(z_k - \Pi_{w_k} z_k\right). \tag{40}$$

We make the following assumption regarding the sampling process, which parallels Assumption 4.

Assumption 5 *(a) Each cycle t consists of m iterations, corresponding to indexes $k = tm, tm + 1, \ldots, (t+1)m - 1$. Iterations within cycle t use constant stepsizes, denoted by*

$$\overline{\alpha}_t = \alpha_k, \qquad \overline{\beta}_t = \beta_k, \qquad \overline{\gamma}_t = \gamma_k = \beta_k(2 - \beta_k), \qquad k = tm, tm+1, \ldots, (t+1)m - 1.$$

However, the sequence $\{\beta_k\}$ satisfies $\limsup_{k\to\infty} \beta_k < 2$.

(b) Within each cycle t,

$$\frac{1}{m} \sum_{k=tm}^{(t+1)m-1} \mathbf{E}\left[f(x, v_k) \mid \mathcal{F}_{tm}\right] = F(x), \qquad \forall\, x \in \Re^n, \qquad w.p.1.$$

(c) Within each cycle t, the sequence of constraint sets $\{X_{w_k}\}$, where $k = tm, tm+1, \ldots, (t+1)m - 1$, is a permutation of $\{X_1, \ldots, X_m\}$.

We refer to the algorithm under the preceding assumption as the *cyclic projection algorithm*. Note that this assumption covers several interesting cases. For example, in the case where $F(x)$ is evaluated without sampling $[f(x, v) \equiv F(x)]$, the algorithm differs from the classical projection method only in the way the constraint projection is performed. For another example, we may let v_k be generated as i.i.d. random variables, so the algorithm chooses samples of F randomly and independently, but chooses samples of the constraint sets cyclically. Also covered by Assumption 5 is the important case where the mapping F is the sum of a large number of component functions. In a more general situation, F may have an arbitrary (possibly infinite) number of component functions:

$$F(x) = \sum_{i \in I} F_i(x),$$

where I is the set of indexes. In this case, we may let $\{I_1, \ldots, I_m\}$ be a partition of I and use the following samples

$$f(x_k, v_k) = \frac{m}{p_{j_k}} F_{j_k}(x_k), \qquad \text{where} \quad j_k \in I_{i_k}.$$

Here $v_k = (i_k, j_k)$, where i_k is selected from $\{1, \ldots, m\}$ cyclically, and j_k is then obtained by sampling from I_{i_k} independently with probability p_{j_k}. Assumption 5 is satisfied in all the cases mentioned above.

We will show that under Assumption 5, as well as assumptions on strong monotonicity, Lipschitz continuity, stepsizes, and linear regularity of the constraints sets (namely Assumptions 1-3), the cyclic projection algorithm (40) converges almost surely to the unique solution x^* of the VI (1). The proof idea is to partition the sequence of iterates $\{x_k\}$ into cycles

$$\{x_{tm}, \ldots, x_{(t+1)m-1}\}, \qquad t = 1, 2, \ldots,$$

and to consider the m iterations within the same cycle as a single step. To do this, we will argue that the iterates $\{x_k\}$ "do not change much" within one cycle. In this way, the m iterations involving $\{X_{w_k}\}$ and $\{f(x_k, v_k)\}$ resemble a single iteration involving X and F. This will show that the mapping $x_{tm} \mapsto x_{(t+1)m}$ is asymptotically contractive in the sense that

$$\mathbf{E}\big[\|x_{(t+1)m} - x^*\|^2 \mid \mathcal{F}_{tm}\big] \leq (1 - 2m\sigma\overline{\alpha}_t + \delta_k)\,\|x_{tm} - x^*\|^2 + \epsilon_k,$$

where δ_k and ϵ_k are nonnegative random variables such that $\sum_{k=0}^{\infty}(\delta_k + \epsilon_k) < \infty$. Then it will follow by the supermartingale convergence argument that $\{x_{tm}\}$ converges to the solution x^* as $t \to \infty$. Finally, since the iterates within one cycle become increasingly close to each other, it will follow that $\{x_k\}$ converges to the same limit.

4.1 Almost Sure Convergence

We will be using Assumptions 1-3 and 5, so Lemmas 1-3 still hold. According to the assumptions on the stepsizes [Assumptions 3 and 5(a)], we can verify that

$$\frac{\alpha_k}{\beta_k} \to 0, \qquad \beta_k \leq O(1), \qquad \gamma_k \leq O(1), \qquad \frac{\beta_k}{\gamma_k} \leq O(1), \qquad \frac{\gamma_k}{\beta_k} \leq O(1).$$

We will frequently use the $O(\cdot)$ notation to simplify the subsequent analysis. The following lemma gives a uniform bound on $\|f(x_k, v_k)\|$ for $k = tm, \ldots, (t+1)m - 1$, within a cycle. The bound is in terms of the distance between the starting iterate x_{tm} and x^*.

Lemma 5 *Under Assumptions 1-3 and 5, for any $t \geq 0$ and $k = tm, \ldots, (t+1)m - 1$,*

$$\mathbf{E}\big[\|f(x_k, v_k)\| \mid \mathcal{F}_{tm}\big]^2 \leq \mathbf{E}\big[\|f(x_k, v_k)\|^2 \mid \mathcal{F}_{tm}\big] \leq O\left(\|x_{tm} - x^*\|^2 + 1\right), \qquad w.p.1. \tag{41}$$

Proof. By applying Lemma 1(a) to algorithm (40), we have

$$\|x_{k+1} - x^*\|^2 \leq \|z_k - x^*\|^2 \leq \big(\|x_k - x^*\| + \alpha_k\|f(x_k, v_k)\|\big)^2.$$

Taking conditional expectation on both sides yields

$$\begin{aligned}
\mathbf{E}\big[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\big] &\leq \|x_k - x^*\|^2 + 2\alpha_k\mathbf{E}\big[\|f(x_k, v_k)\| \mid \mathcal{F}_k\big]\|x_k - x^*\| + \alpha_k^2\mathbf{E}\big[\|f(x_k, v_k)\|^2 \mid \mathcal{F}_k\big] \\
&\leq \|x_k - x^*\|^2 + 2\alpha_k\big(L\|x_k - x^*\| + B\big)\|x_k - x^*\| + \alpha_k^2\big(2L^2\|x_k - x^*\|^2 + 2B^2\big) \\
&= \big(1 + 2\alpha_k L + 2\alpha_k^2 L^2\big)\,\|x_k - x^*\|^2 + 2\alpha_k B\|x_k - x^*\| + 2\alpha_k^2 B^2 \\
&\leq \big(1 + \alpha_k(2L+1) + 2\alpha_k^2 L^2\big)\|x_k - x^*\|^2 + \alpha_k B^2 + 2\alpha_k^2 B^2,
\end{aligned}$$

where the first inequality uses the fact $x_k \in \mathcal{F}_k$, the second inequality uses Lemma 3, and the third inequality uses the relation $2\alpha_k B\|x_k - x^*\| \leq \alpha_k\|x_k - x^*\|^2 + \alpha_k B^2$. Since $\alpha_k \to 0$, it follows that

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

Let $t \geq 0$. By induction we have for all $k = tm, \ldots, (t+1)m - 1$, that

$$\begin{aligned}
\mathbf{E}\big[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_{tm}\big] &\leq \left(\prod_{j=tm}^{(t+1)m-1}(1 + O(\alpha_j))\right)\|x_{tm} - x^*\|^2 + \sum_{j=tm}^{(t+1)m-1}\left(\prod_{i=j}^{(t+1)m-1}(1 + O(\alpha_i))\right)O(\alpha_j) \\
&\leq O\left(\|x_{tm} - x^*\|^2 + 1\right).
\end{aligned}$$

Then by using Lemma 3, we obtain

$$\begin{aligned}\mathbf{E}\big[\|f(x_k,v_k)\|^2 \mid \mathcal{F}_{tm}\big] &= \mathbf{E}\big[\mathbf{E}\big[\|f(x_k,v_k)\|^2 \mid \mathcal{F}_k\big] \mid \mathcal{F}_{tm}\big] \le 2L^2\ \mathbf{E}\big[\|x_k - x^*\|^2 \mid \mathcal{F}_{tm}\big] + 2B^2 \\ &\le O\big(\|x_{tm} - x^*\|^2 + 1\big),\end{aligned}$$

for all $k = tm, \ldots, (t+1)m-1$, with probability 1. Finally, we complete the proof by using the Cauchy-Schwarz inequality. ■

We will now argue that the iterates $\{x_k\}$ do not change "too much" within a cycle. Define the maximal change of iterates within cycle t to be

$$\Delta_t = \max_{tm \le k \le (t+1)m-1} \big\{\|x_k - x_{tm}\|\big\}.$$

The next lemma states that this maximal change per cycle is bounded by a diminishing term, which is determined by the stepsizes, the distances from the starting iterate to the optimal solution and to the constraint set.

Lemma 6 *Under Assumptions 1-3 and 5, for any $t \ge 0$,*

$$\mathbf{E}[\Delta_t^2 \mid \mathcal{F}_{tm}] \le O(m\overline{\beta}_t)\, \mathrm{d}^2(x_{tm}) + O\left(m^4 \overline{\alpha}_t^2\right)\big(\|x_{tm} - x^*\|^2 + 1\big), \qquad w.p.1.$$

Proof. We will use the inequality $\Delta_t \le \sum_{k=tm}^{(t+1)m-1} \|x_k - x_{k+1}\|$ to obtain the upper bound. From the relation

$$x_{k+1} = x_k - \alpha_k f(x_k, v_k) - \beta_k(z_k - \Pi_{w_k} z_k)$$

[cf. Eq. (40)], we obtain

$$\|x_k - x_{k+1}\|^2 \le \big(\alpha_k\|f(x_k,v_k)\| + \beta_k\|z_k - \Pi_{w_k} z_k\|\big)^2 \le 2\alpha_k^2\|f(x_k,v_k)\|^2 + 2\beta_k^2\|z_k - \Pi_{w_k} z_k\|^2.$$

By applying Eq. (12) with $y = \Pi x_{tm}$ we have

$$\|z_k - \Pi_{w_k} z_k\|^2 \le \frac{1}{\gamma_k}\left(\|x_k - \Pi x_{tm}\|^2 - \|x_{k+1} - \Pi x_{tm}\|^2 + \alpha_k^2\|f(x_k,v_k)\|^2 + 2\alpha_k\|f(x_k,v_k)'(x_k - \Pi x_{tm})\|\right).$$

By combining the last two relations, we obtain

$$\begin{aligned}\|x_k - x_{k+1}\|^2 \le{}& \frac{2\beta_k^2}{\gamma_k}\left(\|x_k - \Pi x_{tm}\|^2 - \|x_{k+1} - \Pi x_{tm}\|^2 + 2\alpha_k\|f(x_k,v_k)'(x_k - \Pi x_{tm})\|\right) \\ &+ \left(2 + \frac{2\beta_k^2}{\gamma_k}\right)\alpha_k^2\|f(x_k,v_k)\|^2.\end{aligned}$$

Adding the preceding relations over $k = tm, \ldots, t(m+1)-1$, and using the fact that the stepsizes within one cycle are constant, we obtain

$$\begin{aligned}\sum_{k=tm}^{(t+1)m-1}\|x_k - x_{k+1}\|^2 \le{}& \frac{2\overline{\beta}_t^2}{\overline{\gamma}_t}\left(\|x_{tm} - \Pi x_{tm}\|^2 - \|x_{(t+1)m} - \Pi x_{tm}\|^2 + 2\overline{\alpha}_t \sum_{k=tm}^{(t+1)m-1}\|f(x_k,v_k)'(x_k - \Pi x_{tm})\|\right) \\ &+ \left(2 + \frac{2\overline{\beta}_t^2}{\overline{\gamma}_t}\right)\overline{\alpha}_t^2 \sum_{k=tm}^{(t+1)m-1}\|f(x_k,v_k)\|^2.\end{aligned} \tag{42}$$

Let ϵ be an arbitrary positive scalar. For any $k = tm, \ldots, t(m+1)-1$, we have

$$\begin{aligned}
2\alpha_k \|f(x_k, v_k)'(x_k - \Pi x_{tm})\| &= 2\alpha_k \|f(x_k, v_k)'(x_k - x_{tm}) + f(x_k, v_k)'(x_{tm} - \Pi x_{tm})\| \\
&\leq 2\alpha_k \|f(x_k, v_k)'(x_k - x_{tm})\| + 2\alpha_k \|f(x_k, v_k)'(x_{tm} - \Pi x_{tm})\| \\
&\leq \left(\frac{\alpha_k^2}{\epsilon\gamma_k}\|f(x_k, v_k)\|^2 + \epsilon\gamma_k\|x_k - x_{tm}\|^2\right) + \left(\frac{\alpha_k^2}{\epsilon\gamma_k}\|f(x_k, v_k)\|^2 + \epsilon\gamma_k\|x_{tm} - \Pi x_{tm}\|^2\right) \\
&\leq \frac{2\alpha_k^2}{\epsilon\gamma_k}\|f(x_k, v_k)\|^2 + \epsilon\gamma_k\Delta_t^2 + \epsilon\gamma_k\|x_{tm} - \Pi x_{tm}\|^2,
\end{aligned}$$

where the second inequality uses the fact $2ab \leq a^2 + b^2$ for any real numbers a, b, and the third inequality uses $\|x_k - x_{tm}\| \leq \Delta_t$. By applying the preceding relation to Eq. (42), we obtain

$$\begin{aligned}
\sum_{j=tm}^{(t+1)m-1} \|x_k - x_{k+1}\|^2 &\leq \frac{2\overline{\beta}_t^2}{\overline{\gamma}_t}\left(1 + \epsilon m\overline{\gamma}_t\right)\|x_{tm} - \Pi x_{tm}\|^2 - \frac{2\overline{\beta}_t^2}{\overline{\gamma}_t}\|x_{(t+1)m} - \Pi x_{tm}\|^2 \\
&\quad + 2\overline{\beta}_t^2 m\epsilon\Delta_t^2 + \overline{\alpha}_t^2\left(2 + \frac{2\overline{\beta}_t^2}{\overline{\gamma}_t} + \frac{4\overline{\beta}_t^2}{\epsilon\overline{\gamma}_t^2}\right)\sum_{k=tm}^{(t+1)m-1}\|f(x_k, v_k)\|^2 \qquad (43) \\
&\leq \overline{\beta}_t O\left(1+\epsilon\right)\,\mathrm{d}^2(x_{tm}) + O(m\epsilon)\Delta_t^2 + \overline{\alpha}_t^2 O\left(1 + 1/\epsilon\right)\sum_{k=tm}^{(t+1)m-1}\|f(x_k, v_k)\|^2,
\end{aligned}$$

where the second inequality uses the facts $\mathrm{d}(x_{tm}) = \|x_{tm} - \Pi x_{tm}\|$, $\beta_k/\gamma_k \leq O(1)$, $\beta_k \leq O(1)$ and $\gamma_k \leq O(1)$. By taking ϵ to be sufficiently small so that $O(m\epsilon) < \frac{1}{2m}$ and $O\left(1 + 1/\epsilon\right) \leq O(m^2)$, we obtain

$$\sum_{j=tm}^{(t+1)m-1} \|x_k - x_{k+1}\|^2 \leq O(\overline{\beta}_t)\,\mathrm{d}^2(x_{tm}) + \frac{\Delta_t^2}{2m} + O\left(m^2\overline{\alpha}_t^2\right)\sum_{k=tm}^{(t+1)m-1}\|f(x_k, v_k)\|^2.$$

Combining this relation with the inequality

$$\Delta_t^2 \leq \left(\sum_{k=tm}^{(t+1)m-1}\|x_k - x_{k+1}\|\right)^2 \leq m\left(\sum_{k=tm}^{(t+1)m-1}\|x_k - x_{k+1}\|^2\right),$$

it follows that

$$\Delta_t^2 \leq O(m\overline{\beta}_t)\,\mathrm{d}^2(x_{tm}) + \frac{1}{2}\Delta_t^2 + O\left(m^3\overline{\alpha}_t^2\right)\sum_{k=tm}^{(t+1)m-1}\|f(x_k, v_k)\|^2. \qquad (44)$$

Finally, by taking conditional expectation on both sides and applying Lemma 5, we obtain

$$\mathbf{E}[\Delta_t^2 \mid \mathcal{F}_{tm}] \leq O(m\overline{\beta}_t)\,\mathrm{d}^2(x_{tm}) + \frac{1}{2}\mathbf{E}[\Delta_t^2 \mid \mathcal{F}_{tm}] + m^4\overline{\alpha}_t^2 O\left(\|x_{tm} - x^*\|^2 + 1\right).$$

This implies the desired inequality. ■

The next lemma derives a lower bound for the algorithm's progress towards feasibility within one cycle, and parallels Lemma 4 of the random projection case. Its analysis revolves around properties of cyclic projections, and has a similar flavor as that of [DeH08] Theorem 3.15.

Lemma 7 *Under Assumptions 1-3 and 5, for any $t \geq 0$,*

$$\sum_{k=tm}^{(t+1)m-1}\mathbf{E}[\|z_k - \Pi_{w_k} z_k\|^2 \mid \mathcal{F}_{tm}] \geq \frac{1}{8m\eta}\,\mathrm{d}^2(x_{tm}) - m\overline{\alpha}_t^2\, O\left(\|x_{tm} - x^*\|^2 + 1\right), \qquad w.p.1. \qquad (45)$$

Proof. Let $j \in \{tm, \ldots, (t+1)m-1\}$ be the index that attains the maximum in the linear regularity assumption for x_{tm} (cf. Assumption 2), so that

$$\mathrm{d}^2(x_{tm}) \leq \eta \max_{i=1,\ldots,m} \|x_{tm} - \Pi_{X_i} x_{tm}\|^2 = \eta \|x_{tm} - \Pi_{w_j} x_{tm}\|^2.$$

We have

$$\begin{aligned}
\frac{1}{\sqrt{\eta}}\,\mathrm{d}(x_{tm}) &\leq \|x_{tm} - \Pi_{w_j} x_{tm}\| \\
&\leq \|x_{tm} - \Pi_{w_j} z_j\| \quad \text{(by the definition of } \Pi_{w_j} x_{tm} \text{ and the fact } \Pi_{w_j} z_j \in X_{w_j}) \\
&= \left\| x_{tm} - \frac{1}{\overline{\beta}_t} x_{j+1} + \frac{1-\overline{\beta}_t}{\overline{\beta}_t} z_j \right\| \quad \text{(by the relation } x_{j+1} = z_j - \overline{\beta}_t(z_j - \Pi_{w_j} z_j), \text{ cf. algorithm (40))} \\
&= \left\| \frac{\overline{\beta}_t - 1}{\overline{\beta}_t} \sum_{k=tm}^{j-1} (z_k - x_{k+1}) + \frac{1}{\overline{\beta}_t} \sum_{k=tm}^{j} (z_k - x_{k+1}) - \sum_{k=tm}^{j} (z_k - x_k) \right\| \\
&\leq \left| \frac{\overline{\beta}_t - 1}{\overline{\beta}_t} \right| \sum_{k=tm}^{j-1} \|z_k - x_{k+1}\| + \frac{1}{\overline{\beta}_t} \sum_{k=tm}^{j} \|z_k - x_{k+1}\| + \sum_{k=tm}^{j} \|z_k - x_k\| \\
&\leq \left| \frac{\overline{\beta}_t - 1}{\overline{\beta}_t} \right| \sum_{k=tm}^{(t+1)m-2} \|z_k - x_{k+1}\| + \frac{1}{\overline{\beta}_t} \sum_{k=tm}^{(t+1)m-1} \|z_k - x_{k+1}\| + \sum_{k=tm}^{(t+1)m-1} \|z_k - x_k\| \\
&\leq \frac{2}{\overline{\beta}_t} \sum_{k=tm}^{(t+1)m-1} \|z_k - x_{k+1}\| + \sum_{k=tm}^{(t+1)m-1} \|z_k - x_k\| \quad \text{(since } \overline{\beta}_t \in (0,2)) \\
&= 2 \sum_{k=tm}^{(t+1)m-1} \|z_k - \Pi_{w_k} z_k\| + \overline{\alpha}_t \sum_{k=tm}^{(t+1)m-1} \|f(x_k, v_k)\| \quad \text{(by the definition of algorithm (40))} \\
&\leq \sqrt{2m} \left(4 \sum_{k=tm}^{(t+1)m-1} \|z_k - \Pi_{w_k} z_k\|^2 + \overline{\alpha}_t^2 \sum_{k=tm}^{(t+1)m-1} \|f(x_k, v_k)\|^2 \right)^{1/2},
\end{aligned}$$

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. The preceding relation can be equivalently written as

$$\frac{1}{2m\eta}\,\mathrm{d}^2(x_{tm}) \leq 4 \sum_{k=tm}^{(t+1)m-1} \|z_k - \Pi_{w_k} z_k\|^2 + \overline{\alpha}_t^2 \sum_{k=tm}^{(t+1)m-1} \|f(x_k, v_k)\|^2.$$

By taking expectation on both sides and applying Lemma 5, we obtain Eq. (45). ■

We are ready to present the main result of this section.

Proposition 5 (Convergence of Cyclic Projection Algorithm) *Let Assumptions 1-3 and 5 hold. Then the cyclic projection algorithm* (40) *generates a sequence of iterates* $\{x_k\}$ *that converges almost surely to the unique solution* x^* *of the VI* (1).

Proof. Let $t \geq 0$. By using Lemma 1(a), we have for all k that

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + 2(z_k - x_k)'(x_k - x^*) + \|z_k - x_k\|^2 - \gamma_k \|\Pi_{w_k} z_k - z_k\|^2. \tag{46}$$

In Eq.(46), the cross product term can be decomposed as

$$2(z_k - x_k)'(x_k - x^*) = -2\alpha_k f(x_k, v_k)'(x_k - x^*) = -2\alpha_k f(x_{tm}, v_k)'(x_{tm} - x^*) + 2\alpha_k h_k, \tag{47}$$

where we define

$$h_k = -f(x_k, v_k)'(x_k - x^*) + f(x_{tm}, v_k)'(x_{tm} - x^*). \tag{48}$$

By combining Eqs. (46) and (47), we obtain

$$\|x_{k+1} - x^*\|^2 \le \|x_k - x^*\|^2 - 2\alpha_k f(x_{tm}, v_k)'(x_{tm} - x^*) + 2\alpha_k h_k + \|z_k - x_k\|^2 - \gamma_k \|\Pi_{w_k} z_k - z_k\|^2. \tag{49}$$

We apply Eq. (49) repeatedly for $k = tm, \ldots, (t+1)m - 1$, and obtain

$$\begin{aligned} \|x_{(t+1)m} - x^*\|^2 \le \|x_{tm} - x^*\|^2 - 2\overline{\alpha}_t \left(\sum_{k=tm}^{(t+1)m-1} f(x_{tm}, v_k) \right)' (x_{tm} - x^*) \\ - \overline{\gamma}_t \sum_{k=tm}^{(t+1)m-1} \|\Pi_{w_k} z_k - z_k\|^2 + \sum_{k=tm}^{(t+1)m-1} \left(2\alpha_k h_k + \|z_k - x_k\|^2\right). \end{aligned} \tag{50}$$

We take conditional expectation on both sides of Eq. (50), and then apply Assumption 5(b) and Lemma 7. This yields

$$\mathbf{E}\big[\|x_{(t+1)m} - x^*\|^2 \mid \mathcal{F}_{tm}\big] \le \|x_{tm} - x^*\|^2 - 2m\overline{\alpha}_t F(x_{tm})'(x_{tm} - x^*) - \frac{\overline{\gamma}_t}{8m\eta} \mathrm{d}^2(x_{tm}) + e_t, \tag{51}$$

where we define

$$e_t = \mathbf{E}\left[\sum_{k=tm}^{(t+1)m-1} \left(2\alpha_k h_k + \|z_k - x_k\|^2\right) \;\middle|\; \mathcal{F}_{tm} \right] + m\overline{\alpha}_t^2 \overline{\gamma}_t O\left(\|x_{tm} - x^*\|^2 + 1\right).$$

By Lemma 2 we have

$$-F(x_{tm})'(x_{tm} - x^*) \le -\sigma \|x_{tm} - x^*\|^2 + B\,\mathrm{d}(x_{tm}),$$

so Eq. (51) becomes

$$\mathbf{E}\big[\|x_{(t+1)m} - x^*\|^2 \mid \mathcal{F}_{tm}\big] \le (1 - 2\sigma m\overline{\alpha}_t)\|x_{tm} - x^*\|^2 + 2Bm\overline{\alpha}_t\,\mathrm{d}(x_{tm}) - \frac{\overline{\gamma}_t}{8m\eta}\,\mathrm{d}^2(x_{tm}) + e_t. \tag{52}$$

We will now obtain a bound on the error e_t with the following lemma.

Lemma 8 *Under Assumptions 1-3 and 5, for any $\epsilon > 0$, $t \ge 0$,*

$$e_t \le O\left(\frac{\epsilon \overline{\gamma}_t}{m}\right) \mathrm{d}^2(x_{tm}) + O\left(\frac{m^4 \overline{\alpha}_t^2}{\epsilon}\right) \left(\|x_{tm} - x^*\|^2 + 1\right), \qquad w.p.1.$$

Proof. We note that

$$h_k = -f(x_k, v_k)'(x_k - x^*) + f(x_{tm}, v_k)'(x_{tm} - x^*) = \big(f(x_{tm}, v_k) - f(x_k, v_k)\big)'(x_k - x^*) + f(x_{tm}, v_k)'(x_{tm} - x_k),$$

so that

$$\|h_k\| \le \big\|f(x_{tm}, v_k) - f(x_k, v_k)\big\| \|x_k - x^*\| + \big\|f(x_{tm}, v_k)\big\| \|x_{tm} - x_k\|.$$

By taking conditional expectation of both sides and using the stochastic Lipschitz continuity of $f(\cdot, v)$ (cf. Assumption 1) repeatedly, we have with probability 1,

$$\begin{aligned} \mathbf{E}\big[\|h_k\| \mid \mathcal{F}_k\big] &\le \mathbf{E}\big[\|f(x_{tm}, v_k) - f(x_k, v_k)\| \mid \mathcal{F}_k\big] \|x_k - x^*\| + \mathbf{E}\big[\|f(x_{tm}, v_k)\| \mid \mathcal{F}_k\big] \|x_{tm} - x_k\| \\ &\le L\|x_{tm} - x_k\| \|x_k - x^*\| + (L\|x_{tm} - x^*\| + B)\, \|x_{tm} - x_k\| \\ &\le L\|x_{tm} - x_k\| \left(\|x_{tm} - x^*\| + \|x_{tm} - x_k\|\right) + (L\|x_{tm} - x^*\| + B)\, \|x_{tm} - x_k\| \\ &= B\|x_{tm} - x_k\| + 2L\|x_k - x_{tm}\| \|x_{tm} - x^*\| + L\|x_{tm} - x_k\|^2 \\ &\le B\Delta_t + 2L\Delta_t \|x_{tm} - x^*\| + L\Delta_t^2. \end{aligned}$$

Let $\epsilon > 0$. Then by using the basic inequality $2ab \le a^2 + b^2$ repeatedly and the fact $\alpha_k \to 0$ we obtain

$$\mathbf{E}\big[\alpha_k \|h_k\| \mid \mathcal{F}_k\big] \le O\big(\alpha_k \Delta_t + \alpha_k \Delta_t \|x_{tm} - x^*\| + \alpha_k \Delta_t^2\big) \le O\left((\alpha_k + \epsilon/m^3)\Delta_t^2 + \frac{m^3 \alpha_k^2}{\epsilon}\left(\|x_{tm} - x^*\|^2 + 1\right)\right)$$
$$\le \frac{\epsilon}{m^3} O\left(\Delta_t^2\right) + \frac{m^3 \alpha_k^2}{\epsilon} O\left(\|x_{tm} - x^*\|^2 + 1\right).$$

By applying the preceding bound to the definition of e_t, and then using Lemmas 5 and 6, we obtain

$$\begin{aligned} e_t &\le \frac{\epsilon}{m^2} O\left(\Delta_t^2\right) + \frac{m^4 \overline{\alpha}_t^2}{\epsilon} O\left(\|x_{tm} - x^*\|^2 + 1\right) + m\overline{\alpha}_t^2 (1 + \overline{\gamma}_t) O\big(\|x_{tm} - x^*\|^2 + 1\big) \\ &\le \frac{\epsilon \overline{\beta}_t}{m} O\left(\mathrm{d}^2(x_{tm})\right) + \left(\epsilon m^2 \overline{\alpha}_t^2 + \frac{m^4 \overline{\alpha}_t^2}{\epsilon}\right) O\big(\|x_{tm} - x^*\|^2 + 1\big) \\ &\le O\left(\frac{\epsilon \overline{\gamma}_t}{m}\right) \mathrm{d}^2(x_{tm}) + O\left(\frac{m^4 \overline{\alpha}_t^2}{\epsilon}\right)\big(\|x_{tm} - x^*\|^2 + 1\big). \end{aligned}$$

■

Let us return to the main proof of Prop. 5, and apply Lemma 8 to Eq. (52). We have

$$\begin{aligned} \mathbf{E}\left[\|x_{(t+1)m} - x^*\|^2 \mid \mathcal{F}_{tm}\right] &\le (1 - 2\sigma m \overline{\alpha}_t)\|x_{tm} - x^*\|^2 + 2Bm\overline{\alpha}_t \,\mathrm{d}(x_{tm}) - \frac{\overline{\gamma}_t}{8m\eta}\,\mathrm{d}^2(x_{tm}) \\ &\quad + O\left(\frac{\epsilon \overline{\gamma}_t}{m}\right)\mathrm{d}^2(x_{tm}) + O\left(\frac{m^4 \overline{\alpha}_t^2}{\epsilon}\right)\big(\|x_{tm} - x^*\|^2 + 1\big). \end{aligned} \tag{53}$$

For ϵ and $\overline{\alpha}_t$ sufficiently small, we have

$$2B\overline{\alpha}_t m\,\mathrm{d}(x_{tm}) - \frac{\overline{\gamma}_t}{8m\eta}\,\mathrm{d}^2(x_{tm}) + O\left(\frac{\epsilon \overline{\gamma}_t}{m}\right)\mathrm{d}^2(x_{tm}) \le O\left(\frac{m^3 \overline{\alpha}_t^2}{\overline{\gamma}_t}\right).$$

By summarizing and reordering the terms in Eq. (53), we obtain for some $c_1, c_2 > 0$ that

$$\mathbf{E}\left[\|x_{(t+1)m} - x^*\|^2 \mid \mathcal{F}_{tm}\right] \le \left(1 - 2m\sigma\overline{\alpha}_t + c_1 \frac{m^4 \overline{\alpha}_t^2}{\overline{\gamma}_t}\right)\|x_{tm} - x^*\|^2 + c_2 \frac{m^4 \overline{\alpha}_t^2}{\overline{\gamma}_t}. \tag{54}$$

According to Assumption 3, we have $\sum_{t=0}^{\infty} \overline{\alpha}_t = \infty$, and $\sum_{t=0}^{\infty} \frac{\overline{\alpha}_t^2}{\overline{\gamma}_t} < \infty$. It follows from the Supermartingale Convergence Theorem 1 that $\|x_{tm} - x^*\| \xrightarrow{a.s.} 0$ and $x_{tm} \xrightarrow{a.s.} x^*$ as $t \to \infty$. This also implies that $\mathrm{d}(x_{tm}) \xrightarrow{a.s.} 0$.

There remains to show that $x_k \xrightarrow{a.s.} x^*$ as well. For any $\epsilon > 0$, by using Lemma 5, we have

$$\sum_{k=0}^{\infty} \mathbf{P}\big(\alpha_k \|f(x_k, v_k)\| \ge \epsilon\big) \le \sum_{k=0}^{\infty} \frac{\alpha_k^2 \mathbf{E}\big[\|f(x_k, v_k)\|^2\big]}{\epsilon^2} \le \sum_{k=0}^{\infty} \frac{\alpha_k^2 O\left(\mathbf{E}[\|x_{\lfloor k/m \rfloor m} - x^*\|^2] + 1\right)}{\epsilon^2} \le O\left(\sum_{k=0}^{\infty} \alpha_k^2\right) < \infty.$$

It follows by the Borel-Cantelli lemma that the event $\big\{\alpha_k \|f(x_k, v_k)\| \ge \epsilon\big\}$ cannot happen infinitely often, or equivalently, $\alpha_k \|f(x_k, v_k)\| < \epsilon$ for all k sufficiently large with probability 1. Since ϵ is arbitrary, this further implies that $\alpha_k \|f(x_k, v_k)\| \xrightarrow{a.s.} 0$. Finally, by using the analysis of Lemma 6 [cf. Eq. (44)], we have

$$\Delta_t^2 \le O(m\overline{\beta}_t)\,\mathrm{d}^2(x_{tm}) + O\left(m^3 \overline{\alpha}_t^2\right) \sum_{k=tm}^{(t+1)m-1} \|f(x_k, v_k)\|^2 \xrightarrow{a.s.} 0.$$

Since $x_{tm} \xrightarrow{a.s.} x^*$, we also have

$$\|x_k - x^*\| \le \|x_k - x_{\lfloor k/m \rfloor m}\| + \|x_{\lfloor k/m \rfloor m} - x^*\| \le \Delta_{\lfloor k/m \rfloor} + \|x_{\lfloor k/m \rfloor m} - x^*\| \xrightarrow{a.s.} 0.$$

Therefore $x_k \xrightarrow{a.s.} x^*$ as $k \to \infty$. ■

4.2 Convergence Rate and Constant Stepsize Error Bound

Now we consider the rate of convergence of the cyclic projection algorithm. We will derive convergence rate results for both the case of diminishing stepsizes α_k and the case of constant stepsizes α.

Proposition 6 (Cyclic Projection Algorithm: Convergence Rate for Diminishing $\{\alpha_k\}$) *Let Assumptions 1-3 and 5 hold, let $\{\alpha_k\}$ be bounded by a sufficiently small positive scalar, and let $\{x_k\}$ be generated by the cyclic projection algorithm (40). For any positive scalar ϵ, there exists a random variable N such that*

$$\min_{0\le k\le N}\left\{\|x_k - x^*\|^2 - \delta_k\right\} \le \epsilon, \qquad w.p.1,$$

where $\delta_k = O\left(\frac{m^3\alpha_k}{\sigma\gamma_k}\right)$, and N satisfies

$$\mathbf{E}\left[\sum_{k=0}^{N}\alpha_k\right] \le \frac{\|x_0 - x^*\|^2}{2\sigma\epsilon}.$$

Proof. The analysis is very similar to that of Prop. 2. Let δ_k be given by

$$\delta_k = \frac{\left(c_1 m^4\epsilon + c_2 m^4\right)\alpha_k/\gamma_k}{2m\sigma - c_1 m^4\alpha_k/\gamma_k},$$

where c_1, c_2 are the constants from Eq. (54). For α_k sufficiently small, we have $2m\sigma - c_1 m^4\alpha_k/\gamma_k > 0$ so that $\delta_k > 0$. It can be seen that $\delta_k \le O\left(\frac{m^3\alpha_k}{\sigma\gamma_k}\right)$.

Define a new process $\{\hat{x}_k\}$ which is identical to $\{x_k\}$ except that once $\hat{x}_k$ enters the level set

$$L_k = \left\{x \in \Re^n \mid \|x - x^*\|^2 \le \delta_k + \epsilon\right\},$$

the process stays at $\hat{x}_k = x^*$ for all future k. According to Eq. (54), we have

$$\mathbf{E}[\|\hat{x}_{(t+1)m} - x^*\|^2 \mid \mathcal{F}_{tm}] \le \left(1 - 2m\sigma\overline{\alpha}_t + c_1\frac{m^4\overline{\alpha}_t^2}{\overline{\gamma}_t}\right)\|\hat{x}_{tm} - x^*\|^2 + c_2\frac{m^4\overline{\alpha}_t^2}{\overline{\gamma}_t}.$$

This is equivalent to

$$\mathbf{E}\left[\|\hat{x}_{(t+1)m} - x^*\|^2 \mid \mathcal{F}_{tm}\right] \le \|\hat{x}_{tm} - x^*\|^2 - \xi_t, \tag{55}$$

where we define

$$\xi_t = \begin{cases} \left(2m\sigma\overline{\alpha}_t - c_1\frac{m^4\overline{\alpha}_t^2}{\overline{\gamma}_t}\right)\|\hat{x}_k - x^*\|^2 - c_2\frac{m^4\overline{\alpha}_t^2}{\overline{\gamma}_t} & \text{if } \hat{x}_{tm} \notin L_{tm}, \\ 0 & \text{otherwise.} \end{cases}$$

When $\hat{x}_{tm} \notin L_{tm}$, we can verify by using the definition of δ_t that

$$\xi_t \ge (2\sigma\epsilon m)\overline{\alpha}_t.$$

Hence we have $\xi_t \ge 0$ for all t. By applying Theorem 1 to Eq. (55), we have $\sum_{t=0}^{\infty}\xi_t < \infty$ with probability 1. Therefore ξ_t must terminate at 0 for t sufficiently large with probability 1.

Let N be the smallest integer such that $\hat{x}_k \in L_k$ for all $k \ge N$ with probability 1, implying that $\hat{x}_{m\lceil N/m\rceil} \in L_{m\lceil N/m\rceil}$. We have for all t that

$$\mathbf{E}\left[\|\hat{x}_{tm} - x^*\|^2 \mid \mathcal{F}_k\right] \le \|\hat{x}_0 - x^*\|^2 - \mathbf{E}\left[\sum_{k=0}^{t}\xi_k\right].$$

By letting $t \to \infty$ and using the monotone convergence theorem, we obtain

$$\|x_0 - x^*\|^2 \geq \mathbf{E}\left[\sum_{t=0}^{\infty} \xi_t\right] = \mathbf{E}\left[\sum_{t=0}^{\lceil N/m \rceil} \xi_t\right] \geq (2m\sigma\epsilon)\mathbf{E}\left[\sum_{t=0}^{\lceil N/m \rceil} \overline{\alpha}_t\right].$$

Finally, we have

$$\mathbf{E}\left[\sum_{k=0}^{N} \alpha_k\right] \leq \mathbf{E}\left[\sum_{t=0}^{\lceil N/m \rceil} (m\overline{\alpha}_t)\right] \leq \frac{\|x_0 - x^*\|^2}{2\sigma\epsilon}.$$

■

The next proposition gives an error bound for cyclic projection algorithms with constant stepsizes. It is an almost immediate corollary of Prop. 6.

Proposition 7 (Cyclic Projection Algorithm: Error Bound for Constant $\{\alpha_k\}$ and $\{\beta_k\}$) *Let Assumptions 1, 2, and 5 hold, let the stepsizes be constant scalars satisfying*

$$\alpha_k = \alpha > 0, \qquad \beta_k = \beta \in (0,2), \qquad \gamma_k = \gamma = \beta(2-\beta), \qquad \forall\ k \geq 0,$$

where α is a sufficiently small scalar, and let $\{x_k\}$ be generated by the cyclic projection algorithm (40). *Then*

$$\liminf_{k\to\infty} \|x_k - x^*\|^2 \leq O\left(\frac{m^3\alpha}{\sigma\gamma}\right), \qquad w.p.1.$$

For any positive scalar ϵ, there exists a random variable N such that

$$\min_{1\leq k\leq N} \|x_k - x^*\|^2 \leq \epsilon + O\left(\frac{m^3\alpha}{\sigma\gamma}\right), \qquad w.p.1,$$

where N satisfies

$$\mathbf{E}[N] \leq \frac{\|x_0 - x^*\|^2}{\big(2\sigma - O\left(m^3\alpha/\gamma\right)\big)\epsilon\alpha}.$$

Proof. The proof is identical with that of Prop. 6. ■

Let us compare Props. 3 and 7. In a comparable expected number of iterations, the cyclic algorithm converges to within an error tolerance of the order of m^3, while the random projection algorithm converges to within an error bound that is of the order of m. This suggests an advantage of the random projection algorithm. Intuitively, the analysis shows that the cyclic algorithm may incur an accumulating error within one cycle, due to the correlation of the random process $\big\{(w_k, v_k)\big\}$ within the cycle and across cycles. Of course, the preceding comparison is based on upper bound estimates, and to some extent may be an artifact of our method of analysis. However, the superiority of the random sampling approach over the deterministic cyclic sampling approach is supported by the computational results of the next section, and is consistent with related analyses for incremental subgradient and proximal methods (e.g., [NeB01], [Ber10]).

5 Applications and Computational Experiments

Our algorithm is particularly well-suited for VIs with many linear constraints. As an example consider a linear complementarity problem with

$$F(x) = Ax - b, \qquad X = \{x \in \Re^n \mid Cx \leq d\},$$

where A is an $n \times n$ positive definite matrix, b is a vector in $\Re^n$, C is an $m \times n$ matrix, and d is a vector in $\Re^m$. The constraint set X is an intersection of halfspaces X_i given by

$$X_i = \{x \in \Re^n \mid c_i'x \leq d_i\}, \qquad i = 1, \ldots, m,$$

where c_i' is the ith row of C, and d_i is the ith entry of d. In this case, assuming that $Ax_k - b$ is computed exactly without sampling, our algorithm becomes

$$z_k = x_k - \alpha_k(Ax_k - b), \qquad x_{k+1} = z_k - \beta_k \frac{\max\{c_{i_k}'z_k - d_i, 0\}}{\|c_{i_k}\|^2} c_{i_k}.$$

Thus the set projection portion of the algorithm is very simple.

Linear complementarity problems with a large number of constraints arise among others in important approximate dynamic programming contexts (see e.g., the books [BeT96], [SuB98], and [Ber12]), which motivated in part our work. In one such context, arising in approximate policy evaluation, we aim to approximate the solution of a high-dimensional linear fixed point equation $y = Ay + b$, where A is an $n \times n$ matrix and $b \in \Re^n$, by approximation over a low-dimensional subspace $S = \{\Phi x \mid x \in \Re^s\}$, where Φ is an $n \times s$ matrix (with $s \ll n$) whose columns can be viewed as basis functions for S. In the Galerkin approximation approach (see e.g., Krasnoselskii et al. [Kra72], Saad [Saa03]), the high-dimensional problem $y = Ay + b$ is replaced by the low-dimensional fixed point problem

$$\Phi x = \Pi_S(A\Phi x + b).$$

Here Π_S denotes weighted Euclidean projection onto S, where the projection norm being $\|x\| = \sqrt{x'\Xi x}$, where Ξ is a positive definite symmetric matrix. Aside from classical applications in solving large-scale problems arising from discretization of partial differential equations or from inverse problems, this approach (in combination with randomization and simulation) is central in popular approximate dynamic programming methods, known as projected equation or temporal difference methods, as well as in aggregation methods (see [Ber12]).

In a constrained variant of the Galerkin approach one may improve the quality of approximation if the solution of the original fixed point problem is known (or is required) to belong to some given closed convex set C. Then it typically makes sense to impose the additional constraint $\Phi x \in C$, thereby giving rise to the problem of finding x such that

$$\Phi x = \Pi_{S \cap C}(A\Phi x + b).$$

This constrained projected equation has been discussed in [Ber11b], and was shown to be equivalent to the VI

$$\big((I - A)\Phi x^* + b\big)'\Xi\Phi(x - x^*) \geq 0, \qquad \forall\, x \in X \stackrel{\text{def}}{=} \{x \mid \Phi x \in C\}.$$

A serious difficulty for its solution by projection methods is that while the dimension of x may be small, the constraint set X often consists of the intersection of a large number of constraints. This difficulty also arises in approximate linear programming methods, another major approach for approximate dynamic programming (see e.g, de Farias and Van Roy [FaV03], [FaV04], and Desai et al. [DFM12]). Our proposed method in this paper addresses effectively this difficulty, by using incremental projections on simpler supersets of X.

We will now describe the results of computational experimentation with our method. The test problem is an example based on the constrained Galerkin approximation approach just described.

Example 1: We want to compute a low-dimensional approximation to the invariant distribution ξ of an ergodic 1000-state Markov chain with transition probability matrix P. The approximation has the form Φx, where Φ is an 1000×20 matrix and x is a vector in $\Re^{20}$. We approximate the equation $\xi = P'\xi$ characterizing the invariant distribution by using its projected version

$$\Phi x = \Pi P'\Phi x,$$

where Π denotes the weighted orthogonal projection onto the set of distribution vectors

$$\{\Phi x \mid x \in \Re^{20}, \quad \Phi x \geq 0, \quad e'\Phi x = 1\}$$

with weight vector ξ (other Euclidean projection norms could also be used), e is the vector in $\Re^n$ with all components equal to 1 . As noted earlier, the projected equation is equivalent to the VI

$$(x - x^*)'Ax^* \geq 0, \qquad \forall\, x \in \Re^{20} \;\; \text{s.t.} \;\; \Phi x \geq 0, \quad e'\Phi x = 1, \tag{56}$$

where A takes the form

$$A = \Phi'\Xi(I - P')\Phi,$$

with Ξ being the diagonal matrix with the components of the vector ξ along the diagonal. Note here that there are efficient methods for calculating the matrix A by simulation and low-dimensional calculation (see e.g., [Ber11b]) - such methods could be used to calculate a close approximation to A prior to applying our algorithm to VI (56). Throughout our experiments we assume that A is known. We have chosen the columns of Φ to be sine functions of various frequencies together with the unit vector, and have chosen ξ to be an arbitrary distribution vector (so ξ may not belong to the range of Φ). Figure 1 plots the approximate distribution Φx^*, obtained as the optimal solution of VI (56), and compares it with the underlying true distribution ξ.

To evaluate the proposed incremental projection algorithms, we have experimented with different choices of the stepsizes α_k and β_k, as illustrated in Fig. 2. In this experiment, we have used $f(x_k, v_k) = F(x_k) = Ax_k$, and have also used uniformly distributed independent samples of the constraint set indexes. The left side of Fig. 2 plots $\|x_k - x^*\|$ and $\mathrm{d}(x_k)$ in the cases where $\beta_k = 1$ and $\beta_k = 1/\log k$, with $\alpha_k = k^{-0.55}$ in both cases. The comparison between the two cases indicates an advantage for using a constant β over a diminishing β_k. The right side of Fig. 2 plots the trajectories of iteration errors and feasibility errors in the case where $\alpha_k = k^{-1}$ and in the case where $\alpha_k = k^{-0.55}$, with $\beta_k = 1$ in both cases. Again, the iteration with the larger stepsizes, i.e. $\alpha_k = k^{-0.55}$, converges faster than the iteration with the smaller stepsizes.

The next experiment is to compare the constraint sampling schemes. More specifically, we have tested the independent uniform sampling scheme against the deterministic cyclic sampling scheme, while using $f(x_k, v_k) = F(x_k) = Ax_k$, $\alpha_k = k^{-0.55}$, and $\beta_k = 1$ throughout. As illustrated in Fig. 3, the algorithm that uses random/independent samples converges much faster than the algorithm using deterministic cyclic samples. We have repeated this experiment with other choices of stepsizes, and have observed similar phenomena. These observations are consistent with our analysis in Sections 3 and 4, and support our argument that random sampling is preferable over deterministic cyclic sampling. We have also experimented with the alternative of randomly shuffling the constraint indexes at the beginning of each cycle. This type of constraint sampling is more similar to independent random sampling, and gave comparable results in our experiments (not reported here). This is consistent with earlier observations and analysis by Recht and Re [ReR12], which suggest that the performance of cyclic sampling in incremental subgradient methods is enhanced if the components of the cost function are randomly reshuffled at the beginning of each cycle.

Finally, we have experimented with all possible combinations of random independent sampling and deterministic cyclic sampling, for both the component functions and the constraint sets. The results are plotted in Fig. 4. The "batch" case, using $f(x_k, v_k) = F(x_k) = Ax_k$ and independent uniform samples of the constraint sets, has the fastest rate of convergence. However, for large scale incremental problems, the computation of $F(x_k)$ requires a time complexity on the order of the number of component functions, which makes each iteration of the algorithm very expensive. On the other hand, when F is linear, one may replace the matrix A and the vector b defining F with a simulation-based approximation, in which case the time complexity is reduced. As noted earlier in connection with Galerkin approximation, methods of this type are popular in simulation-based approximate dynamic programming (see [Ber12] for textbook treatment), and lead to more efficient computation than stochastic approximation methods. Methods of this type have also been considered in simulation-based nonlinear optimization, where they are known as sample average approximation methods (see Shapiro, Dentcheva, and Ruszczynski [SDR09] for a recent textbook treatment, and Nemirovski et al. [NJLS09]).

In the remaining four cases of Fig. 4, we consider A as an average of a large number of matrices

$$A = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}, \qquad \text{where} \quad A_{ij} = n^2\xi_i\phi_i(\phi_i - p_{ji}\phi_j)',$$

where ξ_i denotes the ith entry of ξ, p_{ij} denotes the (i,j)th entry of P, and ϕ_i' denotes the ith row of Φ. We will use A_{v_k}, where v_k are sample index pairs (i,j), as samples of A. The last four cases in Fig. 4 use one sample component function $f(x_k, v_k) = A_{v_k} x_k$ per iteration. Among these cases, it can be seen that random sampling generally performs better than cyclic sampling. This is particularly so for constraint sampling, since according to Fig. 4, the two cases using random/independent samples of constraints have much better convergence properties than the other two cases using cyclic samples of constraints.

6 Concluding Remarks

In this paper we have proposed new algorithms for strongly monotone variational inequalities with structure that lends itself to constraint and function sampling. We analyzed the convergence properties of various types of sampling, and we established a substantial rate of convergence advantage for random sampling over cyclic sampling. Our cyclic sampling rule for constraints requires that each constraint is sampled exactly once in a cycle, and allows a lot of freedom on how the constraint indexes are ordered within each cycle; our convergence rate result applies to the worst case. It is therefore possible that a cyclic rule with separate randomization within each cycle yields a performance close to the one of the independent uniform sampling method, and superior to a deterministic cyclic rule; this was observed in the experiments described Section 5. We also note that constraint sampling rules that sample constraints adaptively based on their "importance" and the progress of the algorithm may yield even better performance, and that this is an interesting subject for investigation.

A potential direction of further research is to relax the strong monotonicity assumption on F, either by assuming a special structure or by modification of our algorithms. For example, if F is the gradient of a convex function, the projection method as well as the related methods of Bertsekas [Ber11a] and Nedić [Ned11] do not require strong convexity of the cost function (or equivalently, strong monotonicity of the gradient). Another interesting case arises when X is polyhedral and $F(x) = \Phi' \bar{F}(\Phi x)$, where $\bar{F}$ is strongly monotone but Φ is singular (cf. classical applications in network traffic assignment problems). The convergence of the projection method for this case was shown by Bertsekas and Gafni [BeG82]. Another possibility is to consider the extragradient method of Korpelevich [Kor76] or the recent iterative Tikhonov regularization method and the iterative proximal point method of Kannan et al. [KNS12], which are modifications of the projection method to deal with VIs that are not necessarily strongly monotone.

References

[BaB96] Bauschke, H. H., and Borwein, J. M., 1996. "On projection algorithms for solving convex feasibility problems," SIAM Review, Vol. 38, pp. 367-426.

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

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

[BeG82] Bertsekas, D. P., and Gafni, E. M., 1982. "Projection methods for variational inequalities with application to the traffic assignment problem," Math. Progr. Studies, Vol. 17, pp. 139-159.

[Ber10] Bertsekas, D. P., 2010. "Incremental gradient, subgradient, and proximal methods for convex optimization: a survey," Lab. for Information and Decision Systems Report LIDS-P-2848, MIT; an extended version of a paper in the edited volume "Optimization for Machine Learning," by S. Sra, S. Nowozin, and S. J. Wright, MIT Press, Cambridge, MA, pp. 85-119.

[Ber11a] Bertsekas, D. P., 2011. "Incremental proximal methods for large scale convex optimization," Mathematical Programming, Ser. B, Vol. 129, pp. 163-195.

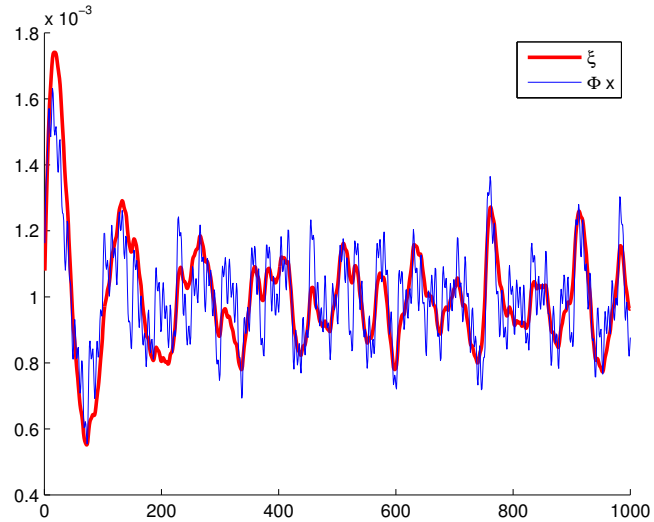

Figure 1: Estimated distribution Φx^* compared against the true invariant distribution ξ (Example 1), with $\Phi x^*, \xi \in \Re^{1000}$.

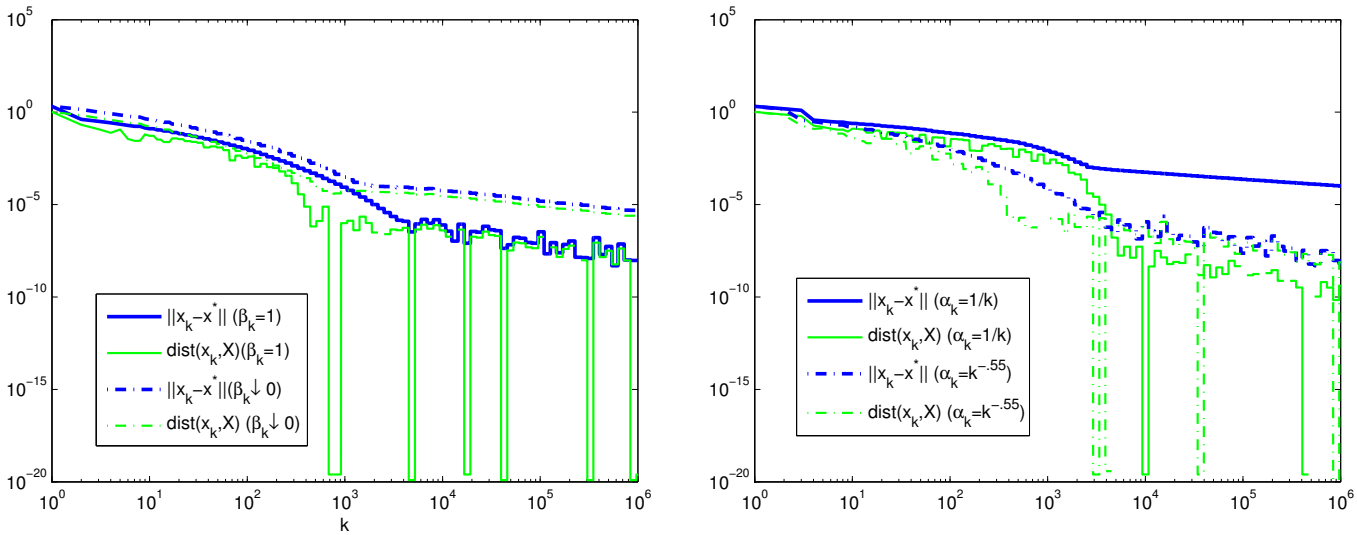

Figure 2: Comparison of different choices of stepsizes $\{\alpha_k\}$ and $\{\beta_k\}$ (Example 1). The left figure plots the trajectories of iteration errors and feasibility errors with $\beta_k = 1$ and $\beta_k = 1/\log k \downarrow 0$, while fixing $\alpha_k = k^{-0.55}$. The right figure plots the trajectories of iteration errors and feasibility errors with $\alpha_k = k^{-1}$ and $\alpha_k = k^{-0.55}$, while fixing $\beta_k = 1$. In both figures, we use $f(x_k, v_k) = Ax_k$ and independent uniformly distributed samples of the constraint sets.

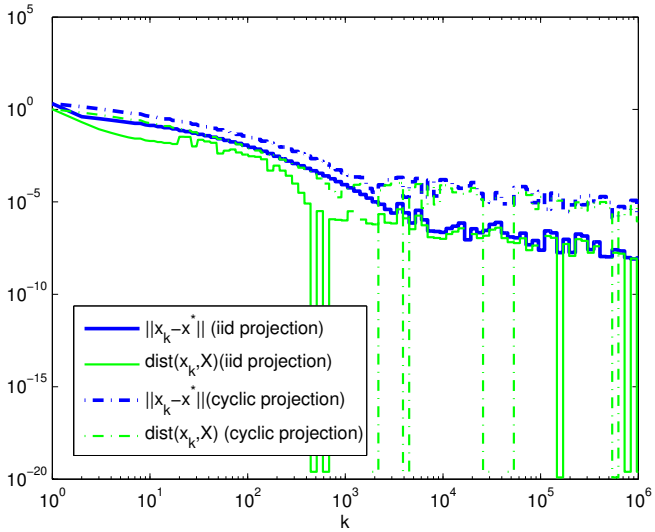

Figure 3: Comparison between independent uniformly distributed and deterministic cyclic orders of constraint sampling (Example 1), with $\alpha_k = k^{-0.55}$, $\beta_k = 1$, and $f(x_k, v_k) = Ax_k$ for all k.

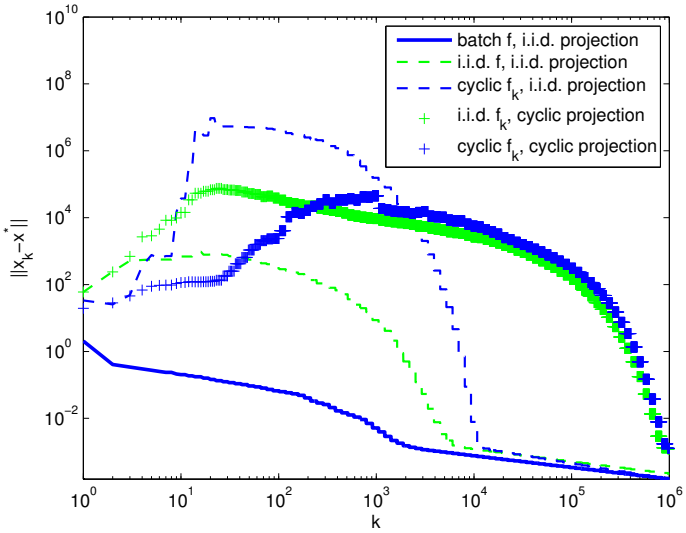

Figure 4: Combinations of independent uniformly distributed and deterministic cyclic orders of function and constraint sampling (Example 1). In the case of "batch f", we use $f(x_k, v_k) = Ax_k$. In the case of "i.i.d. f", we use $f(x_k, v_k) = A_{v_k} x_k$, where A_{v_k} are i.i.d. random variables with mean A. In the case of "cyclic f", we use $f(x_k, v_k) = A_{v_k} x_k$, where A_{v_k} are cyclic samples such that their empirical mean over one cycle equals to A. In all five cases, we use $\alpha_k = 1/k$ and $\beta_k = 1$ for all k.

[Ber11b] Bertsekas, D. P., 2011. "Temporal difference methods for general projected equations," IEEE Trans. on Automatic Control, Vol. 56, pp. 2178-2189.

[Ber12] Bertsekas, D. P., 2012. Dynamic Programming and Optimal Control Vol. II: Approximate Dynamic Programming, Athena Scientific, Belmont, M. A.

[BeT89] Bertsekas, D. P., and Tsitsiklis, J. N., 1989. Parallel and Distributed Computation: Numerical Methods, Prentice-Hall, Englewood Cliffs, N. J.; republished by Athena Scientific, Belmont, M. A, 1997.

[BeT96] Bertsekas, D. P., and Tsitsiklis, J. N., 1996. Neuro-Dynamic Programming, Athena Scientific, Belmont, M. A.

[BeT00] Bertsekas, D. P., and Tsitsiklis, J. N., 2000. "Gradient convergence in gradient methods," SIAM J. Optimization, Vol. 10, pp. 627-642.

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

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

[CeG08] Censor, Y., and Gibali, A., 2008. "Projections onto super-half-spaces for monotone variational inequality problems in finite-dimensional spaces," J. of Nonlinear and Convex Analysis, Vol. 9, pp. 461-475.

[CeS08] Cegielski, A., and Suchocka, A., 2008. "Relaxed alternating projection methods," SIAM J. Optimization, Vol. 19, pp. 1093-1106.

[CGRS12a] Censor, Y., Gibali, A., Reich, S., and Sabach, S., 2012. "Common solutions to variational inequalities," Set-Valued and Variational Analysis, Vol. 20, pp. 229-247.

[CGRS12b] Censor, Y., Gibali, A., and Reich, S., 2012. "A von Neumann alternating method for finding common solutions to variational inequalities," Nonlinear Analysis Series A: Theory, Methods & Applications, Vol. 75, pp. 4596-4603.

[DeH06a] Deutsch, F., and Hundal, H., 2006. "The rate of convergence for the cyclic projections algorithm I: angles between convex sets," J. of Approximation Theory, Vol. 142, pp. 36-55.

[DeH06b] Deutsch, F., and Hundal, H., 2006. "The rate of convergence for the cyclic projections algorithm II: norms of nonlinear operators," J. of Approximation Theory, Vol. 142, pp. 56-82.

[DeH08] Deutsch, F., and Hundal, H., 2008. "The rate of convergence for the cyclic projections algorithm III: regularity of convex sets," J. of Approximation Theory, Vol. 155, pp. 155-184.

[DFM12] Desai, V. V., Farias, V. F., and Moallemi, C. C., 2012. "Approximate dynamic programming via a smoothed approximate linear program," Operations Research, Vol. 60, pp. 655-674.

[FaP03] Facchinei, F., and Pang, J. S., 2003. Finite-Dimensional Variational Inequalities and Complementarity Problems, Vols. I and II, Springer-Verlag, New York.

[Fuk86] Fukushima, M., 1986. "A relaxed projection method for variational inequalities," Mathematical Programming, Vol. 35, pp. 58-70.

[FaV03] de Farias, D. P., and Van Roy, B., 2003. "The linear programming approach to approximate dynamic programming," Operations Research, Vol. 51, pp. 850-865.

[FaV04] de Farias, D. P., and Van Roy, B., 2004. "On constraint sampling in the linear programming approach to approximate dynamic programming," Mathematics of Operations Research, Vol. 29, pp. 462-478.

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

[GOR99] Gürkan, G., Ozge, A. Y., and Robinson, S. M., 1999. "Sample-path solution of stochastic variational inequalities", Mathematical Programming, Vol. 84, pp. 313-333.

[Hal62] Halperin, I., 1962. "The product of projection operators," Acta Scientiarum Mathematicarum, Vol. 23, pp. 96-99.

[HaS66] Hartman, P., and Stampacchia, G., 1966. "On some non-linear elliptic differential functional equations," Acta Mathematica, Vol. 115, pp. 271-310.

[LeL10] Leventhal, D., and Lewis, A. S., 2010. "Randomized methods for linear constraints: convergence rates and conditioning," Mathematics of Operations Research, Vol. 35, pp. 641-654.

[LeM08] Lewis, A. S., and Malick, J., 2008. "Alternating projections on manifolds," Mathematics of Operations Research, Vol. 33, pp. 216-234.

[JiX08] Jiang, H., and Xu, F., 2008. "Stochastic approximation approaches to the stochastic variational inequality problem," IEEE Trans. on Automatic Control, Vol. 53, pp. 1462-1475.

[KaS13] Kannan, A., and Shanbhag, U. V., 2013. "Distributed online computation of Nash equilibria via iterative regularization techniques," SIAM J. Optimization (to appear).

[KiS80] Kinderlehrer, D., and Stampacchia, G., 1980. An Introduction to Variational Inequalities and Their Applications, Academic Press, New York-London.

[KNS12] Koshal, J., Nedić, A., and Shanbhag, U. V., 2012. "Regularized iterative stochastic approximation methods for stochastic variational inequality problems," IEEE Trans. on Automatic Control (to appear).

[Kor76] Korpelevich, G. M., 1976. "An extragradient method for finding saddle points and for other problems," Matecon, Vol. 12, pp. 747-756.

[Kra72] Krasnoselskii, M. A., 1972. Approximate Solution of Operator Equations, translated by Louvish, D., Wolters-Noordhoff Pub., Groningen.

[KuY03] Kushner, H. J., and Yin, G., 2003. Stochastic Approximation Methods, 2nd Edition, Springer-Verlag, N. Y.

[NeB00] Nedić, A., and Bertsekas, D. P., 2000. "Convergence rate of the incremental subgradient algorithm," in Stochastic Optimization: Algorithms and Applications, by S. Uryasev and P. M. Pardalos Eds., Kluwer Academic Publishers, pp. 263-304.

[NeB01] Nedić, A., and Bertsekas, D. P., 2001. "Incremental subgradient methods for nondifferentiable optimization," SIAM J. Optimization, Vol. 12, pp. 109-138.

[Ned10] Nedić, A., 2010. "Random projection algorithms for convex set intersection problems," the 49th IEEE Conference on Decision and Control, Atlanta, Georgia, pp. 7655-7660.

[Ned11] Nedić, A., 2011. "Random algorithms for convex minimization problems," Mathematical Programming, Ser. B, Vol. 129, pp. 225-253.

[Noo04] Noor, M. A., 2004."Some developments in general variational inequalities," Applied Mathematics and Computation, Vol. 152, pp. 197-277.

[NJLS09] Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A., 2009. "Robust stochastic approximation approach to stochastic programming," SIAM J. Optimization, Vol. 19, pp. 1574-1609.

[Pat99] Patriksson, M., 1999. Nonlinear Programing and Variational Inequality Problems: A Unified Approach, Kluwer Academic Publishers, Dordrecht, The Netherlands.

[ReR12] Recht, B., and Re, C., 2012. "Beneath the valley of the noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences," arXiv:1202.4184.

[RoS71] Robbins, H., and Siegmund, D. O., 1971. "A convergence theorem for nonnegative almost supermartingales and some applications," Optimizing Methods in Statistics, pp. 233-257.

[Saa03] Saad, Y., 2003. Iterative Methods for Sparse Linear Systems, SIAM, Philadelphia, P. A.

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

[Sib70] Sibony M., 1970. "Methodes iteratives pour les equations et inequations aux derivees partielles non lineaires de type monotone," Calcolo, Vol. 7, pp. 65-183.

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

[vNe50] von Neumann, J., 1950. Functional Operators, Princeton University Press, Princeton, N. J.

[SuB98] Sutton, R. S., and Barto, A. G., 1998. Reinforcement Learning, MIT Press, Cambridge, M. A.

[XiZ03] Xiu, N., and Zhang, J., 2003. "Some recent advances in projection-type methods for variational inequalities," J. Computational and Applied Mathematics, Vol. 152, pp. 559-585.