Approximate projection methods for monotone stochastic variational inequalities

A. N. Iusem* A. Jofré† P. Thompson‡

June 4, 2014

Abstract

We consider stochastic variational inequalities with monotone operators. The operator F defining the variational inequality depends both on a variable in the finite dimensional Euclidean space and on a random variable. We are interested in finding solutions for the deterministic variational inequality problem whose operator T is defined as the expected value of F, but we do not assume that T is explicitly available; rather we propose a Stochastic Approximation procedure, meaning that at each iteration, a step similar to some variant of the deterministic projection method is taken after sampling the random variable, choosing thus a specific realization of the operator. We consider two variants of the method where the exact orthogonal projection step is replaced by an approximate one. The first variant is a projection method with approximate projections, where the variational inequality satisfies an error bound on the solution set, called *weak sharpness*. We prove that the generated sequence is bounded and its distance to the solution set converges to zero almost surely. In particular, every cluster point of the sequence is, almost surely, a solution. For the case in which the feasible set is compact, we establish a convergence rate and an estimate on the number of iterations required so that any solution of an auxiliary linear program solves the variational inequality. The second variant is an iterative Tykhonov regularization method with approximate projections where, instead of solving a sequence of regularized variational inequality problems, the regularization parameter is updated in each iteration and a single projection step associated with the regularized problem is taken. In this second method, we allow a Cartesian structure on the variational inequality so as to encompass, for example, equilibrium conditions of monotone stochastic Nash games with a limited coordination between the players' stepsize and regularization sequences. Requiring just monotonicity, we prove that the generated sequence converges to the least-norm solution of the variational inequality almost surely.

Key words: Stochastic variational inequalities, projection method, stochastic approximation

Mathematical Subject Classification (2000): 90C15,90C30.

1 Introduction

The standard (deterministic) variational inequality problem, which we will denote as $\mathrm{VI}(T, X)$, is defined as follows: given a closed and convex set $X \subset \mathbb{R}^n$ and an operator $T : \mathbb{R}^n \to \mathbb{R}^n$, find $x^* \in X$ such that

$$\langle T(x^*), x - x^* \rangle \geq 0, \tag{1}$$

for all $x \in X$. The variational inequality problem includes many interesting special cases, such as complementarity and optimization problems.

In the stochastic case, we start with a mapping F which depends not only on $x \in \mathbb{R}^n$ but also on a random variable $v : \Omega \to \Xi$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. So F is of the form $F : \Xi \times \mathbb{R}^n \to \mathbb{R}^n$,

*Instituto de Matemática Pura e Aplicada (IMPA), Rio de Janeiro, RJ, Brazil, iusp@impa.br
†Centro de Modelamiento Matemático, Universidad de Chile, Santiago de Chile, Chile, ajofre@dim.uchile.cl
‡Instituto de Matemática Pura e Aplicada (IMPA), Rio de Janeiro, RJ, Brazil, philipthomp@gmail.com

where for every $x \in \mathbb{R}^n$, $F(\cdot, x) : \Xi \to \mathbb{R}^n$ is an integrable random vector on a measurable space $(\Xi, \mathcal{G})$. The solution crierion analyzed in this paper consists of solving VI(T, X) as defined by (1), where $T : \mathbb{R}^n \to \mathbb{R}^n$ is the expected value of $F(v, \cdot)$, i.e.

$$T(x) = \mathbb{E}[F(v, x)]. \tag{2}$$

In this special case, problem (1)-(2) becomes the *stochastic variational inequality* problem (SVI). We remark that a random solution of the SVI is allowed, i.e., a random variable $x^* : \Omega \to X$ that satisfies (1) almost surely. Recently, [10] considered a more general definition of stochastic variational inequality where the feasible set is also affected by randomness, that is, $X : \Xi \rightrightarrows \mathbb{R}^n$ is a random set-valued function.

Methods for the deterministic VI(T, X) have been extensively studied (see [12]). If T is fully available then SVI can be solved by these methods. However, knowledge of T is often scarce, for various reasons:
a) the probability distribution of v is known, but the calculation of the expected value (2) involves multi-dimensional integration which is computationally expensive if not impossible,
b) the random function F is known but the distribution of v is not, so that the information on v can be only obtained using past data or sampling,
c) $\mathbb{E}[F(v, x)]$ is not observable and must be approximately evaluated with some simulation procedure.

In these cases new methods are required for the stochastic counterpart SVI, that involve a statistical analysis. When $X = \mathbb{R}^n$, the VI(T,X) defined by (1) becomes the problem of finding the zeroes of T: find $x^* \in \mathbb{R}^n$ such that

$$T(x^*) = 0. \tag{3}$$

When T is given by (2), problem (3) is called the *stochastic non-linear equation* problem (SE). Robbins and Monro proposed in [24] a *stochastic approximation* (SA) method for solving SE which uses directly the *random map* F which is fully available, in contrast with the mean operator T. Since this fundamental work, SA approaches to various stochastic optimization problems and, more recently, to stochastic variational inequalities, have been studied (see references in [16]).

1.1 Deterministic projection methods

In the deterministic setting (1), the classical projection method for VI(T, X), akin to the projected gradient method for convex optimization, is

$$x^{k+1} = \Pi[x^k - \alpha_k T(x^k)] \tag{4}$$

where Π is the projection operator onto X and $\{\alpha_k\}$ is an exogenous sequence of positive stepsizes. If T is *strongly monotone*, i.e.,

$$\langle T(z) - T(x), z - x\rangle \geq \sigma\|z - x\|^2 \tag{5}$$

for all $x, z \in \mathbb{R}^n$ and some $\sigma > 0$, and also *Lipschitz continuous*, i.e.,

$$\|T(z) - T(x)\| \leq L\|z - x\|$$

for all $x, z \in \mathbb{R}^n$ and some $L > 0$, then the method converges to the (unique) solution of VI(T, X) (see [5]), assuming that the stepsizes satisfy

$$\alpha_k \in (0, 2\sigma/L^2) \ \ \forall k \in \mathbb{N}, \tag{6}$$

$$\inf_k \alpha_k > 0 \ \text{ or } \ \sum_k \alpha_k = \infty. \tag{7}$$

This choice of stepsizes includes the possibility of a constant stepsize $\alpha_k \equiv \alpha \in (0, 2\sigma/L^2)$.

The strong monotonicity assumption is quite demanding. If we remove the strong monotonicity assumption and require the weaker assumption of *monotonicity* (i.e., $\langle T(z) - T(x), z - x\rangle \geq 0$ for all $x, z \in \mathbb{R}^n$) the situation becomes more complicated, and quite different from the case of convex optimization. Consider $T : \mathbb{R}^2 \to \mathbb{R}^2$ defined as $T(x) = Ax$ with $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$. T is monotone and the unique solution of VI(T,$\mathbb{R}^2$) is $x^* = 0$. However, it is easy to check that $\|x - \alpha T(x)\| > \|x\|$ for all $x \neq 0$ and all $\alpha > 0$, so that the sequence generated by (4) moves away from the solution for any choice of $\{\alpha_k\}$. In order to deal with this situation, Korpelevich proposed in [18] an extra-gradient algorithm of the form

$$z^k = \Pi(x^k - \alpha_k T(x^k)), \quad (8)$$
$$x^{k+1} = \Pi(x^k - \alpha_k T(z^k)). \quad (9)$$

In this case, if T is Lipschitz continuous with constant L and VIP(T,X) has solutions, then the sequence generated by (8)-(9) converges to a solution of VIP(T,X) provided that the α_k's are taken as $\alpha_k \equiv \alpha \in (0, 1/L)$ (see [18],[12]). Observe that all these methods are *explicit*, i.e., the formula for obtaining x^{k+1} is an explicit one, up to the computation of the orthogonal projection Π.

Another possible *implicit* approach for the solution of monotone variational inequalities is through a Tikhonov regularization scheme (see [12], Chapter 12), which dates back to the study of ill-posed variational problems in [26]. This approach typically requires solving a sequence of perturbed variational inequality problems. Precisely, the k-th iteration consists of solving the variational inequality VI($T + \epsilon_k I, X$), where the operator $T + \epsilon_k I$ is a perturbation of the original operator T given by a positive scalar ϵ_k. In this way, each of the variational inequality problems VI($T + \epsilon_k I, X$) is strongly monotone and, hence, it has a unique solution $t^k \in X$. Under suitable conditions, if the regularization sequence $\{\epsilon_k\}$ is decreasing and convergent to zero, the Tikhonov sequence $\{t^k\}$ converges to the least-norm solution of VI(T, X). An alternative to the Tykhonov method is the proximal point method (see [12], Chapter 12 and [25]). In this case, the convergence to a single solution of VI(T, X) is obtained through the addition of a proximal term $\theta(x^k - x^{k-1})$, where θ is a fixed positive parameter. The sequence x^k is defined as the (unique) solution of VI($T + \theta(I - x^{k-1}), X$) and convergence is guaranteed under suitable assumptions. These are implicit methods, since they require the solution of a sequence of regularized variational inequalities.

Excepting in very special cases (e.g., when X is an affine manifold or a ball), the computation of the projection is a computationally expensive task, and hence it is desirable to replace the projection Π in (4) and (8)-(9) by a more easily computable operator. In general, it is natural to assume that X is of the form

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

with all the X_i's closed and convex. When the orthogonal projection onto each X_i, namely $\Pi_i : \mathbb{R}^n \to X_i$, is much easier to compute than Π, a natural idea consists of replacing, at iteration k, Π by one of the Π_i's, say Π_{i_k}, or even by an approximation of Π_i. A natural context for this procedure is the case in which X is a polyhedron and the X_i'a are halfspaces. This procedure is the basis of the so called *sequential row action methods* for solving systems of equations (another option consists of computing separately but simultaneously the projections onto all the X_i's and then taking a convex combination of these projections as the next iterate; these are the so called *parallel row action methods*). See [7] for more details on row action methods.

When dealing with sequential row action methods, it is necessary to specify the order in which the sets X_i are used along the iterations, i.e. the so called *control sequence* $\{i_k\} \subset \{1, \ldots, m\}$. Several options have been considered in the literature:

a) *cyclical control*: $\{i_k\} = \{\sigma(1), \sigma(2), \ldots, \sigma(m), \sigma(1), \ldots\}$ where σ is a permutation of $\{1, \ldots, m\}$.

b) *almost cyclical control*: i_k is chosen arbitrarily, but in such a way that there exists $q \in \mathbb{N}$ such that each set is used al least once in each sequence of q consecutive iterations.

c) *most violated constraint control*: $i_k = \text{argmax}_{i \in M} h_i(z)$, where z is the point to be projected in the k-th iteration and each X_i is assumed to be of the form $X_i = \{x \in \mathbb{R}^n : h_i(x) \leq 0\}$, with $h_i : \mathbb{R}^n \to \mathbb{R}$ convex for all $i \in \{1, \ldots, m\}$.

d) *random control*: i_k is sampled from $\{1, \ldots, m\}$ with a given probability distribution, with some assumption on it ensuring positivity of the frequency with which each index $i \in \{1, \ldots, m\}$ is used.

A negative consequence of the use of any kind of approximate projections in explicit methods for variational inequalities is that the iterates in general cease to be feasible, and therefore, no specific relation exists between them and the solution values, i.e. (1) is not valid any more when x is an element of the generated sequence, because we do not have $x \in X$. In order to preserve the convergence properties in a context of nonfeasible iterates, it becomes necessary to take *small stepsizes*, assuming e.g. that

$$\sum_{k=0}^{\infty} \alpha_k^2 < \infty, \ \sum_{k=0}^{\infty} \alpha_k = \infty. \quad (10)$$

The first condition allows to control the perturbation caused by the lack of feasibility of the iterates, while the second one ensures that the sequence will eventually reach a solution even when the initial iterate lies arbitrarily distant from the solution set. With these small stepsizes linear convergence rates are not attainable, even in the case of the steepest descent method for convex minimization (a special case of $\mathrm{VI}(T, X)$ with $X = \mathbb{R}^n$ and $T = \nabla h$ for a convex $h : \mathbb{R}^n \to \mathbb{R}$). Thus we have a trade-off between the easier computation of the iterates afforded by an approximate projection, and the number of iterations required to achieve a point close enough to the solution set.

For methods using approximate projections, some assumption is needed in order to ensure that the projections onto the sets X_i's are reasonable approximations of the projection onto X. For this, some form of error bound or linear regularity is assumed on the sets X_i, see Assumption 8 in Subsection 2.2 and the comments following it. When each X_i is of the form $X_i = \{x \in \mathbb{R}^n : h_i(x) \leq 0\}$, where the h_i's are convex and differentiable, existence of a Slater point (i.e. a point $\hat{x}$ such that $h_i(\hat{x}) < 0$ for all i) is enough. This is the assumption made for instance in [3].

Explicit methods for monotone variational inequalities using approximate projections were studied e.g. in [13] and [8], imposing rather demanding coercivity assumptions on T, in [2] assuming paramonotonicity of T, and then in [4] assuming just monotonicity of T. Another method of this type, using an Armijo search as in [15] for determining the stepsizes, and approximate projections with the most violated constraint control, can be found in [3].

1.2 Stochastic approximation methods

We now discuss methods for stochastic variational inequalities (1)-(2), and we focus on stochastic approximation methods (SA).

Jiang and Xu proposed the following SA method with exact projections for SVI in [16]:

$$x^{k+1} = \Pi[x^k - \alpha_k(T(x^k) + w^k)], \tag{11}$$

where $\{w^k\}$ is a sequence of random vectors on $(\Omega, \mathcal{F}, \mathbb{P})$, called stochastic error. This formulation includes the important case in which the stochastic error is given by samples v^k of v, i.e.

$$w^k = F(v^k, x^k) - T(x^k), \tag{12}$$

in which case (11) can be written directly in terms of the available random map F:

$$x^{k+1} = \Pi[x^k - \alpha_k F(v^k, x^k)]. \tag{13}$$

When $X = \mathbb{R}^n$ (13) becomes the SA method first proposed by Robbins and Monro in [24] for the SE problem given by (3):

$$x^{k+1} = x^k - \alpha_k F(v^k, x^k). \tag{14}$$

In the deterministic case, this method coincides with Bertsekas' algorithm (4). For the convergence analysis, it is assumed that T is strongly monotone and Lipschitz continuous, as in the case of of its deterministic counterpart (4).

Modeling (11) as a stochastic process with the natural filtration

$$\mathcal{F}_k = \sigma(x^0, w^0, \dots, w^{k-1}),$$

where the initial iterative x^0 is possibly an integrable random vector, this method has been proved in [16] to converge to the (unique) solution of SVI, with stepsizes satisfying (6)-(7), which includes the option of a constant stepsize, and with the following conditions on the stochastic error:

$$\mathbb{E}[w^k | \mathcal{F}_k] = 0, \tag{15}$$

$$\sum_k \alpha_k^2 \mathbb{E}[\|w^k\|^2 | \mathcal{F}_k] < \infty. \tag{16}$$

The first condition means that the error is "stochastically unbiased" and the second one is a "stochastic boundedness" hypothesis on the error variance, both of which are standard in SA methods. If the errors

$\{w^k\}$ are bounded, i.e., for all $k \in \mathbb{N}$, $\|w^k\| \leq C$ for some $C > 0$, then the second condition above is satisfied when $\sum_k \alpha_k^2 < \infty$, i.e., when the steps are square summable. In this case, convergence of the methods requires that $\sum_k \alpha_k = \infty$.

Recently, Wang and Bertsekas in [27] improved upon the method in [16] by allowing approximate projections instead of exact ones, for the case of $X = \cap_{i=1}^m X_i$, with a random control sequence, where both the random map F and the control sequence $\{\omega_k\}$ are jointly sampled, giving rise to the following algorithm:

$$x^{k+1} \quad = \quad \Pi_{\omega_k}[x^k - \alpha_k F(v^k, x^k)]. \tag{17}$$

This method is a stochastic variation of method (4). In [27] the convergence of (17) to the solution of SVI is analyzed under two options for the control sequence $\{\omega_k\}$:
a) cyclic control, according to either a deterministically order or randomly permuted order,
b) random control, assuming that all indices are sampled sufficiently often.

In case (b), $\{\omega_k\}$ is a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $\{1, \ldots, m\}$. The method (17) is modeled as a stochastic process with the natural filtration

$$\mathcal{F}_k = \sigma(x^0, \omega_0, \ldots, \omega_{k-1}, v^0, \ldots, v^{k-1}),$$

with an integrable initial iterate x^0. For convergence, the mean operator T is assumed to be strongly monotone and F is required to be *stochastically Lipschitz*, namely, there exists $L > 0$, such that for all $x, z \in \mathbb{R}^n$ and all $k \in \mathbb{N}$,

$$\mathbb{E}\big[\|F(v^k, z) - F(v^k, x)\|^2 \big| \mathcal{F}_k\big] \leq L^2 \|z - x\|^2 \tag{18}$$

almost surely. The notion of stochastic Lipschitz continuity of the random map F in (18) resembles the standard Lipschitz continuity hypothesis of T used in [16] (see Assumption 4 in Subsection 2.2 and the comments following it). Two additional assumptions on the operator sampling, in the spirit of (15) and (16) are required: first, the sampling must be unbiased, i.e., $\mathbb{E}[F(v^k, x^k)|\mathcal{F}_k] = T(x^k)$ for every $k \in \mathbb{N}$; second, the sampling must be stochastically bounded: if x^* is the unique solution of SVI (1)-(2), then there exists $B_{x^*} > 0$ such that

$$\mathbb{E}\big[\|F(v^k, x^*)\|^2 \big| \mathcal{F}_k\big] \leq B_{x^*}^2 \tag{19}$$

for all $k \in \mathbb{N}$ almost surely. For the convergence of the method (17) small stepsizes, satisfying (10), are required, due to the stochastic approximation and to the use of approximate projections. Finally, the convergence analysis of method (17) requires also a linear regularity condition on the sets X_i, namely our Assumption 8 in Subsection 2.2.

Recently, regularized iterative Tychonov and proximal point methods for monotone stochastic variational inequalities were introduced in [19]. In standard (deterministic) Tykhonov methods, one requires a sequence of exact or approximate solutions of the strongly monotone variational inequalities VI$(T + \epsilon_k I, X)$ for each iteration $k \in \mathbb{N}$. In the stochastic regime, termination criteria are generally much harder to meet. As a consequence, one often provides confidence intervals in practice by generating a fixed number of sample paths. Another observation is that the convergence of methods based upon the Tikhonov regularization requires increasing accuracy of the subproblem solution. The implementation of such algorithms in a stochastic regime is substantially harder, since the mentioned confidence intervals for each regularized problem are obtained via simulation, which require that these intervals get increasingly tighter. Alternatively, a stochastic *iterative* Tykhonov-based scheme is studied in [19], where, instead of solving a sequence of regularized variational inequality problems, the regularization parameter is updated in each iteration and a *single projection step* associated with the regularized problem is taken. Also, the algorithm in [19] allows for a Cartesian structure on the variational inequality, so as to encompass, for example, equilibrium conditions of monotone stochastic Nash games with a limited coordination between the players stepsize and regularization sequences. Namely, the feasible set $X \subset \mathbb{R}^N$ has the the form

$$X = X^1 \times \cdots \times X^m, \tag{20}$$

where each Cartesian component $X^i \subset \mathbb{R}^{n_i}$ is a closed and convex set, and the operator has components $T = (T_1, \ldots, T_m)$ with $T_i : \mathbb{R}^N \to \mathbb{R}^{n_i}$ for $i = 1, \ldots, m$ and $\sum_{i=1}^m n_i = N$. The algorithm in [19] is thus described as follows: given the k-th iterate $x^k \in X$ with components $x_i^k \in X^i$, for $i = 1, \ldots, m$, the next iterate is given by the projection

$$x_i^{k+1} = \Pi_{X^i}[x_i^k - \alpha_{k,i}(T_i(x^k) + \epsilon_{k,i} x_i^k + w_i^k)], \tag{21}$$

for $i = 1, \ldots, m$, where $\{\alpha_{k,1}, \ldots, \alpha_{k,m}\}$ are the stepsize sequences, $\{\epsilon_{k,1}, \ldots, \epsilon_{k,m}\}$ are the regularization parameter sequences, and $w^k = (w_1^k, \ldots, w_m^k)$ is the stochastic error. As mentioned before, in method (21) a single exact projection step associated with the regularized problem is taken at each iteration after the regularization parameter update. This method is shown to converge under monotonicity and Lipschitz-continuity of T and a partial coordination between the stepsize and regularization parameter sequences (see Assumption 14)

The iterative proximal point in [19] follows a similar pattern, where a single exact projection step associated with the regularized problem is taken at every iteration after updating the centering parameter θ_k update. Namely, the proximal point method in [19] is

$$x_i^{k+1} = \Pi_{X^i}[x_i^k - \alpha_{k,i}(T_i(x^k) + \theta_{k,i}(x^k - x^{k-1}) + w_i^k)], \tag{22}$$

for $i = 1, \ldots, m$. Differently from the Tykhonov method, this method requires strong monotonicity, which in particular implies uniqueness of solutions.

An important example where the Cartesian structure mentioned above appears is the so called stochastic Nash game, in which there are N players and the i-th one must solve the parametrized stochastic minimization problem

$$\begin{aligned} \min \quad & \mathbb{E}[f_i(v, x_i, x_{-i})] \\ \text{s.t.} \quad & x_i \in X^i. \end{aligned}$$

Here, x_{-i} denotes the collection $\{x_j : j \neq i\}$ of decisions of players j other than player i, and $x_i \mapsto \mathbb{E}[f_i(v, x_i, x_{-i})]$ is convex for all $x_{-i} \in \prod_{j \neq i} X^j$. The equilibrium conditions for the problem above can be formulated as a variational inequality of the form (1), with X as in (20) and $T_i(x) = \nabla_{x_i} \mathbb{E}[f_i(v, x)]$ for $i = 1, \ldots, m$. In this setting, the use of *different* stepsizes and regularization parameters for each component $i \in \{1, \ldots, m\}$ is motivated by the players' need to choose their stepsize and regularization parameters while abiding by a coordination requirement.

Recently, stochastic extra-gradient methods for SVI have been proposed in [17], [9] and [28]. They take the form

$$z^k = \Pi(x^k - \alpha_k F(v_1^k, x^k)), \tag{23}$$

$$x^{k+1} = \Pi(x^k - \alpha_k F(v_2^k, z^k)), \tag{24}$$

where $\{v_1^k, v_2^k\}$ are samples from the random variable v. In [17], [9] and [28] a class of extra-gradient methods based upon the mirror-prox method introduced by Nemirovski in [21] is studied. In these methods, an additional iterative *averaging* technique is used *after* the extra-gradient steps (23)-(24). In [17], [9] and [28], different sets of weights are studied and convergence rates are obtained. In all these methods, the sampling is assumed to be unbiased, and the following uniform stochastic boundedness condition is imposed: there exists $\sigma > 0$ such that

$$\mathbb{E}[\|F(v, x) - T(x)\|^2] \leq \sigma^2 \tag{25}$$

for all $x \in X$. In [17], X is assumed to be compact, and T Lipschitz continuous and monotone on X. In [28] X is assumed to be compact and T monotone and bounded on X, i.e., $\sup_{x \in X} \|T(x)\| < \infty$. In [9], X is allowed to be unbounded, with different convergence analyses for the cases of bounded and unbounded X. Differently from the methods mentioned above, their convergence results are non-asymptotic, and based on gap functions for VI(T, X). For the analysis, tools from Large Deviations theory are used, meaning that for a given maximum number M of iterations, an explicit convergence rate in terms of M is obtained for the expected value of the the gap function.

Finally, in the second part of [28], another extragradient method for the SVI is presented, without requiring the Lipschitz constant of T. The method uses the iteration (23)-(24), without the averaging output. A smoothing technique is applied. The assumptions are again the stochastic boundedness (25), boundedness of X and T, strict monotonicity of T (which implies uniqueness of solutions) and the following *weak sharpness* assumption: there exists $\rho > 0$ such that

$$\langle T(x^*), x - x^* \rangle \geq \rho \mathrm{d}(x, X^*) \tag{26}$$

for all $x \in X$, $x^* \in X^*$, where X^* is the solution set of $\mathrm{VI}(T, X)$ (see Section 1.4 for comments on the weak sharpness property). Differently from the above mentioned extra-gradient methods based upon the mirror-prox method, the convergence results are asymptotic, as in [16], [27] and [19], using super-martingale convergence theorems and obtaining, the a.s.-convergence of the full sequence and a convergence rate for $\mathbb{E}[\|x^k - x^*\|^2]$, where x^* is the unique solution of $\mathrm{VI}(T, X)$. Concerning the choice of stepsizes, this method uses an adaptive stepsize sequence, which does not demand the Lipschitz constant of T, but requires the knowledge of ρ and of bounds on $\|x\|, \|T(x)\|$ for $x \in X$.

The convergence analysis of the stochastic methods (11) and (17) requires strong monotonicity. As in [19], [17], [9] and [28], we are concerned here with the extension of these methods to the barely monotone case. In [17], [9] and [28], this is achieved through an extra-gradient method and an averaging technique (which allows larger steps of order $O(1/\sqrt{k})$ at the k-th iteration instead of the usual $O(1/k)$), with an optimal rate of convergence. However, the convergence analysis is different: non-asymptotic convergence rates for the expected value of the *gap function* are derived, instead of convergence rates for the *distance* of the sequence to the solution set. Also, these methods require compactness of the feasible set (unless an additional perturbation on the operator is added, see [9], Theorem 3.3). In [28], an additional extra-gradient method with an asymptotic analysis for a class of variational inequalities satisfying the weak sharpness property (26) is presented, but there is still the requirement of compactness of the feasible set, strong monotonicity (implying uniqueness of solutions), and knowledge of the sharpness modulus ρ. In [19], instead of the extra-gradient approach, iterative regularized Tykhonov and proximal point methods for the monotone case are suggested, with the important property that instead of solving a sequence of regularized variational inequalities (whose solution can be complex, specially in the stochastic regime), a single exact projection step associated with the regularized problem is taken at every iteration. In the proximal point method, however, strict monotonicity is required.

1.3 Our methods

In many asymptotic stochastic approximation methods, the stochastic error $w(x) := F(v, x) - T(x)$ (which is present due to the lack of knowledge of the mean operator T) is at most bounded, implying the use of a small stepsizes, satisfying (10), with a slow convergence rate. In this case, the use of easyly computable approximate projections, instead of an exact one, can improve significantly the performance of the algorithm. Excepting for method (17) for strongly monotone stochastic variational inequalities, all the above mentioned methods use exact projections. We thus propose two projection methods which use approximate projections.

The first one is the extension of the projection method (17) with approximate projections, proposed in [27], to the class of monotone variational inequalities satisfying the weak sharpness property (26). This property has been first proposed in [6] for convex minimization problems and latter extended by [20] and [29] to monotone variational inequalities. The geometric structure and algorithmical implications of weak sharpness are developed in [20, 29]. In important cases, this property is equivalent to the existence of an error bound for the solution set of the variational inequality in terms of a gap function, in cases with nonunique solutions (thus the term "weak"). We explore weak sharpness in the stochastic approximation method (17) applied to a class of variational inequalities where: (1) approximate projections are possible, and (2) the solution set is not a singleton. In this sense we improve upon the method in [27], which requires strong monotoncity of the operator, and thus uniqueness of the solution. Under monotonicity and weak sharpness, we prove that the distance from the generated sequence $\{x^k\}$ to the solution set X^* a.s. converges to zero, without the knowledge of the sharpness modulus ρ. In the particular case in which the feasible set X is compact, we present a convergence rate depending upon ρ.

The second method is a variation of the regularized iterative Tychonov method proposed in [19] with approximate projections, requiring just monotonicity. We thus improve upon the results [19], which use exact projections, and upon [27], where strong monotonicity is assumed. Precisely, a single approximate projection step associated with the regularized problem is taken at each iteration, after updating the regularization parameter. We also keep the Cartesian structure of the variational inequality, so that the method can be applied to the computation of equilibrium of monotone stochastic Nash games with limited coordination between the players stepsize and regularization sequences. Due to the use of approximate projections instead of exact ones, an additional coordination requirement is imposed. We mention that it is satisfied by usual choices of stepsizes and regularization parameters (see Section 3.2, Assumption 14 and comments following it and [19], Lemma 4). We prove a.s.-convergence of the generated sequence, under plain monotonicity of the operator.

The paper is organized as follows: Section 2 analyzes the first method and Section 3 deals with the second one. Since the assumptions differ for the two methods, we list the assumptions in each section, along with the algorithm statements and their convergence analysis. In the following subsection we give some preliminary result and notation.

1.4 Some preliminaries and notation

We always assume existence in the solution set X^* of $\mathrm{VI}(T,X)$:

Assumption 1. *Problem* $\mathrm{VI}(T,X)$ *is* consistent*:* $X^*\neq\emptyset$.

For $x,y\in\mathbb{R}^n$, we denote $\langle x,y\rangle$ the standard inner product and $\|x\|=\sqrt{\langle x,x\rangle}$ the correspondent Euclidian norm. We shall denote by $\mathrm{d}(\cdot,C)$ the distance function to a general set C, namely, $\mathrm{d}(x,C)=\inf\{\|x-y\|:y\in C\}$ and by d the distance function $\mathrm{d}(\cdot,X)$ to the feasible set X. We denote the l^1-norm as $\|\cdot\|_1$. For a closed and convex set $C\subset\mathbb{R}^n$, we denote by Π_C the orthogonal projection onto C. The following properties of the projection operator are well known.

Lemma 1. *Take a closed and convex set* $C\subset\mathbb{R}^n$. *Then*

i) For all $x\in\mathbb{R}^n, y\in C$,
$$\langle x-\Pi_C(x),y-\Pi_C(x)\rangle\leq 0.$$

ii) For all $x\in\mathbb{R}^n, y\in C$,
$$\|\Pi_C(x)-y\|^2+\|\Pi_C(x)-x\|^2\leq\|x-y\|^2. \tag{27}$$

iii) For all $x,y\in\mathbb{R}^n$,
$$\|\Pi_C(x)-\Pi_C(y)\|\leq\|x-y\|. \tag{28}$$

iv) For all $x,y\in\mathbb{R}^n$,
$$\|\Pi_C(y)-y\|^2\leq 2\|\Pi_C(x)-x\|^2+8\|y-x\|^2. \tag{29}$$

Proof. Item (i) is just the first order optimality condition for the optimization problem which defines $\Pi_C(x)$, namely the minimization of $\|x-y\|^2$ subject to $y\in C$. Items (ii), (iii) follow immediately from (i). Item (iv) follows from (iii), triangle inequality and $(a+b)^2\leq 2a^2+2b^2$. □

The abbreviation “a.s.” means “almost surely” and the abbreviation “i.i.d.” means “independent and identically distributed”. Given sequences $\{x^k\}$ and $\{y^k\}$ and a point $x\in\mathbb{R}^n$, the notation $x^k=O_x(y^k)$ means that there exists $C_x>0$, depending only oupn x, such that $\|x^k\|\leq C_x\|y^k\|$ for all k. When there is no dependence on x, we use the notation O instead of O_x. Given a σ-algebra $\mathcal{F}$ and a random variable ξ, we denote by $\mathbb{E}[\xi]$, $\mathbb{E}[\xi|\mathcal{F}]$, and $\mathbb{V}[\xi]$ the expectation, conditional expectation and variance, respectively. Also, we write $\xi\in\mathcal{F}$ for “ξ is $\mathcal{F}$-measurable”. We denote by $\sigma(\xi_1,\ldots,\xi_n)$ the σ-algebra generated by the random variables $\xi_1,\ldots,\xi_n$.

As in other stochastic approximation methods, a fundamental tool to be used is the following Non-negative Almost Super-martingale Convergence Theorem of Robbins and Siegmund (see [24]), which can be seen as the stochastic version of the properties of quasi-Féjer convergence sequences.

Theorem 1 ([24], Theorem 1)**.** *Let* $\{y_k\},\{u_k\},\{a_k\},\{b_k\}$ *be sequences of nonnegative random variables, adapted to the filtration* $\{\mathcal{F}_k\}$, *such that*
$$\mathbb{E}\big[y_{k+1}\big|\mathcal{F}_k\big]\leq(1+a_k)y_k-u_k+b_k,\ \forall k\in\mathbb{N},\ a.s.,$$
where $\sum a_k<\infty$ *and* $\sum b_k<\infty$ *almost surely. Then* $\{y_k\}$ *converges and* $\sum u_k<\infty$, *almost surely.*

We will also use the following result, whose proof can be found in Lemma 10 of [22].

Theorem 2. *Let* $\{y_k\}, \{a_k\}, \{b_k\}$ *be sequences of nonnegative random variables, adapted to the filtration* $\{\mathcal{F}_k\}$*, such that*

$$\mathbb{E}\big[y_{k+1}\big|\mathcal{F}_k\big] \leq (1-a_k)y_k + b_k, \ \forall k \in \mathbb{N}, \ a.s.,$$

where $a_k \in [0,1]$*,* $\sum a_k = \infty$*,* $\sum b_k < \infty$ *and* $\lim_{k\to\infty} \frac{b_k}{a_k} = 0$ *almost surely. Then* $\{y_k\}$ *converges to zero, almost surely.*

1.4.1 Weak-sharpness

In the following we denote by

$$\mathbb{N}_X(x) = \{v \in \mathbb{R}^n : \langle v, y - x\rangle \leq 0, \forall y \in X\}$$

the normal cone of X at the point $x \in X$. The tangent cone of X at $x \in X$ is defined as $\mathbb{T}_X(x) = [\mathbb{N}_X(x)]^\circ$ where for a set $Y \subset \mathbb{R}^n$, the polar set Y° is defined as

$$Y^\circ = \{v \in \mathbb{R}^n : \langle v, y\rangle \leq 0, \forall y \in Y\}.$$

In [6], the notion of *weak-sharp* minima for the problem $\min_{x\in X} f(x)$ with solution set X^* was introduced: there exists $\rho > 0$ such that

$$f(x) - f(x^*) \geq \rho \mathrm{d}(x, X^*) \tag{30}$$

for all $x^* \in X^*$ and all $x \in X$. In [6], it is proved that if f is a closed, proper, and differentiable convex function and if the sets X and X^* are nonempty, closed, and convex, then (30) is equivalent to the geometric condition:

$$-\nabla f(x^*) \in \mathrm{int}\left(\bigcap_{x\in X^*} [\mathbb{T}_X(x) \cap \mathbb{N}_{X^*}(x)]^\circ\right) \tag{31}$$

for all $x^* \in X^*$.

In optimization problems, objective values can be used for determining regularity of solutions. In variational inequalities one can use for that purpose the above geometric definition or use the gap function

$$G(x) = \sup_{y\in X} \langle T(y), x - y\rangle.$$

We denote by $B(0,1)$ the unitary ball and X^* the solution set of $\mathrm{VI}(T,X)$. Consider the following statements:

(i) There exists $\rho > 0$, such that

$$-F(x^*) + \rho B(0,1) \in \bigcap_{x\in X^*} [\mathbb{T}_X(x) \cap \mathbb{N}_{X^*}(x)]^\circ \tag{32}$$

for all $x^* \in X^*$.

(ii) There exists $\rho > 0$, such that

$$\langle T(x^*), z\rangle \geq \rho\|z\|, \forall z \in \mathbb{T}_X(x) \cap \mathbb{N}_{X^*}(x) \tag{33}$$

for all $x^* \in X^*$.

(iii)

$$-T(x^*) \in \mathrm{int}\left(\bigcap_{x\in X^*} [\mathbb{T}_X(x) \cap \mathbb{N}_{X^*}(x)]^\circ\right) \tag{34}$$

for all $x^* \in X^*$.

(iv) There exist $\rho > 0$ such that

$$G(x) \geq \rho \mathrm{d}(x, X^*). \tag{35}$$

for all $x \in X$.

Statement (iii) is the definition of a weak-sharp VI(T, X) given in [20]. In Theorem 4.1 of [20], it was proved that (i)-(ii) are equivalent, and that (i)-(iv) are equivalent when X is compact and T is paramonotone i.e.,

$$\langle T(x) - T(y), x - y\rangle = 0 \Rightarrow T(x) = T(y)$$

for all $x, y \in \mathbb{R}^n$ (see [15] for other properties of paramonotone operators).

Relation (35) means that the gap function G provides an error bound on the solution set X^*. Paramonotonicity implies that T is constant on the solution set X^* and important classes of paramonotone operators are, for example, co-coercive, symmetric monotone and strictly monotone composite operators (see [12], Chapter 2)

Recently, the following assumption was introduced in [28]: there exists $\rho > 0$ such that

$$\langle T(x^*), x - x^*\rangle \geq \rho \mathrm{d}(x, X^*). \tag{36}$$

for all $x^* \in X^*$ and all $x \in X$. Clearly, (36) implies (35). We prove next that (36) implies (33) and the converse statement holds when T is constant on X^*. Thus, when T is constant on X^*, (32), (33) and (36) are equivalent, and when T is paramonotone and X is compact, relations (32)-(36) are all equivalent.

Proposition 1. *Condition* (36) *implies* (33). *If T is constant on X^*, then* (33) *implies* (36).

Proof. Suppose that (36) holds and let $x^* \in X^*$. If $\mathbb{T}_X(x^*) \cap \mathbb{N}_{X^*}(x^*) = \{0\}$, then (33) holds trivially. Otherwise, take $d \in \mathbb{T}_X(x^*) \cap \mathbb{N}_{X^*}(x^*)$ with $d \neq 0$. For all $\bar{x} \in X^*$, we have

$$\langle d, \bar{x} - x^*\rangle \geq 0,$$
$$\langle d, \bar{x} - x^*\rangle \leq 0,$$

where the first relation holds because $d \in \mathbb{T}_X(x^*)$ and the second one holds because $d \in \mathbb{N}_X(x^*)$. From the above relations, X^* is a subset of the hyperplane $H_d := \{y : \langle d, y - x^*\rangle = 0\}$. Since $d \in \mathbb{T}_X(x^*)$, there exist sequences $d^k \in \mathbb{R}^n$, $t_k > 0$ such that $x^* + t_k d^k \in X^*$, $d^k \to d$ and $t^k \to 0$. From (36) we get

$$\langle T(x^*), x^* + t_k d^k - x^*\rangle \geq \rho \mathrm{d}(x^* + t_k d^k, X^*) \geq \rho \mathrm{d}(x^* + t_k d^k, H_d) = \rho t_k \frac{\langle d, d^k\rangle}{\|d\|}.$$

Dividing the above relation by t_k and letting $k \to \infty$, we conclude that (33) holds for d.

Now suppose that (33) holds and that T is constant on X^*. Take $x \in X$, $x^* \in X^*$. Let $\bar{x} := \Pi_{X^*}(x)$. Since $x, \bar{x} \in X$ and X is convex, we have $x - \bar{x} \in \mathbb{T}_X(\bar{x})$. On the other hand, since $\bar{x} := \Pi_{X^*}(x)$, we obtain from Lemma 1(i) and closedness and convexity of X^* that $x - \bar{x} \in \mathbb{N}_{X^*}(\bar{x})$. Thus, $x - \bar{x} \in \mathbb{T}_X(\bar{x}) \cap \mathbb{N}_{X^*}(\bar{x})$. We conclude from relation (33) that

$$\langle T(\bar{x}), x - \bar{x}\rangle \geq \rho\|x - \bar{x}\| = \rho \mathrm{d}(x, X^*). \tag{37}$$

Since T is constant on X^*, we have

$$\langle T(\bar{x}), x - \bar{x}\rangle = \langle T(x^*), x - \bar{x}\rangle = \langle T(x^*), x - x^*\rangle + \langle T(x^*), x^* - \bar{x}\rangle \leq \langle T(x^*), x - x^*\rangle,$$

using the fact that $\langle T(x^*), x - x^*\rangle \leq 0$, which holds because $x^* \in X^*$ and $\bar{x} \in X$. The result follows from the above relation and (37). □

Finally, we will need the following result from [20].

Theorem 3. *If T is continuous and $-T(y) \in \mathrm{int}\big(\bigcap_{x \in X^*}[\mathbb{T}_X(x) \cap \mathbb{N}_{X^*}(x)]^\circ\big)$ for a certain $y \in \mathbb{R}^n$, then $\mathrm{argmin}_{x \in X}\langle T(y), x\rangle \subset X^*$.*

Proof. See Theorem 4.2 of [20]. □

As a consequence of Theorem 3, under weak sharpness and uniform continuity of T, any algorithm which generates a sequence $\{x^k\}$ such that $\mathrm{d}(x^k, X^*) \to 0$ has the property that after a *finite* number of iterations M, any solution of the auxiliary linear program

$$\mathrm{argmin}_{x \in X}\langle T(x^M), x\rangle$$

is a solution of the original variational inequality (see Theorem 5.1 in [20]). This result can be used for devising new algorithms under the weak sharpness property. However, in practice, M may be very large. In Theorem 5 of Section 2 we provide an estimate of the average value of M for our first projection method, under the weak sharpness assumption.

2 A projection method with approximate projections

2.1 Statement of the algorithm

As commented upon in Section 1, the expensive step in the projection method is the computation of the two orthogonal projections onto X. Among the several schemes which have been proposed for replacing the orthogonal projection by more easily computable approximate ones, we will follow here the approach presented in [27]. In this reference, the feasible set X is assumed to be of the form

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

where all the X_i's are closed and convex. We assume that the orthogonal projections onto the X_i's, namely $\Pi_i : \mathbb{R}^n \to X_i$, are relatively easy to compute, while the orthogonal projection onto X, namely $\Pi : \mathbb{R}^n \to X$, is not. A prototypical instance of this situation occurs when X is a polyhedron and each X_i is a halfspace. The approximation scheme consists of replacing, in each step of the projection method, Π by one of the Π_i's. If i_k is the index of the constraint used in the k-th iteration, the sequence $\{i_k\}$ is called *control sequence*. Among the various control sequences which have been studied, the method in [27] chooses the random control sequence, i.e., we consider the control sequence as a sequence $\{\omega_k\}$ of random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$, with values in $\{1, \ldots, m\}$. We sample also the random variable v in the definition of the mapping F, selecting a realization $F(v_k, \cdot)$ of F as the operator to be used in the k-th iteration.

Algorithm 1. *1)* **Initialization:** *Choose the initial iterate $x^0 \in \mathbb{R}^n$ such that $\mathbb{E}[\|x^0\|] < \infty$, the stepsizes $\{\alpha_k\}$, the random controls $\{\omega_k\}$ and the operator samples $\{v^k\}$.*

2) **Iterative step:** *Given x^k, define:*

$$x^{k+1} = \Pi_{\omega_k}(x^k - \alpha_k F(v^k, x^k)). \tag{38}$$

2.2 Discussion of the assumptions

In the sequel we consider the natural filtration

$$\mathcal{F}_k = \sigma(x^0, \omega_0, \ldots, \omega_{k-1}, v_0, \ldots, v_{k-1}).$$

Next we present the assumptions necessary for our convergence analysis.

Assumption 2. *The pair (T, X) is weakly sharp, i.e., there exists $\rho > 0$, such that*

$$\langle T(x^*), x - x^* \rangle \geq \rho \mathrm{d}(x, X^*) \tag{39}$$

for all $x^* \in X^*$ and all $x \in X$.

This assumption was proposed recently in [28] with different purposes. We explore weak sharpness in the stochastic approximation method (38) as a property of variational inequalities such that

1) approximate projections are available,

2) solutions are possibly non unique.

By considering these kind of problems, we improve upon the analysis of [27], where strong monotonicity (which implies uniqueness of solutions) is required for convergence of the method in (38), while our approach demands the much weaker condition (39).

Assumption 3. *The mean operator T is* monotone*, i.e.,*

$$\langle T(y) - T(x), y - x \rangle \geq 0$$

for all $y, x \in \mathbb{R}^n$.

Assumption 4. *The random operator F is* stochastically Lipschitz*, i.e., there exists $L > 0$, such that*

$$\mathbb{E}\big[\|F(v^k, y) - F(v^k, x)\|^2 \big| \mathcal{F}_k\big] \le L^2 \|y - x\|^2$$

for all $y, x \in \mathbb{R}^n$, $k \in \mathbb{N}$ almost surely.

The stochastic Lipschitz continuity in Assumption 4 is akin to the Lipschitz continuity of T. Indeed, when the samples $\{v^k\}$ have the same distribution as v, Assumption 4 implies that T is Lipschitz and it is satisfied, for instance, when:
a) The random variable v takes values in a finite sample space.
b) There exists $L > 0$, such that, for all $z, x \in \mathbb{R}^n$ and all $k \in \mathbb{N}$, it holds that $\|F(v^k, z) - F(v^k, x)\| \le L\|z - x\|$ almost surely, or, in other words, when $\{F(v^k, \cdot)\}$ is equi-Lipschitz almost surely.
c) There exists a positive continuous function $L(\cdot)$ such that for every $z, x \in \mathbb{R}^n$, $\|F(v, z) - F(v, x)\| \le L(v)\|z - x\|$ almost surely, and v takes values in a compact sample space.

Indeed, (a)⇒(b)⇒(c). Since only the random operator is supposed to be available, the stochastic Lipschitz continuity of F is a more practical assumption than Lipschitz continuity of T.

Assumption 5. *The operator sampling is* stochastically unbiased*, i.e.,*

$$\mathbb{E}\big[F(v^k, x) \big| \mathcal{F}_k\big] = T(x)$$

for all $x \in \mathbb{R}^n$ and all $k \in \mathbb{N}$, almost surely.

Assumption 6. *The operator sampling is* stochastically bounded *on the solution set, i.e., there exists $B > 0$ such that*

$$\mathbb{E}\big[\|F(v^k, x)\|^2 \big| \mathcal{F}_k\big] \le B^2,$$

for all $x \in X^*$, for all $k \in \mathbb{N}$, almost surely.

Since

$$\|T(x)\| = \|\mathbb{E}[F(v, x)]\| \le \mathbb{E}[\|F(v, x)\|] \le \sqrt{\mathbb{E}[\|F(v, x)\|]^2} \le B$$

for all $x \in X^*$, Assumption 6 implies in particular that the mean T is bounded on X^*. Assumption 6 entails that the operator sampling is bounded and the operator is bounded on $x \in X^*$. Indeed, if the random operator $F : \Xi \times X \to \mathbb{R}^n$ is bounded on $\Xi \times X^*$, then $B := \sup_{(\xi, x) \in \Xi \times X^*} \|F(\xi, x)\|$ satisfies Assumption 6. Sufficient conditions for Assumption 6 to hold are the continuity of F and the compactness of the sample Ξ and the solution set X^*.

We now state the assumptions concerning the approximate projections. The following assumption on the control sequence ω_k is akin to the almost cyclical control used in row action methods (see [7]).

Assumption 7. *Each constraint is sampled sufficiently often, i.e., there exists $\delta \in (0, 1]$, such that, for all $i \in M$,*

$$\mathbb{P}\big(\omega_k = i \big| \mathcal{F}_k\big) \ge \frac{\delta}{m}$$

for all $k \in \mathbb{N}$, almost surely.

This is the same condition imposed in [27]. It means that each constraint is sampled sufficiently often "near independently" of each other and independently of the past. For example, if the constraint samples $\{\omega_k\}$ are i.i.d. with uniform distribution on M and independent of the mapping samples $\{v^k\}$, then this condition holds with $\delta = 1$, since $\mathbb{P}\big(\omega_k = i \big| \mathcal{F}_k\big) = \mathbb{P}(\omega_k = i) = \frac{1}{m}$.

Assumption 8. *The feasible set X satisfies a* linear regularity *condition: there exists $\eta > 0$ such that*

$$\|\Pi(x) - x\|^2 \le \eta \max_{i \in M} \|\Pi_i(x) - x\|^2$$

for all $x \in \mathbb{R}^n$.

This is the same condition imposed in [27]. This condition is satisfied by any polyhedron (Hoffman's Lemma). It has been analyzed by Bauschke and Borwein [1] (Definition 5.6, p. 40), and Deutsch and Hundal [11]. These references provide several other situations where the linear regularity condition holds, and indicate that it is a mild restriction in practice. This assumption means that each Π_i is a not too bad approximation of Π.

Assumption 9. *The stepsizes α_k satisfy:*

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \tag{40}$$

$$\sum_{k=0}^{\infty} \alpha_k^2 < \infty. \tag{41}$$

We remark here that the use of small stepsizes in method (38) is forced by two factors: the use of approximate projections instead of exact ones, and the mapping sampling of SA methods. Indeed, even with *exact projection*, method (38) under Assumption 6 would still require small stepsizes.

2.3 Preliminary results

We state now two lemmas which are needed for our convergence analysis.

Lemma 2. *Almost surely, for all $x \in X^*, y \in \mathbb{R}^n$, $k \in \mathbb{N}$,*

$$\mathbb{E}\big[\|F(v^k, y)\|^2 \big| \mathcal{F}_k\big] \le 2L^2\|y - x\|^2 + 2B^2.$$

Proof. The proof is similar to the one of Lemma 3 in [27]; we present it here for sake of completeness. Let $x \in X^*, y \in \mathbb{R}^n$, $k \in \mathbb{N}$, then

$$\mathbb{E}\big[\|F(v^k, y)\|^2 \big| \mathcal{F}_k\big] \le 2\mathbb{E}\big[\|F(v^k, y) - F(v^k, x)\|^2 \big| \mathcal{F}_k\big] + 2\mathbb{E}\big[\|F(v^k, x)\|^2 \big| \mathcal{F}_k\big] \le 2L^2\|y - x\|^2 + 2B^2,$$

using the bound $(a + b)^2 \le 2a^2 + b^2$ in the first inequality and Assumptions 4, 6 in the second one. □

Lemma 3. *Almost surely,*

$$\mathbb{E}\big[\|\Pi_{\omega_k}(x) - x\|^2 \big| \mathcal{F}_k\big] \ge \frac{\delta}{m\eta} \mathrm{d}^2(x)$$

holds for all $x \in \mathbb{R}^n$ and all $k \in \mathbb{N}$

Proof. The proof is similar to the one of Lemma 4 in [27]. □

2.4 Convergence analysis

Theorem 4. *Under Assumptions 1-9, the method* (38) *generates a bounded sequence $\{x^k\}$ such that*

$$\lim_{k\to\infty} \mathrm{d}(x^k, X^*) = 0, \text{ almost surely.} \tag{42}$$

In particular, all cluster points of $\{x^k\}$ belong to X^* almost surely.

Proof. Take $x \in X^*$ (nonempty by Assumption 1). Denote $y^k := x^k - \alpha_k F(v^k, x^k)$. Since $x \in X_{\omega_k}$, we get from Lemma 1(ii)

$$\begin{aligned}
\|x^{k+1} - x\|^2 &= \|\Pi_{\omega_k}(y^k) - x\|^2 \\
&\le \|y^k - x\|^2 - \|y^k - \Pi_{\omega_k}(y^k)\|^2 \\
&= \|(x^k - x) - \alpha_k F(v^k, x^k)\|^2 - \|y^k - \Pi_{\omega_k}(y^k)\|^2 \\
&= \|x^k - x\|^2 - 2\alpha_k\langle x^k - x, F(v^k, x^k)\rangle + \alpha_k^2\|F(v^k, x^k)\|^2 - \|y^k - \Pi_{\omega_k}(y^k)\|^2 \\
&\le \|x^k - x\|^2 - 2\alpha_k\langle x^k - x, F(v^k, x^k)\rangle + \alpha_k^2\|F(v^k, x^k)\|^2 - \frac{1}{2}\|x^k - \Pi_{\omega_k}(x^k)\|^2 + 4\|y^k - x^k\| \\
&= \|x^k - x\|^2 + 2\alpha_k\langle x - x^k, F(v^k, x^k)\rangle + 5\alpha_k^2\|F(v^k, x^k)\|^2 - \frac{1}{2}\|x^k - \Pi_{\omega_k}(x^k)\|^2,
\end{aligned} \tag{43}$$

using Lemma 1(ii) in the first inequality, Lemma 1(iv) in the second one and simple algebra in the equalities.

From Assumption 5 and the fact that $x^k \in \mathcal{F}_k$, we obtain

$$\mathbb{E}\big[\langle x - x^k, F(v^k, x^k)\rangle \big| \mathcal{F}_k\big] = \langle x - x^k, \mathbb{E}\big[F(v^k, x^k)\big|\mathcal{F}_k\big]\rangle = \langle x - x^k, T(x^k)\rangle. \tag{44}$$

By Lemma 2 and the fact that $x^k \in \mathcal{F}_k$, we have

$$\mathbb{E}\big[\|F(v^k, x^k)\|^2 \big| \mathcal{F}_k\big] \leq 2L^2\|x^k - x\|^2 + 2B^2. \tag{45}$$

By Lemma 3, $x^k \in \mathcal{F}_k$. Denoting $A := \delta/(m\eta)$, we get

$$-\mathbb{E}\big[\|\Pi_{\omega_k}(x^k) - x^k\|^2 \big| \mathcal{F}_k\big] \leq -A\mathrm{d}^2(x^k). \tag{46}$$

Using the fact that $x^k \in \mathcal{F}_k$, taking conditional expectation in (43) and invoking (44)-(46), we get

$$\mathbb{E}\big[\|x^{k+1} - x\|^2 \big| \mathcal{F}_k\big] \leq \big(1 + 10L^2\alpha_k^2\big)\|x^k - x\|^2 + 2\alpha_k\langle x - x^k, T(x^k)\rangle + 10B^2\alpha_k^2 - \frac{A}{2}\mathrm{d}^2(x^k). \tag{47}$$

Concerning the second term in the right hand side of (47), we write

$$\langle T(x^k), x - x^k\rangle = \langle T(x^k) - T(x), x - x^k\rangle + \langle T(x), x - \Pi(x^k)\rangle + \langle T(x), \Pi(x^k) - x^k\rangle. \tag{48}$$

By monotonicity of T (Assumption 3), the first term in the right hand side of (48) satisfies

$$\langle T(x^k) - T(x), x - x^k\rangle \leq 0. \tag{49}$$

Regarding the second term in the right hand side of (48), the weak sharpness property (Assumption 2) and the fact that $x \in X^*$ imply

$$\langle T(x), x - \Pi(x^k)\rangle \leq -\rho\mathrm{d}(\Pi(x^k), X^*). \tag{50}$$

We now observe that $|\mathrm{d}(\Pi(x^k), X^*) - \mathrm{d}(x^k, X^*)| \leq \|\Pi(x^k) - x^k\| = \mathrm{d}(x^k)$, so that

$$\mathrm{d}(\Pi(x^k), X^*) \geq \mathrm{d}(x^k, X^*) - \mathrm{d}(x^k).$$

Using this inequality in (50), we get

$$\langle T(x), x - \Pi(x^k)\rangle \leq -\rho\mathrm{d}(x^k, X^*) + \rho\mathrm{d}(x^k). \tag{51}$$

Concerning the third term in the right hand side of (48), we have

$$\langle T(x), \Pi(x^k) - x^k\rangle \leq \|T(x)\|\|\Pi(x^k) - x^k\| \leq B\mathrm{d}(x^k). \tag{52}$$

Combining (49), (51) and (52) in (48) we finally get

$$\langle T(x^k), x - x^k\rangle \leq -\rho\mathrm{d}(x^k, X^*) + (\rho + B)\mathrm{d}(x^k). \tag{53}$$

We use (53) in (47) and get

$$\begin{aligned}\mathbb{E}\big[\|x^{k+1} - x\|^2 \big| \mathcal{F}_k\big] \leq{}& \big(1 + 10L^2\alpha_k^2\big)\|x^k - x\|^2 - 2\rho\alpha_k\mathrm{d}(x^k, X^*) + 10B^2\alpha_k^2 \\ &+2(\rho + B)\alpha_k\mathrm{d}(x^k) - \frac{A}{2}\mathrm{d}^2(x^k).\end{aligned} \tag{54}$$

Defining $C := 2(\rho + B)$, we use the relation $-a^2 + 2ab = -(a - b)^2 + b^2$ in order to get

$$-\frac{A}{2}\mathrm{d}^2(x^k) + C\alpha_k\mathrm{d}(x^k) = -\left(\sqrt{\frac{A}{2}}\mathrm{d}(x^k) - \frac{C\alpha_k}{\sqrt{2A}}\right)^2 + \frac{C^2\alpha_k^2}{2A} \leq O(\alpha_k^2). \tag{55}$$

In view of (55) and (54), finally we get

$$\mathbb{E}\big[\|x^{k+1} - x\|^2 \big| \mathcal{F}_k\big] \leq \big(1 + 10L^2\alpha_k^2\big)\|x^k - x\|^2 - 2\rho\alpha_k\mathrm{d}(x^k, X^*) + \Big[10B^2 + 2A^{-1}(\rho + B)^2\Big]\alpha_k^2$$

$$= \left[1 + O(\alpha_k^2)\right] \|x^k - x\|^2 - 2\rho\alpha_k \mathrm{d}(x^k, X^*) + O(\alpha_k^2) \tag{56}$$

for all $x \in X^*$.

Choose now some $x^* \in X^*$. By Assumption 9, we have $\sum_k \alpha_k^2 < \infty$. Hence, we conclude from Theorem 1 and (56) that $\{\|x^k - x^*\|\}$ a.s. converges and in particular, $\{x^k\}$ is a.s.-bounded.

Since (56) holds for all $x \in X^*$, we invoke Assumption 1 for choosing $\bar{x}^k := \Pi_{X^*}(x^k)$. Now, $\bar{x}^k \in \mathcal{F}_k$, because since $x^k \in \mathcal{F}_k$). Since $\mathrm{d}(x^k, X^*) = \|x^k - \bar{x}^k\|$, we have

$$\mathbb{E}\left[\mathrm{d}^2(x^{k+1}, X^*)\middle|\mathcal{F}_k\right] \leq \mathbb{E}\left[\|x^{k+1} - \bar{x}^k\|^2\middle|\mathcal{F}_k\right]$$

$$\leq \left[1 + O(\alpha_k^2)\right] \|x^k - \bar{x}^k\|^2 - 2\rho\alpha_k \mathrm{d}(x^k, X^*) + O(\alpha_k^2)$$

$$= \left[1 + O(\alpha_k^2)\right] \mathrm{d}^2(x^k, X^*) - 2\rho\alpha_k \mathrm{d}(x^k, X^*) + O(\alpha_k^2) \tag{57}$$

for all $k \in \mathbb{N}$, using (56) and the fact that $x^k \in \mathcal{F}_k$ in the second inequality.

By Assumption 9, we have $\sum_k \alpha_k^2 < \infty$, so that we conclude from Theorem 1 and (57) that $\{\mathrm{d}(x^k, X^*)\}$ a.s. converges, and almost surely

$$\sum_{k=0}^{\infty} 2\rho\alpha_k \mathrm{d}(x^k, X^*) < \infty.$$

By Assumption 9, we also have $\sum_k \alpha_k = \infty$, so that the above relation implies a.s.

$$\liminf_{k\to\infty} \mathrm{d}(x^k, X^*) = 0.$$

In particular, the sequence $\{\mathrm{d}(x^k, X^*)\}$ has a subsequence that converges to zero almost surely. Since $\{\mathrm{d}(x^k, X^*)\}$ a.s. converges, we conclude that the whole sequence a.s. converges to 0. □

Next we present a convergence rate result for the method with approximate projections (38) under the weak sharpness property (39) and compactness of the feasible set. More precisely, our convergence rate holds for the distance to the solution set $\mathrm{d}(x^k, X^*)$. We apply this result for obtaining an estimate of the average number of iterations required so that any solution of an auxiliary linear program is a solution of the variational inequality.

We require compactness of each set $X_1, \ldots, X_m$ (whose intersection is the feasible set X). In the important case in which X is a compact polyhedron, we can enforce this condition by replacing each X_i by its intersection with a box (i.e., a norm-1 ball) containing X, say X_i' so that $X = \cap X_i'$, and all the X_i''s are clearly compact. If each X_i is a halfspace, the orthogonal projection onto X_i' is reasonably easy to compute (though harder than the projection onto X_i). Corollary 5.6 and Theorem 5.27 of [1] exhibit a large number of situations in which Assumption 8 holds. We continue with our convergence rate result.

Theorem 5. *Let Assumptions 1-9 hold. Assume that $0 < \alpha_k < \rho/(10RL^2)$ for all k and that each X_i is compact. Take $R \geq \max_{i\in M} \max_{x\in X_i} \|x\|$. Let $\{x^k\}$ be the sequence generated by method (38). Then,*

a) Almost surely, $\{x^k\}$ is bounded, $\mathrm{d}(x^k, X^*)$ converges to 0 and, for all $\epsilon > 0$ there exists a finite random variable $M \in \mathbb{N}$, such that

$$\min_{0\leq k\leq M} \{\mathrm{d}(x^k, X^*) - \beta_k\} \leq \epsilon, \tag{58}$$

satisfying

$$\mathbb{E}\left[\sum_{k=0}^{M-1} \alpha_k\right] \leq \frac{\mathrm{d}^2(x^0, X^*)}{2\rho\epsilon}, \tag{59}$$

where

$$\beta_k = \frac{2Rc_1\epsilon + c_2}{2(\rho - Rc_1\alpha_k)}\alpha_k = O(\alpha_k), \tag{60}$$

$$c_1 = 10L^2 \qquad \text{and} \qquad c_2 = 10B^2 + 2(\rho + B)^2\delta m^{-1}\eta^{-1}. \tag{61}$$

b) In particular, for all $\epsilon \in (0, \rho/L)$, and all stepsize sequence satisfying

$$0 < \alpha_k < \min\left\{\frac{\rho}{c_1 R}, \frac{2\rho(\frac{\rho}{L} - \epsilon)}{c_2 + 2Rc_1\frac{\rho}{L}}\right\}$$

for all k, there exist almost surely a finite random $M \in \mathbb{N}$ such that (58) *and* (59) *hold, and any solution of the linear program*

$$\min_{x \in X} \langle T(x^M), x\rangle$$

belongs to X^*.

Proof. First we prove item (a). Fix $\epsilon > 0$. From (57) we have for every $k \in \mathbb{N}$,

$$\mathbb{E}\big[\mathrm{d}^2(x^{k+1}, X^*)\big|\mathcal{F}_k\big] \le \big[1 + c_1\alpha_k^2\big]\mathrm{d}^2(x^k, X^*) - 2\rho\alpha_k \mathrm{d}(x^k, X^*) + c_2\alpha_k^2. \tag{62}$$

We define the following level set and stopping time:

$$L_k = \{x \in \mathbb{R}^n : \mathrm{d}(x^k, X^*) \le \beta_k + \epsilon\} \tag{63}$$

$$M := \inf\{k \in \mathbb{N} : x^k \in L_k\}. \tag{64}$$

We also define the following "stopped" auxiliary processes:

$$u_{k,M} = \begin{cases} 2\rho\alpha_k \mathrm{d}(x^k, X^*) - c_1\alpha_k^2 \mathrm{d}^2(x^k, X^*) - c_2\alpha_k^2 & \text{if } k < M, \\ 0 & \text{if } k \ge M, \end{cases} \tag{65}$$

$$x^{k\wedge M} = \begin{cases} x^k & \text{if } k < M, \\ x^M & \text{if } k \ge M. \end{cases} \tag{66}$$

Since $x^k \in \mathcal{F}_k$ and M is a stopping time, we have that $x^{k\wedge M}$ and $u_{k,M}$ belong to $\mathcal{F}_k$. From (65), (66) and (62) we get for all $k \in \mathbb{N}$,

$$\mathbb{E}\big[\mathrm{d}^2(x^{(k+1)\wedge M}, X^*)\big|\mathcal{F}_k\big] \le \mathrm{d}^2(x^{k\wedge M}, X^*) - u_{k,M} \tag{67}$$

for all $k \in \mathbb{N}$. From (63)-(65), we have for $k < M$:

$$\begin{aligned} u_{k,M} &\ge (2\rho\alpha_k - 2Rc_1\alpha_k^2)\mathrm{d}(x^k, X^*) - c_2\alpha_k^2 \\ &\ge 2(\rho\alpha_k - Rc_1\alpha_k^2)(\beta_k + \epsilon) - c_2\alpha_k^2 = 2\rho\epsilon\alpha_k \end{aligned} \tag{68}$$

for $k < M$, using the bound $d(x^k, X) \le 2R$ in the first inequality, the fact that $\rho\alpha_k - Rc_1\alpha_k^2 > 0$ (which follows from the fact that $0 < \alpha_k < \rho/c_1 R$) in the second one and (60) in the equality. In particular, we have that $u_{k,M} \ge 0$ for all $k \in \mathbb{N}$.

We claim now that $\mathbb{P}(M < \infty) = 1$. This fact, together with (64), implies that $\mathrm{d}(x^M, X^*) \le \delta_M + \epsilon$ in which case (58) holds with probability 1. We proceed to prove the claim. Since $\{u_{k,M}\}$ is nonnegative, we conclude from (67) and Theorem 1 that a.s. $\sum_k u_{k,M} < \infty$. In the event $[M = \infty]$ we have, by (68), $\sum_k u_{k,M} \ge 2\rho\epsilon \sum_k \alpha_k = \infty$. Hence $\mathbb{P}(M = \infty) = 0$.

We take total expectation in (67) and sum from 0 to some ℓ, obtaining

$$\mathbb{E}\big[\mathrm{d}^2(x^{(\ell+1)\wedge M}, X^*)\big] \le \mathrm{d}^2(x^0, X^*) - \mathbb{E}\left[\sum_{k=0}^{\ell} u_{k,M}\right]. \tag{69}$$

Finally, we let $\ell \to \infty$ and use the monotone convergence theorem in (69), obtaining

$$\mathrm{d}^2(x^0, X^*) \ge \mathbb{E}\left[\sum_{k=0}^{\infty} u_{k,M}\right] = \mathbb{E}\left[\sum_{k=0}^{M-1} u_{k,M}\right] \ge 2\rho\epsilon\mathbb{E}\left[\sum_{k=0}^{M-1} \alpha_k\right], \tag{70}$$

using (65) in the equality and (68) in the last inequality. We have proved (59).

For the proof of item (b), we observe that by the choice of ϵ and $\{\alpha_k\}$ we have $\beta_k + \epsilon < \rho/L$ for all k. Hence, invoking item (a), we have $\mathrm{d}(x^M, X^*) \le \beta_M + \epsilon < \rho/L$. Calling $\bar{x}^M := \Pi_{X^*}(x^M)$, we obtain, from the Lipschitz continuity of T,

$$\|T(x^M) - T(\bar{x}^M)\| \le L\|x^M - \bar{x}^M\| = L\mathrm{d}(x^M, X^*) < \rho.$$

From Proposition 1, Assumption 2 and the equivalence between (32) and (33), we get

$$-T(x^M) + \rho B(0,1) \subset \bigcap_{x \in X^*} [\mathbb{T}_X(x) \cap \mathbb{N}_{X^*}(x)]^\circ. \tag{71}$$

Hence, $-T(x^M) \in \mathrm{int}\big(\bigcap_{x\in X^*}[\mathbb{T}_X(x) \cap \mathbb{N}_{X^*}(x)]^\circ\big)$ and we get from Theorem 3

$$\mathrm{argmin}_{x\in X}\langle T(x^M), x\rangle \subset X^*.$$

□

We remark that the above result is also useful when exact projections are taken. In such a case, the constants are smaller: $c_1 = 10L^2$ and $c_2 = 10B^2$.

3 An iterative Tikhonov regularization method with approximate projections

In this section we assume the variational inequality (1) has a Cartesian structure. We consider the decomposition $\mathbb{R}^N = \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m}$, with $N = n_1 + \ldots + n_m$ and furnish this Cartesian space with the standard inner product $\langle x, y\rangle = \sum_{i=1}^m \langle x_i, y_i\rangle$ for $x = (x_1, \ldots, x_m)$ and $y = (y_1, \ldots, y_m)$. We suppose that the feasible set $X \subset \mathbb{R}^N$ has the form

$$X = X^1 \times \cdots \times X^m, \tag{72}$$

where each component $X^i \subset \mathbb{R}^{n_i}$ is a closed and convex set for $i = 1, \ldots, m$. Also, the random operator $F : \Xi \times \mathbb{R}^N \to \mathbb{R}^N$ has the form

$$F = (F_1, \ldots, F_m),$$

where each component is of the form $F_i : \Xi \times \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m} \to \mathbb{R}^{n_i}$ for $i = 1, \ldots, m$. From (2), the mean operator has the form $T = (T_1, \ldots, T_m)$ with $T_i(x) = \mathbb{E}[F_i(\xi, x)]$ for $i = 1, \ldots, m$.

A typical situation where this structure appears is the so called stochastic Nash game, in which there are m players and the i-th one must solve the parametrized minimization problem

$$\begin{aligned} \min \quad & \mathbb{E}[f_i(v, x_i, x_{-i})] \\ \text{s.t.} \quad & x_i \in X^i. \end{aligned}$$

Here, x_{-i} denotes the collection $\{x_j : j \neq i\}$ of decisions of players j other than player i, and $x_i \mapsto \mathbb{E}[f_i(v, x_i, x_{-i})]$ is convex for all $x_{-i} \in \prod_{j\neq i} X^j$. The equilibrium conditions of the above problem can be formulated as a stochastic variational inequality of the form (1)-(2) with X as in (72) and $F_i(v, x) = \nabla_{x_i} f_i(v, x)$ under reasonable assumptions of integrability, differentiability and Lipschitz continuity of f_i for $i = 1, \ldots, m$.

Finally, in order to explore the use of approximate projections, we also assume that for $i = 1, \ldots, m$, each Cartesian component X^i of X in (72) has the following constraint form:

$$X^i = \cap_{j=1}^{n_i} X^i_j \tag{73}$$

where each constraint component $X^i_j \subset \mathbb{R}^{n_i}$ is closed and convex with easy computable projections for $j \in \{1, \ldots, n_i\}$. A typical situation with this formulation occurs when each Cartesian component X^i is a polyhedron and the X^i_j's are half-spaces.

The idea of our iterative Tykhonov method consist of combining the distributed architecture for computing equilibria of the iterative Tykhonov method in [19] (which allows a Cartesian structure of the associated VI) using easily computable approximate projections, with constraints as in (73). Under these two structures (the Cartesian form (72) and constraints of the form (73)), the scheme uses a predefined random control sequence such that, in each iteration of the Tykhonov method, the projection Π_{X^i} is replaced by one of the $\Pi_{X^i_j}$ for every $i = 1, \dots, m$. Simultaneously, the random variable v in the definition of F is also sampled, selecting a realization $F_i(v^k, \cdot)$ of F_i for each $i = 1, \dots, m$, to be used in the k-th iteration. As in [19], we also permit a partial coordination between the stepsize and the regularization parameters of the players, so that each player can choose independently its stepsize and its regularization sequence (see Assumption 14 and comments following it).

We continue some notation. Given $i \in \{1, \dots, m\}$, we denote by $\Pi^i_j : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ the orthogonal projection onto X^i_j for each selection $j \in \{1, \dots, n_i\}$ of the constraint components of X^i. Given a selection $\mathbf{j} = (j_1, \dots, j_m) \in \prod_{i=1}^m \{1, \dots, n_i\}$ of the Cartesian and constraint components, we denote by $\Pi_{\mathbf{j}} : \mathbb{R}^N \to \mathbb{R}^N$ the projection onto $X^1_{j_1} \times \cdots \times X^N_{j_m}$. We emphasize that the orthogonal projection under a Cartesian structure is simple: for $x = (x_1, \dots, x_m) \in \mathbb{R}^N$ and $Y = Y^1 \times \dots \times Y^m \subset \mathbb{R}^N$ with $x_i \in \mathbb{R}^{n_i}$ and $Y^i \subset \mathbb{R}^{n_i}$, we have

$$\Pi_Y(x) = (\Pi_{Y^1}(x_1), \dots, \Pi_{Y^m}(x_m)). \tag{74}$$

3.1 Statement of the algorithm

Algorithm 2. *1)* **Initialization:** *Choose the initial iterate $x^0 \in \mathbb{R}^n$ so that $\mathbb{E}[\|x^0\|] < \infty$, the stepsize sequence $\alpha^k = (\alpha_{k,1}, \dots, \alpha_{k,m}) \in (0, \infty)^m$, the regularization sequence $\epsilon^k = (\epsilon_{k,1}, \dots, \epsilon_{k,m}) \in (0, \infty)^m$, the random control sequence $\omega^k = (\omega_{k,1}, \dots, \omega_{k,m}) \in \prod_{i=1}^m \{1, \dots, n_i\}$ and the operator samples $\{v^k\}$.*

2) **Iterative step:** *Given $x^k = (x^k_1, \dots, x^k_m)$, define, for each $i \in \{1, \dots, m\}$,*

$$x_i^{k+1} = \Pi^i_{\omega_{k,i}}[x^k_i - \alpha_{k,i}(F_i(v^k, x^k) + \epsilon_{k,i} x^k_i)], \tag{75}$$

The iterative step can be written compactly as

$$x^{k+1} = \Pi_{\omega_k}[x^k - D(\alpha_k) \cdot (F(v^k, x^k) + D(\epsilon_k) x^k)]$$

where $D(\lambda)$ is the diagonal matrix with entries $\lambda = (\lambda_1, \dots, \lambda_n) \in \mathbb{R}^n$. In order to simplify the notation, we will write in ther sequel $\Pi_{\omega_{k,i}}$ for $\Pi^i_{\omega_{k,i}}$.

3.2 Discussion of the assumptions

We consider the natural filtration

$$\mathcal{F}_k = \sigma(x^0, \omega^0, \dots, \omega^{k-1}, v^0, \dots, v^{k-1}).$$

Assumption 10. *We request Assumptions 3, 4 and 5, namely, monotonicity of T, stochastic Lipschitz continuity of F and an unbiased operator sampling.*

Assumption 11. *The operator sampling is pointwise* stochastically bounded *on the feasible set, i.e., there exists a locally bounded measurable function $B : \mathbb{R}^n \to (0, \infty)$ such that*

$$\mathbb{E}\big[\|F(v^k, x)\|^2 \big| \mathcal{F}_k\big] \leq B^2(x),$$

for all $x \in X$ and all $k \in \mathbb{N}$, almost surely.

We also have for every $x \in X$,

$$\|T(x)\| = \|\mathbb{E}[F(v, x)]\| \leq \mathbb{E}[\|F(v, x)\|] \leq \sqrt{\mathbb{E}[\|F(v, x)\|]^2} \leq B(x)$$

for all $x \in X$. Assumption 11 means that the operator sampling is pointwise bounded on the feasible set. Indeed, if the sample space Ξ is compact and $F(\cdot, x)$ is continuous for all $x \in X$, then $B(x) := \sup_{\xi \in \Xi} \|F(\xi, x)\|$ satisfies Assumption 11. In [27], strong monotonicity of T implies uniqueness of the solution, and this condition is requested to hold only for such unique solution. Here, with possibly multiple solutions, we request Assumption 6 to hold *uniformly on the solution set*.

We now state the assumptions concerning the approximate projections which accommodate the Cartesian structure. Basically, each Cartesian component $X^i = \cap_{j=1}^{n_i} X_j^i$ should satisfy Assumptions 7, 8 and, as in [19], the step and regularization sequences need the following coordination assumption:

Assumption 12. *For each Cartesian component $i \in \{1, \ldots, m\}$, each constraint component in $X^i = \cap_{j=1}^{n_i} X_j^i$ is sampled sufficiently often, i.e., there exists $\delta_i \in (0, 1]$, such that*

$$\mathbb{P}\big(\omega_{k,i} = j \big| \mathcal{F}_k\big) \geq \frac{\delta_i}{n_i}$$

for all $j \in \{1, \ldots, n_i\}$, and all $k \in \mathbb{N}$, almost surely.

We observe that Assumption 12 requires a sampling coordination between the control sequences $\{\omega_{k,i}\}_{k=0}^{\infty}$ for $i = 1, \ldots, m$ since the filtration $\mathcal{F}_k$ accumulates the history from the control sequence of every Cartesian component. Such assumption is satisfied, for example, when $\delta_i = 1$ for all $i = 1, \ldots, m$ if $\{\omega_{k,1}\}, \ldots \{\omega_{k,m}\}$, the $\{v^k\}$'s are independent, and $\{\omega_{k,i}\}_{k=0}^{\infty}$ are i.i.d. with a uniform distribution on $\{1, \ldots, n_i\}$. for each $i \in \{1, \ldots, m\}$.

Assumption 13. *For every Cartesian component $i \in \{1, \ldots, m\}$, X^i satisfies a* linear regularity *condition: there exists $\eta_i > 0$ such that,*

$$\|\Pi_{X^i}(x) - x\|^2 \leq \eta_i \max_{j \in \{1, \ldots, n_i\}} \|\Pi_{X_j^i}(x) - x\|^2,$$

for all $x \in \mathbb{R}^{n_i}$.

Assumption 14. *Let $\alpha_{k,\min} = \min_{1 \leq i \leq m} \alpha_{k,i}$, $\alpha_{k,\max} = \max_{1 \leq i \leq m} \alpha_{k,i}$, $\epsilon_{k,\min} = \min_{1 \leq i \leq m} \epsilon_{k,i}$ and $\epsilon_{k,\max} = \max_{1 \leq i \leq m} \epsilon_{k,i}$. Then,*

(i) For each $i = 1, \ldots, m$, $\{\epsilon_{k,i}\}_{k=1}^{\infty}$ is a decreasing sequence converging to zero.

(ii) $\lim_{k \to \infty} \frac{\alpha_{k,\max}^2}{\alpha_{k,\min}\epsilon_{k,\min}} = 0$, $\lim_{k \to \infty} \frac{\alpha_{k,\max} - \alpha_{k,\min}}{\alpha_{k,\min}\epsilon_{k,\min}} = 0$ *and* $\lim_{k \to \infty} \alpha_{k,\min}\epsilon_{k,\min} = 0$.

(iii) $\sum_{k=0}^{\infty} \alpha_{k,\min}\epsilon_{k,\min} = \infty$.

(iv)

$$\sum_{k=0}^{\infty} \alpha_{k,\max}^2 < \infty,$$

$$\sum_{k=0}^{\infty} \frac{(\alpha_{k,\max} - \alpha_{k,\min})^2}{\alpha_{k,\min}\epsilon_{k,\min}} < \infty,$$

$$\sum_{k=0}^{\infty} \left(\frac{\epsilon_{k-1,\max} - \epsilon_{k,\min}}{\epsilon_{k,\min}}\right)^2 \left(1 + \frac{1}{\alpha_{k,\min}\epsilon_{k,\min}}\right) < \infty.$$

(v) $\lim_{k \to \infty} \frac{(\epsilon_{k-1,\max} - \epsilon_{k,\min})^2}{\epsilon_{k,\min}^3 \alpha_{k,\min}} \left(1 + \frac{1}{\alpha_{k,\min}\epsilon_{k,\min}}\right) = 0$.

Assumption 14 contains usual conditions on the regularization parameters of Tykhonov algorithms, and conditions on the stepsize sequence for SA algorithms, with certain coordination across stepsizes and regularization parameters. Assumption 14 includes Assumption 2 in [19] with the following addition, due to the use of approximate projections:

$$\sum_{k=0}^{\infty} \frac{(\alpha_{k,\max} - \alpha_{k,\min})^2}{\alpha_{k,\min}\epsilon_{k,\min}} < \infty. \tag{76}$$

We observe that this condition is trivially satisfied when the stepsizes are equally chosen among the players. Lemma 4 of [19] establishes that stepsizes and regularization parameters of the form

$$\alpha_{k,i} = (k + C_i)^{-c},$$
$$\epsilon_{k,i} = (k + D_i)^{-d},$$

satisfy Assumption 2 in [19], when $c, d \in (0,1)$ are such that $c > d$ and $c + d < 1$, the C_i's belong to the interval $[\underline{C}, \overline{C}]$ and the D_i's belong to the interval $[\underline{D}, \overline{D}]$ for some $0 < \underline{C} < \overline{C}$ and $0 < \underline{D} < \overline{D}$. These stepsizes and parameters also satisfy our extra condition (76): indeed, if $C_{\max} = \max_{1\le i\le m} C_i$, $C_{\min} = \min_{1\le i\le m} C_i$ and $D_{\max} = \max_{1\le i\le m} D_i$, then

$$\alpha_{k,\min}\epsilon_{k,\min} = (k + C_{\max})^{-c}(k + D_{\max})^{-d} = k^{-(c+d)}(1 + C_{\max}/k)^{-c}(1 + D_{\max}/k)^{-d} > k^{-(c+d)} > k^{-1},$$

because $0 < c + d < 1$. Therefore,

$$\frac{(\alpha_{k,\max} - \alpha_{k,\min})^2}{\alpha_{k,\min}\epsilon_{k,\min}} < \frac{\alpha_{k,\max}^2}{k} = \frac{1}{k(k + C_{\min})^{2c}} \le \frac{1}{k^{1+2c}}.$$

3.3 Preliminary Results

Lemma 4. *The operator $H_k := D(\alpha_k) \cdot (T + D(\epsilon_k))$ satisfies*

$$\langle H_k(y) - H_k(x), y - x\rangle \ge \sigma_k \|y - x\|^2$$

for all $y, x \in \mathbb{R}^n$, with

$$\sigma_k = \alpha_{k,\min}\epsilon_{k,\min} - L(\alpha_{k,\max} - \alpha_{k,\min}).$$

Proof. We consider the decomposition

$$\langle H_k(y) - H_k(x), y - x\rangle = \langle D(\alpha_k) \cdot (T(y) - T(x)), y - x\rangle + \langle D(\alpha_k)D(\epsilon_k)(y - x), y - x\rangle. \tag{77}$$

Concerning the second term in the right hand side of (77), if D_k is the diagonal matrix with entries $(\alpha_1\epsilon_1, \dots, \alpha_m\epsilon_m)$, then

$$\langle D(\alpha_k)D(\epsilon_k)(y - x), y - x\rangle = \langle D_k(y - x), y - x\rangle \ge \alpha_{k,\min}\epsilon_{k,\min}\|y - x\|^2. \tag{78}$$

The first term in the right hand side of (77) is equal to

$$\begin{aligned}\sum_{i=1}^m \alpha_{k,i}\langle T_i(y) - T_i(x), y_i - x_i\rangle &= \alpha_{k,\min}\sum_{i=1}^m \langle T_i(y) - T_i(x), y_i - x_i\rangle \\ &\quad + \sum_{i=1}^m (\alpha_{k,i} - \alpha_{k,\min})\langle T_i(y) - T_i(x), y_i - x_i\rangle. \end{aligned} \tag{79}$$

The first term in the right hand side of (79) is nonnegative by monotonicity of T. For the second term in the right hand side of (79), we have

$$\begin{aligned}\sum_{i=1}^m (\alpha_{k,i} - \alpha_{k,\min})\langle T_i(y) - T_i(x), y_i - x_i\rangle &\ge -\sum_{i=1}^m (\alpha_{k,i} - \alpha_{k,\min})\|T_i(y) - T_i(x)\|\|y_i - x_i\| \\ &\ge -(\alpha_{k,\max} - \alpha_{k,\min})\sum_{i=1}^m \|T_i(y) - T_i(x)\|\|y_i - x_i\| \\ &\ge -(\alpha_{k,\max} - \alpha_{k,\min})\|T(y) - T(x)\|\|y - x\| \\ &\ge -(\alpha_{k,\max} - \alpha_{k,\min})L\|y - x\|^2, \end{aligned} \tag{80}$$

using Cauchy-Schwartz inequality in the first inequality, Hölder-inequality in the third one and Lipschitz continuity of T in the last one. The result follows from (77)-(80). □

We remark that, in view of Lemma 4 and the second limit condition in Assumption 14(ii), the operator H_k is in fact strongly monotone with modulus $\sigma_k > 0$ for all sufficiently large k.

Lemma 5. *Assume that $X \subset \mathbb{R}^n$ is convex and closed, that the operator $T : \mathbb{R}^n \to \mathbb{R}^n$ is continuous and monotone over X and that Assumption 1 hold. If the sequences $\{\epsilon_{k,i}\}_{k=1}^{\infty}$ for $i = 1, \dots, m$ decrease to 0 and satisfy that $\limsup_{k\to\infty} \frac{\epsilon_{k,\max}}{\epsilon_{k,\min}} < \infty$, with $\epsilon_{k,\max} = \max_i \epsilon_{k,i}$ and $\epsilon_{k,\min} = \min_i \epsilon_{k,i}$, then the Tykhonov sequence $t^k \in SOL(T + D(\epsilon_k), X)$ satisfies that*

(i) $\{t^k\}$ *is bounded and all cluster points of* $\{t^k\}$ *belong to* X^*.

(ii) The following inequality holds for all $k \geq 1$:

$$\|t^k - t^{k-1}\| \leq \frac{\epsilon_{k-1,\max} - \epsilon_{k,\min}}{\epsilon_{k,\min}} M_t,$$

where M_t is an upper bound of $\max_{k\in\mathbb{N}} \|t^k\|$.

(iii) If $\limsup_{k\to\infty} \frac{\epsilon_{k,\max}}{\epsilon_{k,\min}} \leq 1$ then $\{t^k\}$ converges to the least-norm solution in X^*.

Proof. See Lemma 3 in [19]. □

3.4 Convergence Analysis

We present next our convergence result for this method.

Theorem 6. *If Assumptions 1,10-14 hold, then the method* (75) *generates a sequence $\{x^k\}$ such that:*

(i) if $\limsup_{k\to\infty} \frac{\epsilon_{k,\max}}{\epsilon_{k,\min}} < \infty$, then almost surely $\{x^k\}$ is bounded and all cluster points of $\{x^k\}$ belong to the solution set X^*,

(ii) if $\limsup_{k\to\infty} \frac{\epsilon_{k,\max}}{\epsilon_{k,\min}} \leq 1$, then almost surely $\{x^k\}$ converges to the least-norm solution in X^*.

Proof. Let $\{t^k\}$ denote the Tykhonov sequence associated to $\mathrm{VI}(T, X)$ with regularization parameters $\{\epsilon_k\}$, so that $t^k \in \mathrm{SOL}(T+D(\epsilon_k), X)$ for all $k \in \mathbb{N}$. Take $i \in \{1, \dots, m\}$ and denote $y_i^k := x_i^k - \alpha_{k,i}(F_i(v^k, x^k) + \epsilon_{k,i} x_i^k)$. Since $t_i^k \in X_i \subset X_{\omega_{k,i}}$, we get

$$\begin{aligned}
\|x_i^{k+1} - t_i^k\|^2 &= \|\Pi_{\omega_{k,i}}(y_i^k) - t_i^k\|^2 \\
&\leq \|y_i^k - t_i^k\|^2 - \|y_i^k - \Pi_{\omega_{k,i}}(y_i^k)\|^2 \\
&= \|(x_i^k - t_i^k) + (y_i^k - x_i^k)\|^2 - \|y_i^k - \Pi_{\omega_{k,i}}(y_i^k)\|^2 \\
&= \|x_i^k - t_i^k\|^2 + 2\langle x_i^k - t_i^k, y_i^k - x_i^k\rangle + \|y_i^k - x_i^k\|^2 - \|y_i^k - \Pi_{\omega_{k,i}}(y_i^k)\|^2 \\
&\leq \|x_i^k - t_i^k\|^2 + 2\langle t_i^k - x_i^k, x_i^k - y_i^k\rangle + 5\|y_i^k - x_i^k\|^2 - \frac{1}{2}\|x_i^k - \Pi_{\omega_{k,i}}(x_i^k)\|^2,
\end{aligned} \tag{81}$$

using Lemma 1(ii) in the first inequality, Lemma 1(iv) in the second one and simple algebra on the equalities. We sum the inequalities in (81) with i between 1 and m, getting

$$\|x^{k+1} - t^k\|^2 \leq \|x^k - t^k\|^2 + 2\sum_{i=1}^m \langle t_i^k - x_i^k, x_i^k - y_i^k\rangle + 5\|y^k - x^k\|^2 - \frac{1}{2}\sum_{i=1}^m \|x_i^k - \Pi_{\omega_{k,i}}(x_i^k)\|^2. \tag{82}$$

From Assumption 5 and the fact that $x_i^k \in \mathcal{F}_k$, we obtain

$$\mathbb{E}\big[\langle t_i^k - x_i^k, x_i^k - y_i^k\rangle \big| \mathcal{F}_k\big] = \alpha_{k,i}\langle t_i^k - x_i^k, \mathbb{E}\big[F_i(v^k, x^k)\big|\mathcal{F}_k\big] + \epsilon_{k,i} x_i^k\rangle = \alpha_{k,i}\langle t_i^k - x_i^k, T_i(x^k) + \epsilon_{k,i} x_i^k\rangle. \tag{83}$$

Using the fact that $x_i^k \in \mathcal{F}_k$, we have

$$\begin{aligned}
\mathbb{E}\big[\|y_i^k - x_i^k\|^2 \big| \mathcal{F}_k\big] &= \alpha_{k,i}^2 \mathbb{E}\big[\|F_i(v^k, x^k) + \epsilon_{k,i} x_i^k\|^2 \big| \mathcal{F}_k\big] \\
&\leq \alpha_{k,i}^2 \mathbb{E}\big[\|F_i(v^k, x^k) - F_i(v^k, t^k) + \epsilon_{k,i}(x_i^k - t_i^k) + F_i(v^k, t^k) + \epsilon_{k,i} t_i^k\|^2 \big| \mathcal{F}_k\big] \\
&\leq 4\alpha_{k,i}^2 \mathbb{E}\big[\|F_i(v^k, x^k) - F_i(v^k, t^k)\|^2 \big| \mathcal{F}_k\big] + 4\alpha_{k,i}^2 \epsilon_{k,i}^2 \|x_i^k - t_i^k\|^2 \\
&\quad + 4\alpha_{k,i}^2 \mathbb{E}\big[\|F_i(v^k, t^k)\|^2 \big| \mathcal{F}_k\big] + 4\alpha_{k,i}^2 \epsilon_{k,i}^2 \|t_i^k\|^2 \\
&\leq 4\alpha_{k,\max}^2 \mathbb{E}\big[\|F_i(v^k, x^k) - F_i(v^k, t^k)\|^2 \big| \mathcal{F}_k\big] + 4\alpha_{k,\max}^2 \epsilon_{k,\max}^2 \|x_i^k - t_i^k\|^2 \\
&\quad + 4\alpha_{k,\max}^2 \mathbb{E}\big[\|F_i(v^k, t^k)\|^2 \big| \mathcal{F}_k\big] + 4\alpha_{k,\max}^2 \epsilon_{k,\max}^2 \|t_i^k\|^2.
\end{aligned}$$

Summing the inequalities in (84) with i between 1 and m, we get from Assumptions 4, 11,

$$\begin{aligned}
\mathbb{E}\big[\|y^k - x^k\|^2 \big| \mathcal{F}_k\big] &= \sum_{i=1}^m \mathbb{E}\big[\|y_i^k - x_i^k\|^2 \big| \mathcal{F}_k\big] \\
&\leq 4\alpha_{k,\max}^2 \mathbb{E}\big[\|F(v^k, x^k) - F(v^k, t^k)\|^2 \big| \mathcal{F}_k\big] + 4\alpha_{k,\max}^2 \epsilon_{k,\max}^2 \|x^k - t^k\|^2 \\
&\quad + 4\alpha_{k,\max}^2 \mathbb{E}\big[\|F(v^k, t^k)\|^2 \big| \mathcal{F}_k\big] + 4\alpha_{k,\max}^2 \epsilon_{k,\max}^2 \|t^k\|^2 \\
&\leq 4L^2 \alpha_{k,\max}^2 \|x^k - t^k\|^2 + 4\alpha_{k,\max}^2 \epsilon_{k,\max}^2 \|x^k - t^k\|^2 \\
&\quad + 4\alpha_{k,\max}^2 B^2(t^k) + 4\alpha_{k,\max}^2 \epsilon_{k,\max}^2 \|t^k\|^2 \\
&\leq 4(L^2 + \epsilon_{k,\max}^2) \alpha_{k,\max}^2 \|x^k - t^k\|^2 + 4\alpha_{k,\max}^2 (B_t^2 + \epsilon_{k,\max}^2 M_t^2), \qquad (84)
\end{aligned}$$

using the triangular inequality and the fact that $(\sum_{i=1}^4 a_i)^2 \leq 4 \sum_{i=1}^4 a_i^2$ in the first inequality, while the the last inequality follows from the fact that B_t and M_t are positive constants (depending on the Tykhonov sequence) satisfying $\max_{k\in\mathbb{N}} \|B(t^k)\| \leq B_t$ and $\max_{k\in\mathbb{N}} \|t^k\| \leq M_t$, because $\{t^k\}$ is a bounded sequence by Lemma 5, and B is a nonnegative locally bounded function by Assumption 11.

Denoting $A_i := \delta_i/(n_i\eta_i)$, we get from Lemma 3 and the fact that $x_i^k \in \mathcal{F}_k$,

$$\begin{aligned}
\sum_{i=1}^m \mathbb{E}\big[\|\Pi_{\omega_{k,i}}(x_i^k) - x_i^k\|^2 \big| \mathcal{F}_k\big] &\geq \sum_{i=1}^m A_i \|\Pi_{X^i}(x_i^k) - x_i^k\|^2 \\
&\geq A_{\min} \sum_{i=1}^N \|\Pi_{X^i}(x_i^k) - x_i^k\|^2 \\
&= A_{\min} \mathrm{d}^2(x^k), \qquad (85)
\end{aligned}$$

where $A_{\min} = \min_{1\leq i\leq m} A_i$. Now we use again the fact that $x^k \in \mathcal{F}_k$, take conditional expectation in (82) and combine the result with (83)-(85), in order to obtain

$$\mathbb{E}\big[\|x^{k+1} - t^k\|^2 \big| \mathcal{F}_k\big] \leq \big[1 + 20(L^2 + \epsilon_{k,\max}^2)\alpha_{k,\max}^2\big] \|x^k - t^k\|^2 + 2\sum_{i=1}^m \alpha_{k,i} \langle t_i^k - x_i^k, T_i(x^k) + \epsilon_{k,i} x_i^k \rangle$$

$$+20(B_t^2 + M_t^2 \epsilon_{k,\max}^2)\alpha_{k,\max}^2 - \frac{A_{\min}}{2} \mathrm{d}^2(x^k). \qquad (86)$$

The sum in the second term of the right hand side of (86) is equal to

$$\begin{aligned}
\langle D(\alpha_k) \cdot (T + D(\epsilon_k))(x^k), t^k - x^k \rangle &= \langle D(\alpha_k) \cdot (T + D(\epsilon_k))(x^k) - D(\alpha_k) \cdot (T + D(\epsilon_k))(t^k), t^k - x^k \rangle \\
&\quad + \langle D(\alpha_k) \cdot (T + D(\epsilon_k))(t^k), t^k - \Pi(x^k) \rangle \\
&\quad + \langle D(\alpha_k) \cdot (T + D(\epsilon_k))(t^k), \Pi(x^k) - x^k \rangle. \qquad (87)
\end{aligned}$$

Calling $\Delta_k := \alpha_{k,\max} - \alpha_{k,\min}$, it follows from Lemma 4 that the first term in the right hand side of (87) satisfies

$$\langle D(\alpha_k) \cdot (T + D(\epsilon_k))(x^k) - D(\alpha_k) \cdot (T + D(\epsilon_k))(t^k), t^k - x^k \rangle \leq -(\alpha_{k,\min}\epsilon_{k,\min} - L\Delta_k)\|x^k - t^k\|^2. \quad (88)$$

The second term in the right hand side of (87) is equal to

$$\sum_{i=1}^m \alpha_{k,i}\langle T_i(t^k)+\epsilon_{k,i}t_i^k, t_i^k-\Pi_{X^i}(x_i^k)\rangle = \alpha_{k,\min}\sum_{i=1}^m\langle T_i(t^k)+\epsilon_{k,i}t_i^k, t_i^k-\Pi_{X^i}(x_i^k)\rangle + \sum_{i=1}^m(\alpha_{k,i}-\alpha_{k,\min})\langle T_i(t^k)+\epsilon_{k,i}t_i^k, t_i^k-\Pi_{X^i}(x_i^k)\rangle. \quad (89)$$

The first term in the right hand side of (89) satisfies

$$\sum_{i=1}^m\langle T_i(t^k)+\epsilon_{k,i}t_i^k, t_i^k-\Pi_{X^i}(x_i^k)\rangle = \langle (T+D(\epsilon_k))(t^k), t^k-\Pi(x^k)\rangle \le 0, \quad (90)$$

since $t^k\in \mathrm{SOL}(T+D(\epsilon_k),X)$. Regarding the second term in the right hand side if (89), we invoke the fact that $\Pi_{X^i}(t_i^k)=t_i^k$, so that for each $\mu\in(0,1)$ we have

$$\begin{aligned}\sum_{i=1}^m(\alpha_{k,i}-\alpha_{k,\min})\langle T_i(t^k)+\epsilon_{k,i}t_i^k, t_i^k-\Pi_{X^i}(x_i^k)\rangle &\le \sum_{i=1}^m(\alpha_{k,i}-\alpha_{k,\min})\|T_i(t^k)+\epsilon_{k,i}t_i^k\|\|\Pi_{X^i}(t_i^k)-\Pi_{X^i}(x_i^k)\|\\ &\le \Delta_k\sum_{i=1}^m(\|T_i(t^k)\|+\epsilon_{k,i}\|t_i^k\|)\|t_i^k-x_i^k\|\\ &\le \Delta_k(B_t+\epsilon_{k,\max}M_t)\|t^k-x^k\|\\ &= 2\frac{(B_t+\epsilon_{k,\max}M_t)\Delta_k}{2\sqrt{\mu\alpha_{k,\min}\epsilon_{k,\min}}}\cdot\sqrt{\mu\alpha_{k,\min}\epsilon_{k,\min}}\|t^k-x^k\|\\ &\le \frac{(B_t+\epsilon_{k,\max}M_t)^2\Delta_k^2}{4\mu\alpha_{k,\min}\epsilon_{k,\min}}+\mu\alpha_{k,\min}\epsilon_{k,\min}\|t^k-x^k\|^2, \quad (91)\end{aligned}$$

using Cauchy-Schwartz inequality in the first inequality, Lemma 1(iii) for Π_{X^i}, in the second one, the fact that $\|T(t^k)\|\le B(t^k)\le B_t$ and $\|t^k\|\le M_t$ for all $k\in\mathbb{N}$ in the third one, and the relation $2ab=-(a-b)^2+a^2+b^2$ in the forth one. Putting together (89)-(91), we finally get that the second term in the right hand side of (87) is bounded by

$$\langle D(\alpha_k)\cdot(T+D(\epsilon_k))(t^k), t^k-\Pi(x^k)\rangle \le \frac{(B_t+\epsilon_{k,\max}M_t)^2\Delta_k^2}{4\mu\alpha_{k,\min}\epsilon_{k,\min}}+\mu\alpha_{k,\min}\epsilon_{k,\min}\|x^k-t^k\|^2. \quad (92)$$

For the third term in the right hand side of (87), we have

$$\begin{aligned}\langle D(\alpha_k)\cdot(T+D(\epsilon_k))(t^k), \Pi(x^k)-x^k\rangle &\le \|D(\alpha_k)\|\|T(t^k)+\epsilon_k t^k\|\|\Pi(x^k)-x^k\|\\ &\le \alpha_{k,\max}(B_t+\epsilon_{k,\max}M_t)\mathrm{d}(x^k). \quad (93)\end{aligned}$$

Combining (88), (92) and (93) with (87), we obtain

$$2\langle D(\alpha_k)\cdot(T+D(\epsilon_k))(x^k), t^k-x^k\rangle \le \Big[-2(1-\mu)\alpha_{k,\min}\epsilon_{k,\min}+2L\Delta_k\Big]\|x^k-t^k\|^2 + \frac{(B_t+\epsilon_{k,\max}M_t)^2\Delta_k^2}{2\mu\alpha_{k,\min}\epsilon_{k,\min}}+2\alpha_{k,\max}(B_t+\epsilon_{k,\max}M_t)\mathrm{d}(x^k). \quad (94)$$

Now we use (94) in (86), getting

$$\mathbb{E}\big[\|x^{k+1}-t^k\|^2\big|\mathcal{F}_k\big] \le q_k\|x^k-t^k\|^2 +20(B_t^2+M_t^2\epsilon_{k,\max}^2)\alpha_{k,\max}^2+\frac{(B_t+\epsilon_{k,\max}M_t)^2\Delta_k^2}{2\mu\alpha_{k,\min}\epsilon_{k,\min}} +2\alpha_{k,\max}(B_t+\epsilon_{k,\max}M_t)\mathrm{d}(x^k)-\frac{A_{\min}}{2}\mathrm{d}^2(x^k), \quad (95)$$

where

$$q_k := 1 - 2(1-\mu)\alpha_{k,\min}\epsilon_{k,\min} + 20(L^2 + \epsilon_{k,\max}^2)\alpha_{k,\max}^2 + 2L\Delta_k. \tag{96}$$

Denoting by $C_k := 2(B_t + \epsilon_{k,\max}M_t)$, the last term in the right hand side of (95) becomes

$$-\frac{A_{\min}}{2}\mathrm{d}^2(x^k) + C_k\alpha_{k,\max}\mathrm{d}(x^k) = -\left(\sqrt{\frac{A_{\min}}{2}}\mathrm{d}(x^k) - \frac{C_k\alpha_{k,\max}}{\sqrt{2A_{\min}}}\right)^2 + \frac{C_k^2\alpha_{k,\max}^2}{2A_{\min}} \le \frac{C_k^2\alpha_{k,\max}^2}{2A_{\min}} = O(\alpha_{k,\max}^2). \tag{97}$$

Using (97) in (95) we get that

$$\mathbb{E}\big[\|x^{k+1} - t^k\|^2 \big| \mathcal{F}_k\big] \le q_k\|x^k - t^k\|^2$$
$$+\left[20(B_t^2 + M_t^2\epsilon_{k,\max}^2) + \frac{2(B_t + M_t\epsilon_{k,\max})^2}{A_{\min}}\right]\alpha_{k,\max}^2 + \frac{(B_t + \epsilon_{k,\max}M_t)^2\Delta_k^2}{2\mu\alpha_{k,\min}\epsilon_{k,\min}} \tag{98}$$

for all k.

Next we relate $\|x^k - t^k\|^2$ with $\|x^k - t^{k-1}\|^2$, using the properties of the Tykhonov sequence (Lemma 5).

$$\begin{aligned}
\|x^k - t^k\|^2 &\le (\|x^k - t^{k-1}\| + \|t^k - t^{k-1}\|)^2 \\
&= \|x^k - t^{k-1}\|^2 + \|t^k - t^{k-1}\|^2 + 2\|x^k - t^{k-1}\|\|t^k - t^{k-1}\| \\
&\le \|x^k - t^{k-1}\|^2 + \left(M_t\frac{\epsilon_{k-1,\max} - \epsilon_{k,\min}}{\epsilon_{k,\min}}\right)^2 + 2M_t\frac{\epsilon_{k-1,\max} - \epsilon_{k,\min}}{\epsilon_{k,\min}}\|x^k - t^{k-1}\|.
\end{aligned}$$

Using the fact that $2ab = -(a-b)^2 + a^2 + b^2$, the last term in the rightmost expression in (99) can be estimated as

$$\begin{aligned}
2M_t\frac{\epsilon_{k-1,\max} - \epsilon_{k,\min}}{\epsilon_{k,\min}}\|x^k - t^{k-1}\| &= 2\sqrt{\alpha_{k,\min}\epsilon_{k,\min}}\|x^k - t^{k-1}\| \cdot M_t\frac{\epsilon_{k-1,\max} - \epsilon_{k,\min}}{\sqrt{\alpha_{k,\min}\epsilon_{k,\min}}\epsilon_{k,\min}} \\
&\le \alpha_{k,\min}\epsilon_{k,\min}\|x^k - t^{k-1}\|^2 + M_t^2\frac{(\epsilon_{k-1,\max} - \epsilon_{k,\min})^2}{\alpha_{k,\min}\epsilon_{k,\min}^3}.
\end{aligned}$$

The inequality in (99) yields

$$\|x^k - t^k\|^2 \le (1 + \alpha_{k,\min}\epsilon_{k,\min})\|x^k - t^{k-1}\|^2 + \left(M_t\frac{\epsilon_{k-1,\max} - \epsilon_{k,\min}}{\epsilon_{k,\min}}\right)^2\left(1 + \frac{1}{\alpha_{k,\min}\epsilon_{k,\min}}\right). \tag{99}$$

We combine (98) and (99) in order to get

$$\mathbb{E}\big[\|x^{k+1} - t^k\|^2 \big| \mathcal{F}_k\big] \le q_k(1 + \alpha_{k,\min}\epsilon_{k,\min})\|x^k - t^{k-1}\|^2$$
$$+\left[20(B_t^2 + M_t^2\epsilon_{k,\max}^2) + \frac{2(B_t + M_t\epsilon_{k,\max})^2}{A_{\min}}\right]\alpha_{k,\max}^2 + \frac{(B_t + \epsilon_{k,\max}M_t)^2\Delta_k^2}{2\mu\alpha_{k,\min}\epsilon_{k,\min}}$$
$$q_k\left(M_t\frac{\epsilon_{k-1,\max} - \epsilon_{k,\min}}{\epsilon_{k,\min}}\right)^2\left(1 + \frac{1}{\alpha_{k,\min}\epsilon_{k,\min}}\right). \tag{100}$$

We now estimate the coefficient $q_k(1 + \alpha_{k,\min}\epsilon_{k,\min})$ in (100). In view of (96), we have

$$q_k = 1 - \alpha_{k,\min}\epsilon_{k,\min}\left(2 - 2\mu - \frac{20(L^2 + \epsilon_{k,\max}^2)\alpha_{k,\max}^2}{\alpha_{k,\min}\epsilon_{k,\min}} - \frac{2L\Delta_k}{\alpha_{k,\min}\epsilon_{k,\min}}\right). \tag{101}$$

Assumption 14(ii) guarantees that

$$\frac{20(L^2 + \epsilon_{k,\max}^2)\alpha_{k,\max}^2}{\alpha_{k,\min}\epsilon_{k,\min}} + \frac{2L\Delta_k}{\alpha_{k,\min}\epsilon_{k,\min}} \to 0,$$

and, since $\mu \in (0,1)$ is arbitrary, we ensure the existence of $c \in (0,1)$ such that

$$c_k := 2\mu + \frac{20(L^2 + \epsilon_{k,\max}^2)\alpha_{k,\max}^2}{\alpha_{k,\min}\epsilon_{k,\min}} + \frac{2L\Delta_k}{\alpha_{k,\min}\epsilon_{k,\min}} < c \tag{102}$$

for all sufficiently large k. Next we show that $q_k \in (0,1)$ for large k. Indeed, from (102) and the fact that $c \in (0,1)$ we have that $1 < 2 - c_k < 2$ for large enough k, so that we obtain, from (101),

$$1 - 2\alpha_{k,\min}\epsilon_{k,\min} < q_k < 1 - \alpha_{k,\min}\epsilon_{k,\min}. \tag{103}$$

Finally, $\lim_{k\to\infty} \alpha_{k,\min}\epsilon_{k,\min} = 0$ by Assumption 14(ii), so that (103) implies that $q_k \in (0,1)$ for sufficiently large k. Using this fact and (102) we get the following estimate:

$$\begin{aligned} 0 < q_k(1 + \alpha_{k,\min}\epsilon_{k,\min}) &\leq q_k + \alpha_{k,\min}\epsilon_{k,\min} \\ &= 1 - \alpha_{k,\min}\epsilon_{k,\min}(2 - c_k) + \alpha_{k,\min}\epsilon_{k,\min} \\ &= 1 - \alpha_{k,\min}\epsilon_{k,\min}(1 - c_k) \\ &\leq 1 - \alpha_{k,\min}\epsilon_{k,\min}(1 - c), \end{aligned} \tag{104}$$

using (102) in the last inequality.

Combining (100) and (104) we obtain

$$\mathbb{E}\big[\|x^{k+1} - t^k\|^2 \big| \mathcal{F}_k\big] \leq (1 - a_k)\|x^k - t^{k-1}\|^2 + b_k \tag{105}$$

for all sufficiently large k, with $a_k := \alpha_{k,\min}\epsilon_{k,\min}(1 - c)$ and

$$\begin{aligned} b_k &:= \left[20(B_t^2 + M_t^2\epsilon_{k,\max}^2) + \frac{2(B_t + M_t\epsilon_{k,\max})^2}{A_{\min}}\right]\alpha_{k,\max}^2 + \\ &\frac{(B_t + \epsilon_{k,\max}M_t)^2\Delta_k^2}{2\mu\alpha_{k,\min}\epsilon_{k,\min}} q_k\Big(M_t\frac{\epsilon_{k-1,\max} - \epsilon_{k,\min}}{\epsilon_{k,\min}}\Big)^2\Big(1 + \frac{1}{\alpha_{k,\min}\epsilon_{k,\min}}\Big). \end{aligned} \tag{106}$$

From (104) and the fact that $c \in (0,1)$, we conclude that $a_k \in [0,1]$, while from Assumption 14(iii) we have that $\sum_k a_k = \infty$. and from Assumption 14(iv) and (106) we get $\sum_k b_k < \infty$. Finally, we obtain from (106):

$$0 \leq \frac{b_k}{a_k} = c_1\frac{\alpha_{k,\max}^2}{\alpha_{k,\min}\epsilon_{k,\min}} + c_2\left(\frac{\Delta_k}{\alpha_{k,\min}\epsilon_{k,\min}}\right)^2 + c_3\frac{(\epsilon_{k-1,\max} - \epsilon_{k,\min})^2}{\epsilon_{k,\min}^3\alpha_{k,\min}}\Big(1 + \frac{1}{\alpha_{k,\min}\epsilon_{k,\min}}\Big)$$

for some positive constants c_1, c_2 and c_3. Therefore, we get that $\lim_{k\to\infty} b_k/a_k = 0$ from Assumption 14(ii) and (v). These conditions, Theorem 2 and (105) imply that $\lim_{k\to\infty} \|x^k - t^{k-1}\| = 0$ almost surely. The result follows from this fact and Lemma 5. □

References

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

[2] Bello Cruz, J.Y. and Iusem, A.N., 2010. "Convergence of direct methods for paramonotone variational inequalities", Computational Optimization and Applications Vol. 46, pp. 247-263.

[3] Bello Cruz, J.Y. and Iusem, A.N., 2012. "An explicit algorithm for monotone variational inequalities", Optimization, Vol. 61, pp. 855-871.

[4] Bello Cruz, J.Y. and Iusem A.N. "Full convergence of an approximate projections method for nonsmooth variational inequalities", to be published in Mathematics and Computers in Simulation.

[5] Bertsekas, D.P, 1995 "Nonlinear Programming". Athena Scientific, Belmont.

[6] Burke, J.V. and Ferris, M.C., 1993. “Weak sharp minima in mathematical programming”, SIAM Journal on Control and Optimization, Vol. 31, pp. 1340-1359.

[7] Censor, Y., 1981. “Row-action methods for huge and sparse systems and its applications”, SIAM Review, Vol. 23, pp. 444-464.

[8] Censor, Y. and Gibali, A., 2008. “Projections onto super-half-spaces ofor monotone variational inequality problems in finite-dimensional spaces”, Journal of Nonlinear and Convex Analysis, Vol. 9, pp. 461-474.

[9] Chen, Y., Lan, G. and Ouyang, Y., “Accelerated schemes for a class of variational inequalities”, to be published.

[10] Chen, X., Wets, R.J-B and Zhang, Y., 2012. “Stochastic Variational Inequalities: Residual Minimization Smoothing/Sample Average approximations”. SIAM Journal on Optimization, Vol. 22, pp. 649-673.

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

[12] Facchinei, F. and Pang, J.-S., 2003. “Finite-Dimensional Variational Inequalities and Complementarity Problems”, Springer, New York.

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

[14] Iusem, A.N., 1998, “On some properties of paramonotone operators”, Journal of Convex Analysis, Vol. 5, pp. 269-278.

[15] Iusem, A.N. and Svaiter, B.F., 1997. “A variant of Kopelevich’s method for variational inequalities with a new search strategy”, Optimization, Vol. 42, pp.309-321.

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

[17] Juditsky, A., Nemirovski, A. and Tauvel, C., 2011. “Solving variational inequalities with stochastic mirror-prox algorithm”, Stochastic Systems, Vol. 1, pp. 17–58.

[18] Korpelevich, G.M., 1976. “The extragradient method for finding saddle points and other problems”, Ekonomika i Matematcheskie Metody, Vol. 12, pp. 747-756.

[19] Koshal, J., Nedić, A. and Shanbhag, U.V., “Regularized Iterative Stochastic Approximation Methods for Stochastic Variational Inequality Problems”, to be published in IEEE Transactions on Automatic Control.

[20] Marcotte, P. and Zhu, D., 1998. “Weak sharp solutions of variational inequalities”. SIAM Journal on Optimization, Vol. 9, pp. 179189.

[21] Nemirovski, A., 2004. “Prox-method with rate of convergence O(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems”. SIAM Journal on Optimization, Vol. 15, pp. 229-251.

[22] Polyak, B., 1987. “Introduction to Optimization”. Optimization Software, New York.

[23] Robbins, H. and Monro, S., 1951. “A Stochastic Approximation Method”, The Annals of Mathematical Statistics, Vol. 22, pp. 400-407.

[24] Robbins, H. and Siegmund, D.O., 1971. “A convergence theorem for nonnegative almost super- martingales and some applications”, Optimizing Methods in Statistics, pp. 233-257.

[25] Rockafellar, T.R., 1976. “Augmented Lagrangians and applications of the proximal point algorithm in convex programming”, Mathematics of Operations Research, Vol. 1 pp. 97-116.

[26] Tikhonov, A., 1963. “On the solution of incorrectly put problems and the regularisation method”. In Outlines Joint Symposia. Partial Differential Equations (Novosibirsk). Academy of Sciences of the USSR, Siberian Branch, Moscow, 1963, pp. 261-265.

[27] Wang, M. and Bertsekas, D., “Incremental Constraint Projection Methods for Variational Inequalities”, to be published in Mathematical Programming.

[28] Yousefian, F., Nedić, A. and Shanbhag, U.V., “Optimal robust smoothing extragradient algorithms for stochastic variational inequality problems”, to be published.

[29] Zhang, J., Wan. C. and Xiu, N. 2003. “The Dual Gap Function for Variational Inequalities”, Applied Mathematics and Optimization, Vol. 48, pp. 129-148.