EXTRAGRADIENT METHOD WITH VARIANCE REDUCTION FOR STOCHASTIC VARIATIONAL INEQUALITIES[*]

A. N. IUSEM[†], A. JOFRÉ[‡], R. I. OLIVEIRA[†],
AND P. THOMPSON[†]

Abstract. We propose an extragradient method with stepsizes bounded away from zero for stochastic variational inequalities requiring only pseudomonotonicity. We provide convergence and complexity analysis, allowing for an unbounded feasible set, unbounded operator, and nonuniform variance of the oracle, and, also, we do not require any regularization. Alongside the stochastic approximation procedure, we iteratively reduce the variance of the stochastic error. Our method attains the optimal oracle complexity $\mathcal{O}(1/\epsilon^2)$ (up to a logarithmic term) and a faster rate $\mathcal{O}(1/K)$ in terms of the mean (quadratic) natural residual and the D-gap function, where K is the number of iterations required for a given tolerance $\epsilon > 0$. Such convergence rate represents an acceleration with respect to the stochastic error. The generated sequence also enjoys a new feature: the sequence is bounded in L^p if the stochastic error has finite p-moment. Explicit estimates for the convergence rate, the oracle complexity, and the p-moments are given depending on problem parameters and distance of the initial iterate to the solution set. Moreover, sharper constants are possible if the variance is uniform over the solution set or the feasible set. Our results provide new classes of stochastic variational inequalities for which a convergence rate of $\mathcal{O}(1/K)$ holds in terms of the mean-squared distance to the solution set. Our analysis includes the distributed solution of pseudomonotone Cartesian variational inequalities under partial coordination of parameters between users of a network.

Key words. stochastic variational inequalities, pseudomonotonicity, extragradient method, stochastic approximation, variance reduction

AMS subject classifications. 65K15, 90C33, 90C15, 62L20

DOI. 10.1137/15M1031953

1. Introduction. The standard (deterministic) variational inequality problem, which we will denote as VI(T, X) or simply VI, is defined as follows: given a closed and convex set $X \subset \mathbb{R}^n$ and a single-valued operator $T : \mathbb{R}^n \to \mathbb{R}^n$, find $x^* \in X$ such that for all $x \in X$,

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

We shall denote by X^* the solution set of VI(T, X). The variational inequality problem includes many interesting special classes of variational problems with applications in economics, game theory, and engineering. The basic prototype is smooth convex optimization, when T is the gradient of a smooth function. Other problems which can be formulated as variational inequalities, include *complementarity problems* (when $X = \mathbb{R}^n_+$), *systems of equations* (when $X = \mathbb{R}^n$), *saddle-point problems*, and many

---

[†]Instituto de Matemática Pura e Aplicada (IMPA), Rio de Janeiro, RJ, Brazil (iusp@impa.br, rimfo@impa.br, philip@impa.br).

[‡]Centro de Modelamiento Matemático (CMM & DIM), Santiago, Chile (ajofre@dim.uchile.cl).

equilibrium problems. We refer the reader to Chapter 1 of [11] and [12] for an extensive review of applications of the VI problem in engineering and economics. The complementarity problem and system of equations are important classes of problems where the feasible set is unbounded.

In the stochastic case, we start with a measurable space $(\Xi, \mathcal{G})$, a measurable (random) operator $F : \Xi \times \mathbb{R}^n \to \mathbb{R}^n$, and a random variable $\xi : \Omega \to \Xi$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ which induces an expectation $\mathbb{E}$ and a distribution $\mathbb{P}_\xi$ of ξ. When no confusion arises, we sometimes use ξ to also denote a random sample $\xi \in \Xi$. We assume that for every $x \in \mathbb{R}^n$, $F(\xi, x) : \Omega \to \mathbb{R}^n$ is an integrable random vector. The solution criterion analyzed in this paper consists of solving $\text{VI}(T, X)$ as defined by (1), where $T : \mathbb{R}^n \to \mathbb{R}^n$ is the expected value of $F(\xi, \cdot)$, i.e.,

$$(2) \qquad T(x) = \mathbb{E}[F(\xi, x)] = \int_\Omega F(\xi(\omega), x) \, d\mathbb{P}(\omega) \quad \forall x \in \mathbb{R}^n.$$

For clarity, we state such formulation of the *stochastic variational inequality* (SVI) problem in the following definition.

DEFINITION 1.1 (SVI). *Assuming that $T : \mathbb{R}^n \to \mathbb{R}^n$ is given by $T(x) = \mathbb{E}[F(\xi, x)]$ for all $x \in \mathbb{R}^n$, the* SVI *problem consists of finding $x^* \in X$ such that $\langle T(x^*), x - x^* \rangle \geq 0$ for all $x \in X$.*

Such formulation of SVI is often called the *expected value* formulation. The introduction of this formulation goes back to [23, 14], as a natural generalization of stochastic optimization problems (SP). We remark here that different formulations for the SVI problem exist in which the randomness is treated differently. For instance, in the so-called expected residual minimization (ERM) formulation, one defines a suitable nonnegative function $x \mapsto h(\xi, x)$ whose zeros are solutions of $\text{VI}(F(\xi, \cdot), X)$. The ERM formulation is then defined as the problem $\min_{x \in X} \mathbb{E}[h(\xi, x)]$. Variants of the ERM formulation can include random constraints as well. Both expected value and ERM formulations have relevance in modeling stochastic equilibrium problems in different settings. See, e.g., [34, 6].

Methods for the deterministic $\text{VI}(T, X)$ have been extensively studied (see [11]). If T is fully available, then SVI can be solved by these methods. As in the case of SP, the SVI in Definition 1.1 becomes very different from the deterministic setting when T is *not available.* This is often the case in practice due to expensive computation of the expectation in (2), unavailability of $\mathbb{P}_\xi$, or no closed form for $F(\xi, \cdot)$. This requires sampling the random variable ξ and the use of values of $F(\eta, x)$ given a sample η of ξ and a current point $x \in \mathbb{R}^n$ (a procedure often called a "stochastic oracle" call). In this context, there are two current methodologies for solving the SVI problem: *sample average approximation* (SAA) and *stochastic approximation* (SA). In this paper we focus on the SA approach. For analysis of the SAA methodology for SP, see e.g., [37] and references therein. For the analysis of the SAA methodology for solving SVIs see e.g. [14, 40, 41, 42].

We make some remarks regarding the solution of the SVI problem in Definition 1.1 when using SA. The SA methodology for SP or SVI can be seen as a projection-type method where the exact mean operator T is replaced along the iterations by a random sample of F. This approach induces a stochastic error $F(\xi, x) - T(x)$ for $x \in X$ in the trajectory of the method. In this solution method the generated sequence $\{x^k\}$ is unavoidably a stochastic process which evolves recursively according to the chosen projection algorithm and the sampling information used in every iteration. As a consequence, asymptotic convergence of the SA method guarantees a solution of

Definition 1.1 with *total probability*. Precisely, limit points of the sequence $\{x^k\}$ are typically a random variable x^* such that, with *total probability*, $x^* \in X^*$.

The first analysis of SA methodology for SVI was carried out recently in [18]. When $X = \mathbb{R}^n$, Definition 1.1 becomes the stochastic equation (SE) problem, that is to say, under (2), almost surely (a.s.) finds $x^* \in \mathbb{R}^n$ such that $T(x^*) = 0$. The SA methodology was first proposed by Robbins and Monro in [35] for the SE problem in the case in which T is the gradient of a strongly convex function under specific conditions. Since this fundamental work, SA approaches for SP and, more recently, for SVI have been carried out [18, 19, 43, 26, 39, 17, 7, 44, 21, 22, 45]. See also [27, 1] for other problems where the SA procedure is relevant (such as machine learning, online optimization, repeated games, queueing theory, signal processing, and control theory).

1.1. Related work on SA and contributions. The first SA method for SVI was analyzed in [18]. Their method is

$$(3) \qquad x^{k+1} = \Pi[x^k - \alpha_k F(\xi^k, x^k)],$$

where Π is the Euclidean projection onto X, $\{\xi^k\}$ is a sample of ξ and $\{\alpha_k\}$ is a sequence of positive steps. In [18], the almost sure convergence is proved assuming L-Lipschitz continuity of T, strong monotonicity or strict monotonicity of T, stepsizes satisfying $\sum_k \alpha_k = \infty, \sum_k \alpha_k^2 < \infty$ (with $0 < \alpha_k < 2\rho/L^2$, assuming that T is ρ-strongly monotone), and an unbiased oracle with uniform variance, i.e., there exists $\sigma > 0$ such that for all $x \in X$,

$$(4) \qquad \mathbb{E}\left[\|F(\xi, x) - T(x)\|^2\right] \leq \sigma^2.$$

After the above mentioned work, recent research on SA methods for SVI have been developed in [19, 43, 26, 39, 17, 7, 44, 21, 22, 45]. Two of the main concerns in these papers were the extension of the SA approach to the general monotone case and the obtention of (optimal) convergence rate and complexity results with respect to known metrics associated to the VI problem. In order to analyze the monotone case, SA methodologies based on the extragradient method of Korpelevich [25] and the mirror-prox algorithm of Nemiroviski [31] were used in [19, 7, 44, 21, 22], and iterative Tykhonov and proximal regularization procedures (see [20, 24]) were used in [43, 26, 17, 45]. Other objectives in some of these papers were the use of incremental constraint projections in the case of difficulties accessing the feasible set [39, 17], the convergence analysis in the absence of the Lipschitz constant [43, 44, 45], and the distributed solution of Cartesian variational inequalities [43, 26, 17, 20].

In Cartesian variational inequalities, a network of m agents is associated to a coupled variational inequality with constraint set $X = X^1 \times \cdots \times X^m$ and operator $F = (F_1, \ldots, F_m)$, where the ith agent is associated to a constraint set $X^i \subset \mathbb{R}^{n_i}$ and a map $F_i : \Xi \times \mathbb{R}^n \to \mathbb{R}^{n_i}$ such that $n = \sum_{i=1}^m n_i$. Two important problems which can be formulated as stochastic Cartesian variational inequalities are the stochastic *Nash equilibria* and the stochastic *multiuser optimization* problem. See [26] for a precise definition. In these problems, the ith agent has only access to constraint set X^i and F_i (which depends on other agents' decision sets) so that a distributed solution of the SVI is required. As an example, the distributed variant of method (3) studied in [43] takes the following form: for all $i = 1, \ldots, m$,

$$x_i^{k+1} = \Pi_i \left[x_i^k - \alpha_{k,i} F_i(\xi_i^k, x^k)\right],$$

where Π_i is the Euclidean projection onto X^i. Thus, the ith agent updates its decision evaluating his operator F_i and projecting onto its decision set X_i.

In this paper we propose the following extragradient method: given x^k, define

$$(5) \qquad z^k = \Pi\left[x^k - \frac{\alpha_k}{N_k}\sum_{j=1}^{N_k} F(\xi_j^k, x^k)\right],$$

$$(6) \qquad x^{k+1} = \Pi\left[x^k - \frac{\alpha_k}{N_k}\sum_{j=1}^{N_k} F(\eta_j^k, z^k)\right],$$

where $\{N_k\} \subset \mathbb{N}$ is a nondecreasing sequence and $\{\xi_j^k, \eta_j^k : k \in \mathbb{N}, j = 1, \ldots, N_k\}$ are independent identically distributed (i.i.d.) samples of ξ. We call $\{N_k\}$ the *sample rate* sequence.

Next we make some observations regarding merit functions and complexity estimates. A merit function for $\text{VI}(T, X)$ is a nonnegative function f over X such that $X^* = X \cap f^{-1}(0)$. An unrestricted merit function f for $\text{VI}(T, X)$ is a merit function such that $X^* = f^{-1}(0)$. For any $\alpha > 0$ we consider the *natural residual function* r_α, defined, for any $x \in \mathbb{R}^n$, by $r_\alpha(x) := \|x - \Pi(x - \alpha T(x))\|$. It is well known that r_α is an unrestricted merit function for $\text{VI}(T, X)$. Given $\epsilon > 0$, we consider an iteration index $K = K_\epsilon$ (whose existence will be proved in section 3.4), such that $\mathbb{E}[r_\alpha(x^K)^2] < \epsilon$, and we look at $\mathbb{E}[r_\alpha(x^K)^2]$ as a *nonasymptotic convergence rate*. In particular, we will have an $\mathcal{O}(1/K)$ convergence rate if $\mathbb{E}[r_\alpha(x^K)^2] \le Q/K$ for some constant $Q > 0$ (depending on the initial iterate and the parameters of the problem and the method). The (stochastic) *oracle complexity* will be defined as the total number of oracle calls needed for $\mathbb{E}[r_\alpha(x^K)^2] < \epsilon$ to hold, i.e., $\sum_{k=1}^{K} 2N_k$.

Beside the natural residual, other merit functions were considered in prior work on SVI. Given a compact feasible set X, the *dual gap function* of $\text{VI}(T, X)$ is defined as $G(x) := \sup_{y \in X}\langle T(y), x - y\rangle$ for $x \in X$. In [19, 7, 44, 45], rates of convergence were given in terms of the expected value of G when X is compact or, when X is unbounded, in terms of the relaxed dual gap function $\tilde{G}(x, v) := \sup_{y \in X}\langle T(y) - v, x - y\rangle$, introduced by Monteiro and Svaiter [29, 30], based on the enlargement of monotone operators introduced in [3]. When X is compact, the dual gap function is a modification of the *primal gap function*, defined as $g(x) := \sup_{y \in X}\langle T(x), x - y\rangle$ for $x \in X$. Both the primal and dual gap functions are continuous only if X is compact. A variation suitable for unbounded feasible sets is the *regularized gap function*, defined, for fixed $a > 0$, as $g_a(x) := \sup_{y \in X}\{\langle T(x), x - y\rangle - \frac{a}{2}\|x - y\|^2\}$, for $x \in \mathbb{R}^n$. The regularized gap function is continuous over $\mathbb{R}^n$. Another variation is the so-called D-gap function. It is defined, for fixed $b > a > 0$, as $g_{a,b}(x) := g_a(x) - g_b(x)$, for $x \in \mathbb{R}^n$. It is well known that $g_{a,b} : \mathbb{R}^n \to \mathbb{R}_+$ is a continuous unrestricted merit function for $\text{VI}(T, X)$. Moreover, the quadratic natural residual and the D-gap function are equivalent merit functions in the sense that, given $b > a > 0$, there are constants $c_1, c_2 > 0$ such that for all $x \in \mathbb{R}^n$, $c_1 r_{b^{-1}}(x)^2 \le g_{a,b}(x) \le c_2 r_{a^{-1}}(x)^2$ (see [11, Theorems 10.2.3, 10.3.3, Proposition 10.3.7]). These properties hold independently of the compactness of X.

Next we resume the contributions of the algorithm presented in this paper. (i) *Asymptotic convergence*. Assuming *pseudomonotonicity* of F, and using an extragradient scheme, without regularization, we prove that, a.s., the generated sequence is bounded, its distance to the solution set converges to zero, and its natural residual value converges to zero a.s. and in L^2. Note that monotonicity implies pseudomonotonicity. See [22] for examples where the more general setting of pseudomonotonicity is

relevant (stochastic fractional programming, stochastic optional pricing, and stochastic economic equilibria). The sequence generated by our method also possesses a new stability feature: for $p = 2$ or any $p \geq 4$, if the random operator has finite p-moment, then the sequence is bounded in L^p, and we are able to provide explicit upper bounds in terms of the problem parameters. Previous work required a bounded monotone operator, specific forms of pseudomonotonicity (monotonicity with acute angle, pseudomonotonicity-plus, strict pseudomonotonicity, symmetric pseudomonotonicity, or strong pseudomonotonicity as in [21, 22]) or regularization procedures. The disadvantage of regularization procedures in the absence of strong monotonicity is the need to introduce additional coordination between the stepsize sequence and the regularization parameters. Also, the regularization induces a suboptimal performance in terms of rate and complexity (see [45]).

(ii) *Faster convergence rate with oracle complexity efficiency.* To the best of our knowledge, our work is the first SA method for SVI with *stepsizes bounded away from zero.* Such a feature allows our method to achieve a faster convergence rate $\mathcal{O}(1/K)$ in terms of the mean-squared natural residual under plain pseudomonotonicity (with no regularization requirements). As a consequence, our method achieves a convergence rate of $\mathcal{O}(1/K)$ in terms of the mean D-gap function value of the generated sequence. In previous works, methods with diminishing stepsizes satisfying $\sum_k \alpha_k = \infty$, $\sum_k \alpha_k^2 < \infty$ were used, achieving a $\mathcal{O}(1/K)$ rate in terms of the mean-squared distance to X^*, with more demanding monotonicity assumptions (namely, bounded strongly pseudomonotone operators and bounded monotone weak-sharp VI) and a rate $\mathcal{O}(1/\sqrt{K})$ in terms of mean gap function values of the ergodic average of the generated sequence in the case of bounded monotone operators. Importantly, our method preserves the optimal oracle complexity $\mathcal{O}(\epsilon^{-2})$ up to a first order logarithmic term. By accelerating the rate, we reduce the computational complexity (in terms of projection computations), preserving a near-optimal oracle complexity. It should be noted that such acceleration represents the closing of the gap from the stochastic to the deterministic and it is distinct in nature from the acceleration of differentiable convex optimization problems using Nesterov-type gradient methods. We provide explicit upper bounds for the rate and complexity in terms of the problem parameters. As a corollary of our result we provide new classes of SVIs for which a convergence rate of $\mathcal{O}(1/K)$ holds in terms of the mean-squared distance to the solution set (see section 4). We remark that for compact X, it is possible to show that the proposed extragradient method achieves a rate $\mathcal{O}(\frac{\ln K}{K})$ in terms of the mean dual gap function value of the ergodic average of the generated sequence with an optimal oracle complexity (up to a logarithmic factor). If a different set of weights is used in the ergodic average (such as window-based averaging, Nesterov-like extrapolation, and other schemes), then our method achieves a rate $\mathcal{O}(\frac{1}{K})$ with an optimal oracle complexity. See, e.g., [32, 44, 7]. In the context of large dimension data ($n \gg 1$), our algorithm complexity is independent of the dimension n (see Proposition 3.23).

(iii) *Unbounded setting.* The results in items (i)–(ii) are valid for an *unbounded feasible set* and an *unbounded operator.* Important examples of such a setting include complementarity problems and systems of equations. Asymptotic convergence for an unbounded feasible set is analyzed in [43, 39, 22, 45] with more demanding monotonicity hypotheses and in [26, 17] for the monotone case, but with an additional regularization procedure. To the best of our knowledge, convergence rates in the case of an unbounded feasible set were treated only in [39, 7]. In [39], a convergence rate

is given only for strongly monotone operators. In [7], assuming *uniform variance* over X (in the sense of (4)), a convergence rate of $\mathcal{O}(1/\sqrt{K})$ for the ergodic average of the iterates is achieved in terms of the mean value of a relaxed gap function recently introduced by Monteiro and Svaiter [29, 30]. It should be noted, however, that, even when assuming uniform variance, the sequence of the iterates generated by their method may diverge to ∞ (see Example 3.10). Our convergence analysis in items (i)–(ii) does not depend upon boundedness assumptions, and we prove the faster rate of $\mathcal{O}(1/K)$ in terms of the mean (quadratic) natural residual and the mean D-gap function, which are new results. The natural residual and the D-gap function are better behaved than the (standard) gap function: the former are finite valued and Lipschitz continuous over $\mathbb{R}^n$, while the later is finite valued and continuous only for a compact X.

(iv) *Nonuniform variance*. According to what we know, all previous works require that the variance of the oracle error be *uniform* over X (in the sense of (4)), except in [39] for the strongly monotone case, and in [17] for the case of a weak-sharp monotone operator, and also for the monotone case with an iterative Tykhonov regularization (with no convergence rate results). Such uniform variance assumption holds for bounded operators, but not for unbounded ones, on an unbounded feasible set. Typical situations where this assumption fails to hold include affine complementarity problems and systems of equations. In such cases, the variance of the oracle error tends (quadratically) to ∞ in the horizon (see Example 3.9). The performance of our method, in terms of the oracle complexity, depends on the point $x^* \in X^*$ with *minimal* trade-off between variance and distance to initial iterates "ignoring" points with high variance (see comments after Theorem 3.22 and section 3.4.1). This result also improves over the case in which (4) *does* holds but $\sigma(x^*)^2 \ll \sigma^2$ or over the case in which X is compact but $\|x^0 - x^*\| \ll \text{diam}(X)$. In conclusion, the performance of method (5)–(6) depends on solution points x^* with minimal variance, compared to the conservative upper bound σ^2, and minimal distance to initial iterates. In the case of uniform variance over X^* or X, we obtain sharper estimates of rate and complexity in item (ii).

(v) *Distributed solution of multiagent system*. The analysis in items (i)–(iv) also holds true for the distributed solution of stochastic Cartesian variational inequalities, in the spirit of [43, 26, 17, 20]. In our framework (see Algorithm 2, (11)–(12)), agents update synchronous stepsizes bounded away from zero over a range $(0, \mathcal{O}(1)L^{-1})$. An advantage of the extragradient approach in the distributed case is that we do not require iterative regularization procedures as in [26, 17, 20] for coping with the merely monotone case. This implies that the faster convergent rate of $\mathcal{O}(1/K)$ is achievable with a near-optimal oracle complexity under weaker conditions (such as an unbounded set and nonuniform variance). As discussed later on, our algorithm requires the choice of a sampling rate for dealing with the setting of items (i)–(iv). Hence, in the distributed solution case, agents should have the choice of sharing their oracle calls or not, and we allow both options. In the later case of fully distributed sampling, the oracle complexity has higher order dependence in terms of the network dimension m, which may be demanding in the context of large networks ($m \gg 1$). For this case, if an estimate of m is available and a decreasing sequence of (deterministic) parameters $\{b_i\}_{i=1}^m$ is shared (in any order) among agents, then our algorithm has oracle complexity of order $m(a^{-1}\epsilon^{-1})^{2+a}$ for arbitrary $a > 0$ (up to a scaling factor in the sample rate). See Proposition 3.28. Further dimension reduction possibilities will be the subject of future work.

For achieving the results of items (i)–(v), we employ an iterative *variance reduction* procedure. This means that instead of calling the oracle once per iteration (as in previous SA methods for SVI studied so far), our method calls the oracle N_k times at iteration k and uses the associated empirical average of the values of the random operator F at the current iterates x^k and z^k (see (5)–(6)). Since the presence of the stochastic error destroys the strict Fejér property (satisfied by the generated sequence in the deterministic setting), the mentioned variance reduction procedure is the mechanism that allows our extragradient method to converge in an unbounded setting with stepsizes bounded away from zero and to achieve an accelerated rate in terms of the natural residual. Such variance reduction scheme is efficient since we maintain a near-optimal oracle complexity when compared to the classical SA method. Precisely, given a prescribed tolerance $\epsilon > 0$, the classical SA method requires $\mathcal{O}(\epsilon^{-2})$ iterations, $\mathcal{O}(\epsilon^{-2})$ samples and a final ergodic average of size $\mathcal{O}(\epsilon^{-2})$. As will be seen in Proposition 3.23, our method requires $K := \mathcal{O}(\epsilon^{-1})$ iterations and $\mathcal{O}(\epsilon^{-2})$ samples and, for $k \leq K$, the kth iteration computes an empirical average of size k (up to a first order logarithmic factor). Hence the total cost in averaging is also $\mathcal{O}(\epsilon^{-2})$ (again, up to a first order logarithmic factor). In conclusion, our method uses the same amount of total samples and same effort in averaging as in the classical SA method but with empirical averages with smaller sizes distributed along iterations instead of one ergodic average at the final iteration. This is the reason for improving the required number of iterations from $\mathcal{O}(\epsilon^{-2})$ to $\mathcal{O}(\epsilon^{-1})$ and thus reducing the number of projections by one order.[1] The use of empirical averages along iterations is also the reason we can include unbounded operators and oracles with nonuniform variance and give estimates which depend on the variance at points of the trajectory of the method and at points of X^* (but not on the whole X). Such results are not shared by the SAA method and SA with constant N_k. In order to obtain these results, we use martingale moment inequalities and a supermartingale convergence theorem (see section 2.2). Our sampling procedure also possesses a *robust* property: a scaling factor on the sampling rate maintains the progress of the algorithm with proportional scaling in the convergence rate and oracle complexity (see Propositions 3.23, 3.26, and 3.28 and [32] for robust methods). In Examples 3.9 and 3.10 we show typical situations where such variance reduction procedure is relevant or even necessary.

To the best of our knowledge the variance reduction procedure mentioned above is new for SA solution of SVI. Moreover, it seems that the derivation of the faster rate of $\mathcal{O}(1/K)$ with a near-optimal stochastic oracle complexity, an unbounded feasible set, and an oracle with nonuniform variance is also new for convex stochastic programming. During the preparation of this paper we became aware of references [9, 5, 13, 16, 8, 38], where variable sample-size methods are studied for stochastic optimization. We treat the general case of pseudomonotone variational inequalities with weaker assumptions. Also, our analysis differs somewhat from these works relying on martingale and optimal stopping techniques. In [9, 5, 13, 38] the SA approach is studied for convex SP. In [5, 38], the focus is on gradient descent methods applied to strongly convex optimization problems. In [13] the strong convexity property is slightly weakened by assuming a special error bound on the solution set (which is satisfied by strongly convex optimization problems in particular). In [5, 13] the

---

[1] The possibility of distributing the empirical averages along iterations is possible due to the on-line nature of the SA method. This is not shared by the SAA methodology, which is an off-line procedure.

optimization problem is unconstrained, while in [38] the problem has a compact feasible set. In [5], second order information is assumed and an adaptive sample size selection is used. In [13] uniform boundedness assumptions are required. In [9], a variant of the dual averaging method of Nesterov [33] is applied for solving nonsmooth stochastic convex optimization, assuming a compact feasible set and uniform variance. A constant oracle call per iteration $N_k \equiv N > 1$ is used, obtaining a convergence rate of $\mathcal{O}(1/\sqrt{KN})$ for the ergodic average of the sequence, while we typically use $N_k = \mathcal{O}(k(\ln k)^{1+b})$ with $b > 0$ obtaining a rate of $\mathcal{O}(1/K)$ for the generated sequence. In [16, 8], the SAA approach for stochastic optimization is studied. This is an implicit method, unlike the SA methodology. Also, uniform boundedness assumptions are required. In [8] the focus is on unconstrained optimization, with second order information, using Bayesian analysis for an adaptive choice of N_k. See also [15].

The paper is organized as follows. In section 2 we present notation and preliminaries, including the required probabilistic tools. In section 3 we present the algorithm and its convergence analysis. In subsection 3.1 the algorithm is formally presented, while in subsection 3.2 the assumptions required for its analysis are discussed. Subsection 3.3 presents the convergence analysis, while subsection 3.4 focuses on convergence rates and complexity results.

2. Preliminaries.

2.1. Projection operator and notation. 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 Euclidean norm. Given $C \subset \mathbb{R}^n$ and $x \in \mathbb{R}^n$, we use the notation $\mathrm{d}(x, C) := \inf\{\|x - y\| : y \in C\}$. For a closed and convex set $C \subset \mathbb{R}^n$, we use the notation $\Pi_C(x) := \operatorname{argmin}_{y \in C} \|y - x\|^2$ for $x \in \mathbb{R}^n$. Given $H : \mathbb{R}^n \to \mathbb{R}^n$, $\mathrm{S}(H, C)$ denotes the solution set of $\mathrm{VI}(H, C)$. For a matrix $B \in \mathbb{R}^{n \times n}$, we use the notation $\|B\| := \sup_{x \neq 0} \|Bx\|/\|x\|$. We use the notation $[m] := \{1, \ldots, m\}$ for $m \in \mathbb{N}$ and $(\alpha_i)_{i=1}^m := (\alpha_1, \ldots, \alpha_m)$ for $\alpha_i \in \mathbb{R}$ and $i \in [m]$. We also use the notation $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. We use the abbreviation "RHS" for "right-hand side." Given sequences $\{x^k\}$ and $\{y^k\}$, we use the notation $x^k = \mathcal{O}(y^k)$ or $\|x^k\| \lesssim \|y^k\|$ to mean that there exists a constant $C > 0$ such that $\|x^k\| \leq C\|y^k\|$ for all k. The notation $\|x^k\| \sim \|y^k\|$ means that $\|x^k\| \lesssim \|y^k\|$ and $\|y^k\| \lesssim \|x^k\|$. 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. We denote by $\mathrm{cov}[B]$ the covariance of a random vector B. Also, we write $\xi \in \mathcal{F}$ for "ξ is $\mathcal{F}$-measurable." We denote by $\sigma(\xi_1, \ldots, \xi_k)$ the σ-algebra generated by the random variables $\xi_1, \ldots, \xi_k$. Given the random variable ξ and $p \geq 1$, $|\xi|_p$ is the L^p-norm of ξ and $|\xi|\mathcal{F}|_p := \sqrt[p]{\mathbb{E}[|\xi|^p |\mathcal{F}]}$ is the L_p-norm of ξ conditional to the σ-algebra $\mathcal{F}$. $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. Given $x \in \mathbb{R}$, we denote by $x_+ := \max\{0, x\}$ its positive part and by $\lceil x \rceil$ the smallest integer greater than or equal to x.

The following properties of the projection operator are well known; see Chapter 1 of [11].

LEMMA 2.1. *Take a nonempty closed and convex set $C \subset \mathbb{R}^n$. Then*
(i) *given $x \in \mathbb{R}^n$, $\Pi_C(x)$ is the unique point of C satisfying the following property: $\langle x - \Pi_C(x), y - \Pi_C(x) \rangle \leq 0$, for all $y \in C$;*
(ii) *for all $x \in \mathbb{R}^n, y \in C$; $\|\Pi_C(x) - y\|^2 + \|\Pi_C(x) - x\|^2 \leq \|x - y\|^2$;*
(iii) *for all $x, y \in \mathbb{R}^n$ $\|\Pi_C(x) - \Pi_C(y)\| \leq \|x - y\|$;*
(iv) *given $\alpha > 0$ and $H : \mathbb{R}^n \to \mathbb{R}^n$, $\mathrm{S}(H, C) = \{x \in \mathbb{R}^n : x = \Pi_C[x - \alpha H(x)]\}$.*

2.2. Probabilistic tools. As in other SA methods, a fundamental tool to be used is the following convergence theorem of Robbins and Siegmund (see [36]).

THEOREM 2.2. *Let* $\{y_k\}, \{u_k\}, \{a_k\}, \{b_k\}$ *be sequences of nonnegative integrable random variables, adapted to the filtration* $\{\mathcal{F}_k\}$, *such that a.s. for all* $k \in \mathbb{N}$, $\mathbb{E}\big[y_{k+1}\big|\mathcal{F}_k\big] \leq (1 + a_k)y_k - u_k + b_k$, $\sum a_k < \infty$, *and* $\sum b_k < \infty$. *Then, a.s.* $\{y_k\}$ *converges and* $\sum u_k < \infty$.

If a.s. for all $k \in \mathbb{N}$, $\mathbb{E}[y_{k+1}|\mathcal{F}_k] = y_k$, then $\{y_k, \mathcal{F}_k\}$ is called a *martingale*. We shall also need the following moment inequality; see [4, 28].

THEOREM 2.3 (the inequality of Burkholder, Davis, -and Gundy). *We denote as* $\|\cdot\|$ *the Euclidean norm in* $\mathbb{R}^n$. *For all* $\mathfrak{q} \geq 1$, *there exists* $C_{\mathfrak{q}} > 0$ *such that for any vector-valued martingale* $(y_i, \mathcal{F}_i)_{i=0}^N$ *taking values in* $\mathbb{R}^n$ *with* $y_0 = 0$, *it holds that*

$$
(7) \qquad \|\|y_N\|\|_{\mathfrak{q}} \leq \left\| \sup_{i \leq N} \|y_i\| \right\|_{\mathfrak{q}} \leq C_{\mathfrak{q}} \left| \sqrt{\sum_{k=1}^N \|y_i - y_{i-1}\|^2} \right|_{\mathfrak{q}}.
$$

For $\mathfrak{q} \geq 2$, we will use the rightmost inequality in (7) in the following simpler form, which follows from applying Minkowski's inequality ($\mathfrak{q}/2 \geq 1$):

$$
(8) \qquad \|\|y_N\|\|_{\mathfrak{q}} \leq \left\| \sup_{i \leq N} \|y_i\| \right\|_{\mathfrak{q}} \leq C_{\mathfrak{q}} \sqrt{\sum_{k=1}^N \|\|y_i - y_{i-1}\|\|_{\mathfrak{q}}^2}.
$$

3. An extragradient method with stepsizes bounded away from zero.

3.1. Statement of the algorithm. Our extragradient method takes the following form.

ALGORITHM 1 (stochastic extragradient method).
1. Initialization. *Choose the initial iterate* $x^0 \in \mathbb{R}^n$, *a positive stepsize sequence* $\{\alpha_k\}$, *the sample rate* $\{N_k\}$, *and initial samples* $\{\xi_j^0\}_{j=1}^{N_0}$ *and* $\{\eta_j^0\}_{j=1}^{N_0}$ *of the random variable* ξ.
2. Iterative step. *Given iterate* x^k, *generate samples* $\{\xi_j^k\}_{j=1}^{N_k}$ *and* $\{\eta_j^k\}_{j=1}^{N_k}$ *of* ξ *and define*

$$
(9) \qquad z^k = \Pi\left[x^k - \frac{\alpha_k}{N_k} \sum_{j=1}^{N_k} F(\xi_j^k, x^k) \right],
$$

$$
(10) \qquad x^{k+1} = \Pi\left[x^k - \frac{\alpha_k}{N_k} \sum_{j=1}^{N_k} F(\eta_j^k, z^k) \right].
$$

In (9) and (10), Π is the Euclidean projection operator onto X. Method (9)–(10) is designed so that at iteration k the random variable ξ is sampled $2N_k$ times and the empirical average of F at x is used as the approximation of $T(x)$ at each projection step.

In order to incorporate the distributed case mentioned in section 1.1, item (v), we will also analyze the case in which the SVI has a Cartesian structure. We consider the decomposition $\mathbb{R}^n = \prod_{i=1}^m \mathbb{R}^{n_i}$, with $n = \sum_{i=1}^m n_i$, and furnish this space with the direct inner product $\langle x, y \rangle := \sum_{i=1}^m \langle x_i, y_i \rangle$ for $x = (x_i)_{i=1}^m$ i and $y = (y_i)_{i=1}^m$. We suppose that the feasible set has the form $X = \prod_{i=1}^m X^i$, where $X^i \subset \mathbb{R}^{n_i}$ is a closed

and convex set for $i \in [m]$. The random operator $F : \Xi \times \mathbb{R}^n \to \mathbb{R}^n$ has the form $F = (F_1, \ldots, F_m)$, where $F_i : \Xi \times \mathbb{R}^n \to \mathbb{R}^{n_i}$ for $i \in [m]$. Given $i \in [m]$, we denote by $\Pi_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ the orthogonal projection onto X^i. We emphasize that the orthogonal projection under a Cartesian structure has a simple form: for $x = (x_i)_{i=1}^m \in \mathbb{R}^n$, we have $\Pi_X(x) = (\Pi_{X^1}(x_1), \ldots, \Pi_{X^m}(x_m))$.

In such a setting, the method takes the following form.

ALGORITHM 2 (stochastic extragradient method: distributed case).
1. Initialization. *Choose the initial iterate $x^0 \in \mathbb{R}^n$, the stepsize sequence $\alpha_k > 0$, the sample rates $N_k = (N_{k,i})_{i=1}^m \in \mathbb{N}^m$, and, for each $i \in [m]$, generate the initial samples $\{\xi_{j,i}^0\}_{j=1}^{N_{0,i}}$ and $\{\eta_{j,i}^0\}_{j=1}^{N_{0,i}}$ of the random variable ξ.*
2. Iterative step. *Given $x^k = (x_i^k)_{i=1}^m$, for each $i \in [m]$, generate samples $\{\xi_{j,i}^k\}_{j=1}^{N_{k,i}}$ and $\{\eta_{j,i}^k\}_{j=1}^{N_{k,i}}$ of ξ and define*

$$(11) \qquad z_i^k = \Pi_i \left[x_i^k - \frac{\alpha_k}{N_{k,i}} \sum_{j=1}^{N_{k,i}} F_i(\xi_{j,i}^k, x^k) \right],$$

$$(12) \qquad x_i^{k+1} = \Pi_i \left[x_i^k - \frac{\alpha_k}{N_{k,i}} \sum_{j=1}^{N_{k,i}} F_i(\eta_{j,i}^k, z^k) \right].$$

Method (9)–(10) is a particular case of method (11)–(12) with $m = 1$. The only additional requirement when $m > 1$ is the sampling coordination between agents (Assumption 3.6). We define next the stochastic errors: for each $i \in [m]$,

$$(13) \qquad \epsilon_{1,i}^k := \frac{1}{N_{k,i}} \sum_{j=1}^{N_{k,i}} F_i(\xi_{j,i}^k, x^k) - T_i(x^k),$$

$$(14) \qquad \epsilon_{2,i}^k := \frac{1}{N_{k,i}} \sum_{j=1}^{N_{k,i}} F_i(\eta_{j,i}^k, z^k) - T_i(z^k),$$

in which case method (11)–(12) is expressible in a compact form as

$$(15) \qquad z^k = \Pi[x^k - \alpha_k(T(x^k) + \epsilon_1^k)],$$

$$(16) \qquad x^{k+1} = \Pi[x^k - \alpha_k(T(z^k) + \epsilon_2^k)],$$

where $\Pi : \mathbb{R}^n \to \mathbb{R}^n$ is the projection operator onto X and $\epsilon_l^k := (\epsilon_{l,i}^k)_{i=1}^m$ for $l \in \{1, 2\}$.

3.2. Discussion of the assumptions. For simplicity of notation, we aggregate the samples as

$$\xi_i^k := \{\xi_{j,i}^k : j \in [N_{k,i}]\}, \ \xi^k := \{\xi_i^k : i \in [m]\},$$
$$\eta_i^k := \{\eta_{j,i}^k : j \in [N_{k,i}]\}, \ \eta^k := \{\eta_i^k : i \in [m]\}.$$

In the method (15)–(16), the sample $\{\xi^k\}$ is used in the first projection while $\{\eta^k\}$ is used in the second projection. In the case of a Cartesian SVI, $\{\xi_i^k\}$ and $\{\eta_i^k\}$ are the samples used in the first and second projections in (11)–(12) by the ith agent, respectively.

We shall study the stochastic process $\{x^k\}$ with respect to the filtrations

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

We observe that by induction, $x^k \in \mathcal{F}_k$ and $z^k \in \widehat{\mathcal{F}}_k$ but $z^k \notin \mathcal{F}_k$. The filtration $\mathcal{F}_k$ corresponds to the information carried until iteration k, to be used on the computation of iteration $k + 1$. The filtration $\widehat{\mathcal{F}}_k$ corresponds to the information carried until iteration k plus the information produced at the first projection step of iteration $k+1$, namely, $\widehat{\mathcal{F}}_k = \sigma(\mathcal{F}_k \cup \sigma(\xi^k))$. The way information evolves according to filtrations $\{\mathcal{F}_k, \widehat{\mathcal{F}}_k\}$ is natural in applications. Also, the use of two filtrations will be important since even though $z^k \notin \mathcal{F}_k$ we have $z^k \in \widehat{\mathcal{F}}_k$, so that, given $i \in [m]$,

$$
\begin{aligned}
\mathbb{E}[\epsilon_{2,i}^k | \widehat{\mathcal{F}}_k] &= \mathbb{E}\left[\frac{1}{N_{k,i}} \sum_{j=1}^{N_{k,i}} F_i(\eta_{j,i}^k, z^k) - T_i(z^k) \middle| \widehat{\mathcal{F}}_k \right] \\
&= \frac{1}{N_{k,i}} \sum_{j=1}^{N_{k,i}} \mathbb{E}\left[F_i(\eta_{j,i}^k, z^k) | \widehat{\mathcal{F}}_k \right] - T_i(z^k) \\
&= \frac{1}{N_{k,i}} \sum_{j=1}^{N_{k,i}} T_i(z^k) - T_i(z^k) = 0
\end{aligned}
$$

(17)

if for every $i \in [m]$, $\{\eta_{j,i}^k : j \in [N_{k,i}]\}$ is independent of $\widehat{\mathcal{F}}_k$ and identically distributed as ξ. We exploit (17) for avoiding first order moments of the stochastic errors, which drastically diminishes the complexity by an order of one, and for using martingale techniques.[2] We remark that, with some minor extra effort, the same samples can be used in both projections in method (11)–(12). Next we describe the assumptions required in our convergence analysis.

Assumption 3.1 (consistency). The solution set $X^* := \mathrm{S}(T, X)$ is nonempty.

Assumption 3.2 (stochastic model). $X \subset \mathbb{R}^n$ is closed and convex, $(\Xi, \mathcal{G})$ is a measurable space such that $F : \Xi \times X \to \mathbb{R}^n$ is a Carathéodory map,[3] $\xi : \Omega \to \Xi$ is a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $\mathbb{E}[\|F(\xi, x)\|] < \infty$ for all $x \in X$.

Assumption 3.3 (Lipschitz continuity). The mean operator $T : X \to \mathbb{R}^n$ defined by (2) is Lipschitz continuous with modulus $L > 0$.

Assumption 3.4 (pseudomonotonicity). The mean operator $T : \mathbb{R}^n \to \mathbb{R}^n$ is pseudomonotone,[4] i.e., $\langle T(x), z - x \rangle \geq 0 \implies \langle T(z), z - x \rangle \geq 0$ for all $z, x \in \mathbb{R}^n$.

Assumption 3.5 (sampling rate). Given $\{N_k\}$ as in Algorithm 2, define $N_{k,\min} := \min_{i \in [m]} N_{k,i}$ and $\frac{1}{\mathcal{N}_k} := \sum_{i=1}^m \frac{1}{N_{k,i}}$. Then one of the conditions is satisfied:
  (i) $\sum_{k=0}^\infty \frac{1}{\mathcal{N}_k} < \infty$,
  (ii) $\sum_{k=0}^\infty \frac{1}{N_{k,\min}} < \infty$.

Typically a sufficient choice is, for $i \in [m]$,

$$
N_{k,i} = \Theta_i \cdot (k + \mu_i)^{1+a_i} \cdot \left(\ln\left(k + \mu_i \right) \right)^{1+b_i}
$$

---

[2]If also $\{\xi_{j,i}^k : j \in [N_{k,i}]\}$ is independent of $\mathcal{F}_k$ and identically distributed as ξ, then, for $i \in [m]$, $\mathbb{V}[\epsilon_{1,i}^k] = N_{k,i}^{-1} \mathbb{V}[F_i(\xi, x^k)]$ and $\mathbb{V}[\epsilon_{2,i}^k] = N_{k,i}^{-1} \mathbb{V}[F_i(\xi, z^k)]$, so that our method iteratively reduces the variance of the oracle error as long as $\{N_{k,i}\}_{k \in \mathbb{N}}$ increases.

[3]That is, $F(\xi, \cdot) : X \to \mathbb{R}^n$ is continuous for a.e. $\xi \in \Xi$ and $F(\cdot, x) : \Xi \to \mathbb{R}^n$ is measurable.

[4]Pseudomonotonicity is a weaker assumption than monotonicity, i.e., $\langle T(z) - T(x), z - x \rangle \geq 0$ for all $x, z \in \mathbb{R}^n$.

for any $\Theta_i > 0$, $\mu_i > 0$ with $a_i > 0$, $b_i \geq -1$ or $a_i = 0$, $b_i > 0$ (the latter is the minimum requirement). It is essential to specify choices of the above parameters that induce a practical complexity of method (11)–(12), i.e., practical upper bounds on the total oracle complexity $\sum_{k=1}^{K} \sum_{i=1}^{m} 2N_{k,i}$, where K is an estimate of the total number of iterations needed for achieving a given specified tolerance $\epsilon > 0$. A convergence rate in terms of K is also desirable. As commented after Theorem 3.22, our algorithm achieves an optimal accelerated rate $\mathcal{O}(1/K)$ and an optimal complexity $\mathcal{O}(\epsilon^{-2})$ up to a first order logarithmic term $\ln(\epsilon^{-1})$.[5]

We offer two options of sampling coordination among the agents.

Assumption 3.6 (sampling coordination). For each $i \in [m]$ and $k \in \mathbb{N}_0$, $\{\xi_i^k\}$ and $\{\eta_i^k\}$ are i.i.d. samples of ξ such that $\{\xi_i^k\}$ and $\{\eta_i^k\}$ are independent of each other. Also, one of the two next coordination conditions is satisfied:
  (i) (Centralized sampling) For all $i \in [m]$, $N_{k,i} \equiv N_k$, $\xi_i^k \equiv \xi^k$, and $\eta_i^k \equiv \eta^k$.
  (ii) (Distributed sampling) $\{\xi^k, \eta^k : k \in \mathbb{N}\}$ is an i.i.d. sample of ξ.

We remark that, with some extra effort, it is possible to use the same samples in each projection step of the method (11)–(12), that is, $\xi_i^k \equiv \eta_i^k$ for $k \in \mathbb{N}_0$ and $i \in [m]$. We ask for independence in Assumption 3.6 in order to simplify the analysis. Both conditions (i) and (ii) in Assumption 3.6 are the same for $m = 1$.[6] Assumption 3.6 implies in particular that $\{\xi^k\}$ is independent of $\mathcal{F}_k$, $\{\eta^k\}$ is independent of $\widehat{\mathcal{F}}_k$, and both are identically distributed as ξ. In particular, for any $x \in \mathbb{R}^n$, $k \in \mathbb{N}$, $i \in [m]$, $j \in [N_{k,i}]$,

$$\mathbb{E}\left[F_i(\xi_{j,i}^k, x)\Big|\mathcal{F}_k\right] = \mathbb{E}\left[F_i(\eta_{j,i}^k, x)\Big|\widehat{\mathcal{F}}_k\right] = T_i(x).$$

Assumption 3.7 (stepsize bounded away from zero). The stepsize sequence $\{\alpha_k\}$ in Algorithm 2 satisfies

$$0 < \inf_{k \in \mathbb{N}} \alpha_k \leq \hat{\alpha} := \sup_{k \in \mathbb{N}} \alpha_k < \frac{1}{\sqrt{6}L}.$$

The following two sets of assumptions ensure that the variance of the error $F(\xi, x) - T(x)$ is controlled, so that (together with Assumption 3.5 on the sampling rate) boundedness is guaranteed, even in the case of an unbounded operator.

Assumption 3.8 (variance control). There exists $p \geq 2$ such that one of the following three conditions holds:
  (i) There exist $x^* \in X^*$ and $\sigma(x^*) > 0$ such that for all $x \in X$,

$$\left|\|F(\xi, x) - T(x)\|\right|_p \leq \sigma(x^*)\left(1 + \|x - x^*\|\right).$$

  (ii) There exists a locally bounded and measurable function $\sigma : X^* \to \mathbb{R}_+$ such that for all $x^* \in X^*$, $x \in X$, the inequality in (i) is satisfied.

---

[5]In large-scale problems such as in machine learning, the dependence of the rate and complexity estimates on the dimension is relevant in the case of large constraint dimension ($n_i \gg 1$) or large networks ($m \gg 1$). We show that our method has a rate of $\mathcal{O}(\sigma^4)$ which is independent of dimension, where σ^2 is the variance, even in the case of an unbounded feasible set and a nonuniform variance. Sharper constants are available in the case of uniform variance (see Proposition 3.26).

[6]When $m > 1$, item (i) corresponds to the case where one stochastic oracle is centralized. In this case, fewer samples are required but the sampling process needs total coordination. Item (ii) corresponds to the other extreme case, where the agents have completely distributed oracles so that the sampling process of each agent is conducted independently. We do not explore the intermediate possibilities between (i) and (ii).

(iii) There exists a positive sequence $\{\sigma_{l,i} : i \in [m], l \in [n_i]\}$ such that for all $i \in [m]$, $l \in [n_i]$, $x \in X$, $|F_{\ell,i}(\xi, x) - T_{\ell,i}(x)|_p \leq \sigma_{\ell,i}$, where $F_{\ell,i}$ and $T_{\ell,i}$ are the components of F_i and T_i respectively.

In item (iii) we define $\sigma^2 := \sum_{i=1}^m \sum_{\ell=1}^{n_i} \sigma_{\ell,i}^2$. Note that when $p = 2$, $\sigma(x^*)^2 (1 + \|x - x^*\|)^2$ in the case of (i)–(ii) and σ^2 in the case of item (iii), are, respectively, upper bounds on the variance of the components of $F(\xi, x)$. Items (i) and (ii) are essentially the same, excepting that (i) only requires the condition to hold at just one point $x^* \in X^*$ rather than on the entire solution set. Item (i) is sufficient for the analysis, but (ii) allows for sharper estimates in the case of unbounded feasible set and operator. Item (iii) allows for even sharper ones. In what follows we shall denote $q := p/2$.

For the important case in which the random operator F is Lipschitz, both items (i)–(ii) are satisfied with a continuous $\sigma : X^* \to \mathbb{R}_+$. Namely, if for any $x, y \in \mathbb{R}^n$,

$$(18) \qquad \|F(\xi, x) - F(\xi, y)\| \leq \mathsf{L}(\xi)\|x - y\|,$$

for some measurable $\mathsf{L} : \Xi \to \mathbb{R}_+$ with finite L_p-norm for some $p \geq 2$, then Assumptions 3.2–3.3 and 3.8 hold with $L := \mathbb{E}[\mathsf{L}(\xi)]$ and

$$(19) \qquad \sigma(x^*) := \max\{|\|F(\xi, x^*) - T(x^*)\||_p, |\mathsf{L}(\xi)|_p + L\},$$

for $x^* \in X^*$. Indeed, Assumption 3.3 with $L := \mathbb{E}[\mathsf{L}(\xi)]$ follows from Jensen's inequality and (18). For establishing (19), note that by Minkowski's inequality

$$\begin{aligned}
\||F(\xi, x) - T(x)\||_p &\leq \||F(\xi, x) - F(\xi, x^*)\||_p + \||F(\xi, x^*) - T(x^*)\||_p \\
&\quad + \|T(x) - T(x^*)\| \\
&\leq (|\mathsf{L}(\xi)|_p + L)\|x - x^*\| + \||F(\xi, x^*) - T(x^*)\||_p \\
&\leq \sigma(x^*)(\|x - x^*\| + 1),
\end{aligned}$$

using (18) and the fact that T is L-Lipschitz continuous in the second inequality and (19) in the third inequality. Thus, Assumption 3.8(i)–(ii) is merely a *finite* variance assumption even for the case of an unbounded feasible set. Assumption 3.8(iii) means that the variance is *uniformly bounded* over the feasible set X. It has been assumed in most of the past literature [19, 26, 43, 7, 44, 21, 22, 45] on SA algorithms for SVI and stochastic programming.[7] Assumptions 3.8(i)–(ii) are much weaker than Assumption 3.8(iii) and, to the best of our knowledge, seem to be new for monotone operators or convex functions without regularization.

The next examples provide instances where Assumption 3.8(i)–(ii) and the iterative variance reduction in method (11)–(12) are relevant or even *necessary* for the asymptotic convergence of the generated sequence, in the case of an unbounded feasible set (e.g., stochastic equations and stochastic complementarity problems).

Example 3.9 (linear SVI with unbounded feasible set). The following example is a typical situation of a nonuniform variance over a unbounded feasible set. It includes the cases of stochastic linear equations and complementarity problem. Let the random operator be

$$F(\xi, x) = A(\xi)x,$$

---

[7]Assumption 3.8(iii) has been weakened in previous works only in situations in which the operator satisfies more demanding monotonicity conditions (strongly monotone operator in [39] and weak-sharp monotone operator in [17]) or when the operator is merely monotone, but with additional Tykhonov regularization (as in [17], without convergence rate results).

for all $x \in \mathbb{R}^n$, where $A(\xi)$ is a random matrix whose entries have finite mean and variance, such that $\bar{A} := \mathbb{E}[A(\xi)]$ is nonnull and positive semidefinite. In this case, $T(x) = \bar{A}x$ $(x \in \mathbb{R}^n)$ is monotone and linear. For all $x \in \mathbb{R}^n$, $\mathbb{V}[F(\xi, x)] = x^t Bx$, where $B := \sum_{i=1}^m \text{cov}[A_i(\xi)]$ is positive semidefinite and $A_1(\xi), \ldots, A_m(\xi)$ are the rows of $A(\xi)$. We denote by $N(B)$ the kernel of B and by $N(B)^\perp$ its orthogonal complement. Given $x \in \mathbb{R}^n$, let x_B be the orthogonal projection of x onto $N(B)^\perp$. Then for all $x \in \mathbb{R}^n$ we have

$$\mathbb{V}[F(\xi, x)] \geq \lambda_+(B) \|x_B\|^2,$$

where $\lambda_+(B)$ is the smallest nonnull eigenvalue of B. In particular, if B is positive definite, then for all $x \in \mathbb{R}^n$, $\mathbb{V}[F(\xi, x)] \geq \lambda_{\min}(B) \|x\|^2$, where $\lambda_{\min}(B)$ is the smallest eigenvalue of B. This shows that Assumption 3.8(iii) does not hold if X is unbounded (in fact, the variance grows quadratically in the infinite horizon).

Example 3.10 (equation problem for zero mean random constant operator). The following example presents a simple situation where, in the case of an unbounded feasible set and an oracle with *uniform variance*, the method in [7] may possess an undesirable property: for the null operator $T \equiv 0$, the method generates a sequence whose final ergodic average converges but the sequence itself a.s. diverges to ∞.

The method in [7] says that given a prescribed number of iterations K, for $k \in [K]$ compute

$$z^k = \Pi \left[x^k - \alpha_k^K F(\xi_k^K, x^k) \right],$$
$$x^{k+1} = \Pi \left[x^k - \alpha_k^K F(\eta_k^K, z^k) \right]$$

and give as final output the ergodic average $\bar{z}^K = \sum_{k=1}^K p_k^K z^k$, where $\{p_k^K\}$ is a positive sequence of weights such that $\sum_{k=1} p_k^K = 1$ and $\{\alpha_k^K\}$ is a sequence of positive stepsizes. For an unbounded X, assuming uniformly bounded variance (Assumption 3.8(iii)) and a single oracle call per iteration, it is shown that there exists $\{v^K\} \subset \mathbb{R}^n$ such that $\mathbb{E}[\tilde{G}(\bar{z}^K, v^K)] \lesssim 1/\sqrt{K}$ and $\mathbb{E}[\|v^K\|] \lesssim \sqrt{K}$ (see [7, Corollary 3.4]). In these statements, for $z, v \in \mathbb{R}^n$, $\tilde{G}(z, v) = \sup_{y \in X} \langle T(y) - v, z - y \rangle$, as mentioned in subsection 1.1. The following example shows that $\limsup_{K \to \infty} \|z^K\| = \infty$ with total probability. We shall consider $n = 1$, but one can easily generalize the argument for any $n > 1$. Consider $X = \mathbb{R}$ and the random operator given by

$$F(\xi, x) = \xi,$$

for all $x \in \mathbb{R}$, where ξ is a random variable with zero mean, finite variance σ^2, and finite third moment (one could generalize the argument assuming finite q-moment for any $q > 2$). In this case, trivially $T \equiv 0$, $X^* = \mathbb{R}$ and Assumption 3.8(iii) holds. It is easy to check that the mirror-prox method in [7] gives, after K iterations, for $k \in [K]$:

$$(20) \qquad z^k = x^1 - \sum_{i=1}^k \alpha_i^K \xi_i^K, \quad \bar{z}^K = \sum_{k=1}^K p_k^K z^k,$$

where $p_k^K = c_0 \Gamma_K \alpha_k^K$, $\gamma_k (\Gamma_k \alpha_k^K)^{-1} \equiv c_0$ is a constant, $\gamma_k := 2(1 + k)^{-1}$, $\{\Gamma_k\}$ is defined recursively as $\Gamma_1 := 1$, $\Gamma_k := (1 - \gamma_k)\Gamma_{k-1}$, and the stepsize is

$$\alpha_k^K := \frac{k}{3LK + \sigma K \sqrt{K - 1}},$$

(see [7, Corollary 3.4]). Using the expression of $\{p_k^K\}$ and $\sum_{k=1}^K p_k^K = 1$, we get

$$(21) \qquad \bar{z}^K = x^1 - \sum_{k=1}^K \theta_k^K \cdot \xi_k^K,$$

where $\theta_k^K := c_0 \Gamma_K \alpha_k^K \sum_{i=k}^K \alpha_i^K$. Note that $\Gamma_k = \frac{2}{k(k+1)}$ and

$$\theta_k^K = \frac{c_0 \Gamma_K k}{\left(3LK + \sigma K \sqrt{K-1}\right)^2} \sum_{i=k}^K i = \frac{c_0 k(K-k+1)(K+k)}{K(K+1)\left(3LK + \sigma K \sqrt{K-1}\right)^2}.$$

We have the following estimates:

$$\bar{s}_K^2 := \sum_{k=1}^K \left(\theta_k^K\right)^2 \sim K^{-3}, \quad s_K^2 := \sum_{k=1}^K \left(\alpha_k^K\right)^2 \sim 1, \quad \sum_{k=1}^K \left(\alpha_k^K\right)^3 \sim K^{-\frac{1}{2}} \quad (\text{as } K \to \infty).$$

We will now invoke Lyapounov's criteria [2, Theorem 7.3] with $\delta = 1$ for the sum $\sum_{k=1}^K \alpha_k^K \cdot \xi_k^K$ of independent random variables, obtaining

$$\lim_{K \to \infty} \frac{\mathbb{E}\left[|\xi|^3\right]}{s_K^3} \sum_{k=1}^K \left(\alpha_k^K\right)^3 = \lim_{K \to \infty} \mathbb{E}\left[|\xi|^3\right] K^{-\frac{1}{2}} = 0.$$

Hence $(\sigma s_K)^{-1} \sum_{k=1}^K \alpha_k^K \xi_k^K$ converges in distribution to $N(0,1)$. Therefore, there exists some constant $C > 0$ such that for any $R > 0$,

$$\mathbb{P}\left(\limsup_{K \to \infty} z^K \geq R\right) = \mathbb{P}\left(\limsup_{K \to \infty} \sum_{k=1}^K \frac{\alpha_k^K}{\sigma s_K} \cdot \xi_k^K \geq CR\right)$$

$$(22) \qquad \geq \limsup_{K \to \infty} \mathbb{P}\left(\sum_{k=1}^K \frac{\alpha_k^K}{\sigma s_K} \cdot \xi_k^K \geq CR\right) > 0,$$

using (20) and $s_K \sim 1$ in the equality and Portmanteau's theorem [10, Theorem 3.2.5] in the inequality. For every $R > 0$, the event $A_R := [\limsup_{K \to \infty} z^K \geq R]$ is a tail event with positive probability and, hence, has total probability, invoking Kolmogorov's zero-one law [10, Theorem 2.5.1]. We conclude from (22) that

$$\mathbb{P}\left(\limsup_{K \to \infty} z^K = \infty\right) = \lim_{R \to \infty} \mathbb{P}\left(A_R\right) = 1.$$

This shows that $\{z^K\}$ diverges with total probability. From the 1-series theorem [10, Theorem 2.5.3], $\sum_{K=1}^\infty \bar{s}_K^2 < \infty$, and (21), we have that a.s. $\{\bar{z}^K\}$ converges.

3.3. Convergence analysis. For any $x = (x_i)_{i=1}^m \in \mathbb{R}^n$ and $\alpha > 0$, we denote the (quadratic) natural residual function by

$$r_\alpha(x)^2 := \|x - \Pi[x - \alpha T(x)]\|^2 = \sum_{i=1}^m \|x_i - \Pi_i[x_i - \alpha T_i(x)]\|^2.$$

We start with two key lemmas whose proofs are given in the appendix. Define

$$(23) \qquad \rho_k := 1 - 6L^2\alpha_k^2,$$

for any $k \in \mathbb{N}_0$. We define recursively, for $k \in \mathbb{N}_0$, $A_0 := 0$,

$$(24) \qquad A_{k+1} := A_k + (8 + \rho_k)\alpha_k^2 \|\epsilon_1^k\|^2 + 8\alpha_k^2 \|\epsilon_2^k\|^2,$$

and, for $x^* \in X^*$, $M_0(x^*) := 0$,

$$(25) \qquad M_{k+1}(x^*) := M_k(x^*) + 2\langle x^* - z^k, \alpha_k \epsilon_2^k \rangle.$$

LEMMA 3.11 (recursive relation). *Suppose that Assumptions* 3.1, 3.3, *and* 3.7 *hold. Then, a.s. for all* $k \in \mathbb{N}$ *and* $x^* \in X^*$,

$$\|x^{k+1} - x^*\|^2 \le \|x^k - x^*\|^2 - \frac{\rho_k}{2} r_{\alpha_k}(x^k)^2 + M_{k+1}(x^*) - M_k(x^*) + A_{k+1} - A_k.$$

LEMMA 3.12 (error decay). *Consider Assumptions* 3.2–3.8. *For each* $i \in [m]$ *and* $N_i \in \mathbb{N}$, *let* $\xi_i := \{\xi_{j,i} : j \in [N_i]\}$ *be an i.i.d. sample of* ξ *with* $1/\mathcal{N} := \sum_{i=1}^m 1/N_i$ *and* $N_{\min} := \min_{i \in [m]} N_i$. *For any* $x \in X$ *set*

$$\epsilon_i(x) := \sum_{j=1}^{N_i} \frac{F_i(\xi_{j,i}, x) - T_i(x)}{N_i}, \quad \epsilon(x) := (\epsilon_1(x), \ldots, \epsilon_m(x)).$$

If Assumption 3.8(i) *hold for some* $x^* \in X^*$, *then for all* $x \in X, v \in \mathbb{R}^n$,

$$\big| \|\epsilon(x)\| \big|_p \le \sqrt{\frac{\mathsf{A}}{\mathcal{N}}} C_p f(x, x^*), \quad |\langle v, \epsilon(x) \rangle|_p \le \|v\| \sqrt{\frac{\mathsf{B}}{\mathcal{N}}} C_p f(x, x^*),$$

where $f(x, x^*) := \sigma(x^*)(1 + \|x - x^*\|)$ *and*

1. $\mathsf{A} = 1$ *if* $m = 1$ *and* $\mathsf{A} = 2$ *if* $m > 1$,
2. $\mathsf{B} = 2$ *if* $m > 1$ *and* $\{\xi_{j,i} : 1 \le i \le m, 1 \le j \le N_i\}$ *is i.i.d.*,
3. $\mathsf{B} = 1$ *if* $m = 1$ *or if* $m > 1$ *with* $N_i \equiv N$, $\xi_{j,i} \equiv \xi_j$ *for all* $i \in [m]$.

Moreover, if Assumption 3.8(iii) *holds, then for all* $x \in X, v \in \mathbb{R}^n$,

$$\big| \|\epsilon(x)\| \big|_p \le \frac{C_p \sigma}{\sqrt{N_{\min}}}, \quad |\langle v, \epsilon(x) \rangle|_p \le \|v\| \frac{C_p \sigma}{\sqrt{N_{\min}}}.$$

In the remainder of the paper, we take A and B as given in Lemma 3.12 with $\{\mathcal{N}_k\}$, $\{N_{k,\min}\}$, $\sigma(\cdot)$, and σ as given in Assumptions 3.5, 3.6, and 3.8.

The following two results will establish upper bounds on $A_{k+1} - A_k$ and $M_{k+1}(x^*) - M_k(x^*)$ in terms of $\|x^k - x^*\|^2$ for any $x^* \in X^*$. Under the Assumptions 3.8(i)–(ii) of nonuniform variance, we need first a bound of $\|x^* - z^k\|^2$ in terms of $\|x^k - x^*\|^2$. We define

$$(26) \qquad \mathsf{G}_{k,p}(x^*) := \alpha_k C_p \sigma(x^*),$$

$$(27) \qquad \mathsf{H}_{k,p}(x^*) := \mathsf{G}_{k,p}(x^*) \sqrt{\frac{\mathsf{A}}{\mathcal{N}_k}},$$

PROPOSITION 3.13. *Consider Assumptions* 3.1–3.8. *If Assumption* 3.8(i) *holds for some* $x^* \in X^*$, *then*

$$\big| \|z^k - x^*\| \big| \mathcal{F}_k \big|_p \le [1 + L\alpha_k + \mathsf{H}_{k,p}(x^*)] \|x^k - x^*\| + \mathsf{H}_{k,p}(x^*).$$

Moreover, if Assumption 3.8(iii) *holds, then*

$$\big| \|z^k - x^*\| \big| \mathcal{F}_k \big|_p \le (1 + \mathcal{L}\alpha_k) \|x^k - x^*\| + \alpha_k \left(\mathcal{M} + \frac{C_p \sigma}{\sqrt{N_{k,\min}}} \right),$$

with $\mathcal{L} = L$ *and* $\mathcal{M} = 0$ *or, alternatively, if* $\sup_{x \in X} \|T(x)\| \le M < \infty$ *with* $\mathcal{L} = 0$ *and* $\mathcal{M} = 2M$.

Proof. Recall that $z^k = \Pi[x^k - \alpha_k(T(x^k) + \epsilon_1^k)]$. By Lemma 2.1(iv), we have $x^* = \Pi[x^* - \alpha_k T(x^*)]$.

Consider first Assumption 3.8(i). By Lemma 2.1(iii),

$$
\begin{aligned}
\|x^* - z^k\| &\leq \|x^* - x^k - \alpha_k(T(x^*) - T(x^k)) + \alpha_k \epsilon_1^k\| \\
&\leq \|x^* - x^k\| + \alpha_k\|T(x^k) - T(x^*)\| + \alpha_k \left\|\epsilon_1^k\right\| \\
&\leq (1 + L\alpha_k)\|x^* - x^k\| + \alpha_k \left\|\epsilon_1^k\right\|,
\end{aligned}
$$

(28)

using the Lipschitz continuity of T in the last inequality. Using $x^k \in \mathcal{F}_k$ and taking $|\cdot|\mathcal{F}_k|_p$ in (28) we get from Minkowski's inequality,

$$
(29) \qquad \left|\|z^k - x^*\|\big|\mathcal{F}_k\right|_p \leq (1 + L\alpha_k)\|x^* - x^k\| + \alpha_k \left|\left\|\epsilon_1^k\right\| \left|\mathcal{F}_k\right|_p\right..
$$

We now recall the definition of ϵ_1^k in (13). We have

$$
(30) \qquad \left|\left\|\epsilon_1^k\right\| \left|\mathcal{F}_k\right|_p\right. \leq \left(\frac{\mathsf{A}}{\mathcal{N}_k}\right)^{\frac{1}{2}} C_p \sigma(x^*)(1 + \|x^k - x^*\|),
$$

using Lemma 3.12, $x^k \in \mathcal{F}_k$, and the independence of ξ^k with $\mathcal{F}_k$. Relations (29)–(30) prove the required claim.

We now consider Assumption 3.8(iii). In this case, (28) may be replaced by

$$
(31) \qquad \|x^* - z^k\| \leq (1 + \mathcal{L}\alpha_k)\|x^* - x^k\| + \alpha_k\left(\mathcal{M} + \left\|\epsilon_1^k\right\|\right),
$$

with $\mathcal{L}$ and $\mathcal{M}$ as stated in the proposition. Using $x^k \in \mathcal{F}_k$ and taking $|\cdot|\mathcal{F}_k|_p$ in (31) we get from Minkowski's inequality

$$
(32) \qquad \left|\|z^k - x^*\|\big|\mathcal{F}_k\right|_p \leq (1 + \mathcal{L}\alpha_k)\|x^* - x^k\| + \alpha_k\left(\mathcal{M} + \left|\left\|\epsilon_1^k\right\| \left|\mathcal{F}_k\right|_p\right.\right).
$$

By Lemma 3.12, relation (30) is replaced by $\left|\|\epsilon_1^k\|\big|\mathcal{F}_k\right|_p \leq \frac{C_p \sigma}{\sqrt{N_{k,\min}}}$. The claim results from the previous inequality and (32). $\qquad \square$

The following proposition gives bounds on the increments of $\{A_k\}$ and $\{M_k(x^*)\}$ in terms of $\|x^k - x^*\|^2$. Recall definitions (23)–(25) and (26)–(27).

PROPOSITION 3.14 (bounds on increments). *Consider Assumptions* 3.1–3.8. *If Assumption* 3.8(i) *holds for some* $x^* \in X^*$, *then, for all* $k \in \mathbb{N}_0$,

$$
\begin{aligned}
|A_{k+1} - A_k|\mathcal{F}_k|_q &\leq \left[32\left(1 + L\alpha_k + \mathsf{H}_{k,p}(x^*)\right)^2 + 2(8 + \rho_k)\right]\mathsf{H}_{k,p}(x^*)^2\|x^k - x^*\|^2 \\
&\quad + \left[32\mathsf{H}_{k,p}(x^*)^2 + 16 + 2(8 + \rho_k)\right]\mathsf{H}_{k,p}(x^*)^2,
\end{aligned}
$$

$$
\begin{aligned}
|M_{k+1}(x^*) - M_k(x^*)|\mathcal{F}_k|_q &\leq \sqrt{\frac{\mathsf{B}}{\mathsf{A}}}\mathsf{H}_{k,p}(x^*)\left[1 + L\alpha_k + \mathsf{H}_{k,p}(x^*)\right]^2\|x^k - x^*\|^2 \\
&\quad + \sqrt{\frac{\mathsf{B}}{\mathsf{A}}}\mathsf{H}_{k,p}(x^*)\left[1 + L\alpha_k + (3 + 2L\alpha_k)\mathsf{H}_{k,p}(x^*) + 2\mathsf{H}_{k,p}(x^*)^2\right]\|x^k - x^*\| \\
&\quad + \sqrt{\frac{\mathsf{B}}{\mathsf{A}}}\mathsf{H}_{k,p}(x^*)\left[\mathsf{H}_{k,p}(x^*) + \mathsf{H}_{k,p}(x^*)^2\right].
\end{aligned}
$$

Moreover, if Assumption 3.8(iii) *holds, then, for all* $k \in \mathbb{N}_0$,

$$|A_{k+1} - A_k|\mathcal{F}_k|_q \leq (16 + \rho_k)\,\alpha_k^2 \frac{C_p^2 \sigma^2}{N_{k,\min}},$$

$$|M_{k+1}(x^*) - M_k(x^*)|\mathcal{F}_k|_q \leq (1 + \mathcal{L}\alpha_k)\,\alpha_k \frac{C_p \sigma}{\sqrt{N_{k,\min}}}\|x^k - x^*\|$$

$$+ \left(\mathcal{M} + \frac{C_p \sigma}{\sqrt{N_{k,\min}}}\right)\alpha_k^2 \frac{C_p \sigma}{\sqrt{N_{k,\min}}}.$$

Proof. Assume first that Assumption 3.8(i) holds. We start with the bound on $A_{k+1} - A_k$. Definition (13), Lemma 3.12, $x^k \in \mathcal{F}_k$, the independence of $\{\xi^k\}$ and $\mathcal{F}_k$, and the fact that $(a + b)^2 \leq 2a^2 + 2b^2$ imply

$$(33) \qquad \left|\|\epsilon_1^k\|^2 \,\Big|\mathcal{F}_k\right|_q = \left|\|\epsilon_1^k\| \,\Big|\mathcal{F}_k\right|_p^2 \leq 2\frac{\mathsf{A}}{\mathcal{N}_k} C_p^2 \sigma(x^*)^2 \left(1 + \|x^k - x^*\|^2\right).$$

We proceed similarly for a bound of ϵ_2^k as defined in (14), but with the use of the filtration $\widehat{\mathcal{F}}_k$. Lemma 3.12, $z^k \in \widehat{\mathcal{F}}_k$, and the independence of $\{\eta^k\}$ and $\widehat{\mathcal{F}}_k$ imply

$$(34) \qquad \left|\|\epsilon_2^k\| \,\Big|\widehat{\mathcal{F}}_k\right|_p \leq \left(\frac{\mathsf{A}}{\mathcal{N}_k}\right)^{\frac{1}{2}} C_p \sigma(x^*)\left(1 + \|z^k - x^*\|\right).$$

We condition (34) with $||\cdot|\widehat{\mathcal{F}}_k|_p|\mathcal{F}_k|_p = |\cdot|\mathcal{F}_k|_p$ and then take squares, getting

$$(35) \qquad \left|\|\epsilon_2^k\|^2 \,\Big|\mathcal{F}_k\right|_q = \left|\|\epsilon_2^k\| \,\Big|\mathcal{F}_k\right|_p^2 \leq 2\frac{\mathsf{A}}{\mathcal{N}_k} C_p^2 \sigma(x^*)^2 \left(1 + \left|\|z^k - x^*\| \,\Big|\mathcal{F}_k\right|_p^2\right).$$

Finally we use (33), (35), (24), Proposition 3.13, and relation $(a + b)^2 \leq 2a^2 + 2b^2$, obtaining the required bounds on $A_{k+1} - A_k$.

Now we deal with $M_{k+1}(x^*) - M_k(x^*)$. Definition (14), Lemma 3.12, $z^k \in \widehat{\mathcal{F}}_k$, and the independence of $\{\eta^k\}$ and $\widehat{\mathcal{F}}_k$ imply

$$(36) \qquad \left|\langle x^* - z^k, \alpha_k \epsilon_2^k\rangle \,\Big|\widehat{\mathcal{F}}_k\right|_p \leq \|z^k - x^*\|\alpha_k \sqrt{\frac{\mathsf{B}}{\mathcal{N}_k}} C_p \sigma(x^*)(1 + \|z^k - x^*\|).$$

In (36), we first use $|\cdot|\widehat{\mathcal{F}}_k|_q \leq |\cdot|\widehat{\mathcal{F}}_k|_p$ and then take $|\cdot|\mathcal{F}_k|_q$, obtaining

$$\left|\langle x^* - z^k, \alpha_k \epsilon_2^k\rangle \,\Big|\mathcal{F}_k\right|_q \leq \alpha_k \sqrt{\frac{\mathsf{B}}{\mathcal{N}_k}} C_p \sigma(x^*) \left(\left|\|x^* - z^k\| \,\Big|\mathcal{F}_k\right|_q + \left|\|x^* - z^k\|^2 \,\Big|\mathcal{F}_k\right|_q\right)$$

$$(37) \qquad \leq \alpha_k \sqrt{\frac{\mathsf{B}}{\mathcal{N}_k}} C_p \sigma(x^*) \left(\left|\|x^* - z^k\| \,\Big|\mathcal{F}_k\right|_p + \left|\|x^* - z^k\| \,\Big|\mathcal{F}_k\right|_p^2\right),$$

using the fact that $||\cdot|\widehat{\mathcal{F}}_k|_q|\mathcal{F}_k|_q = |\cdot|\mathcal{F}_k|_q$ in the first inequality and the fact that $|\cdot|\mathcal{F}_k|_q \leq |\cdot|\mathcal{F}_k|_p$ in the second inequality. Definition (25), (37), Proposition 3.13, and the fact that $(a + b)^2 \leq 2a^2 + 2b^2$ entail the required bound on $M_{k+1}(x^*) - M_k(x^*)$.

Suppose now that Assumption 3.8(iii) holds. First we prove the bound on $\{A_{k+1} - A_k\}$. The proof is similar to the previous case, but (33) and (35) are replaced respectively by

$$(38) \qquad \left|\|\epsilon_1^k\|^2 \,\Big|\mathcal{F}_k\right|_q \leq \frac{C_p^2 \sigma^2}{N_{k,\min}}, \quad \left|\|\epsilon_2^k\|^2 \,\Big|\mathcal{F}_k\right|_q \leq \frac{C_p^2 \sigma^2}{N_{k,\min}},$$

using Lemma 3.12. From (24) and (38) we obtain the required bound on $A_{k+1} - A_k$. We deal now with $\{M_{k+1}(x^*) - M_k(x^*)\}$. The proof is similar to the previous case, but instead of (36) now we have

$$
(39) \qquad \left|\langle x^* - z^k, \alpha_k \epsilon_2^k \rangle \Big| \widehat{\mathcal{F}}_k \right|_p \leq \|z^k - x^*\| \alpha_k \frac{C_p \sigma}{\sqrt{N_{k,\min}}},
$$

using Lemma 3.12. In (39), we use $|\cdot| \widehat{\mathcal{F}}_k|_p \leq |\cdot| \widehat{\mathcal{F}}_k|_p$ and then take $|\cdot| \mathcal{F}_k|_q$, getting

$$
(40) \qquad \left|\langle x^* - z^k, \alpha_k \epsilon_2^k \rangle \Big| \mathcal{F}_k \right|_q \leq \alpha_k \frac{C_p \sigma}{\sqrt{N_{k,\min}}} \Big| \|z^k - x^*\| \Big| \mathcal{F}_k \Big|_q,
$$

using the fact that $\|\cdot| \widehat{\mathcal{F}}_k|_q | \mathcal{F}_k|_q = |\cdot| \mathcal{F}_k|_q$. The definition (25), Proposition 3.13 with $\| \cdot |\widehat{\mathcal{F}}_k|_q | \mathcal{F}_k|_q = | \cdot | \mathcal{F}_k|_q$, and (40) imply the required bound on $M_{k+1}(x^*) - M_k(x^*)$. □

Now, we combine Lemma 3.11 and Proposition 3.14 in the following recursive relation. Recall definitions (26)–(27).

PROPOSITION 3.15 (stochastic quasi-Fejér property). *Consider Assumptions* 3.1–3.8. *Then for all* $k \in \mathbb{N}_0$ *and for* $x^* \in X^*$ *as described in Assumption* 3.8, *it holds that*

$$
(41) \qquad \mathbb{E}\left[\|x^{k+1} - x^*\|^2 | \mathcal{F}_k\right] \leq \|x^k - x^*\|^2 - \frac{\rho_k}{2} r_{\alpha_k}(x^k) + \mathsf{C}_k(x^*) \frac{\mathcal{I}\|x^k - x^*\|^2 + 1}{\mathcal{N}_k'},
$$

where, under Assumption 3.8(i)–(ii), $\mathcal{I} = 1$, $\mathcal{N}_k' = \mathcal{N}_k$, *and*

$$
(42) \qquad \mathsf{C}_k(x^*) := \mathsf{AG}_{k,2}(x^*)^2 \left[32\left(1 + L\alpha_k + \mathsf{H}_{k,2}(x^*)\right)^2 + 18\right],
$$

while, under Assumption 3.8(iii), $\mathcal{I} = 0$, $\mathcal{N}_k' = N_{k,\min}$, *and*

$$
(43) \qquad \mathsf{C}_k(x^*) := \mathsf{C}_k = (16 + \rho_k)\alpha_k^2 C_2^2 \sigma^2.
$$

Proof. We first note that from definition (25), for any $x^* \in X^*$, $\{M_{k+1}(x^*) - M_k(x^*), \mathcal{F}_k\}$ defines a martingale difference sequence, that is, $\mathbb{E}[M_{k+1}(x^*) - M_k(x^*)|\mathcal{F}_k] = 0$ for all $k \in \mathbb{N}_0$. Indeed, $z^k \in \widehat{\mathcal{F}}_k$ and the independence between η^k and $\widehat{\mathcal{F}}_k$ imply that $\mathbb{E}[\epsilon_2^k | \widehat{\mathcal{F}}_k] = 0$. This equality and the fact that $z^k \in \widehat{\mathcal{F}}_k$ imply further that

$$
\mathbb{E}\left[M_{k+1}(x^*) - M_k(x^*) \Big| \widehat{\mathcal{F}}_k\right] = 2\left\langle x^* - z^k, \alpha_k \mathbb{E}\left[\epsilon_2^k \Big| \widehat{\mathcal{F}}_k\right]\right\rangle = 0.
$$

We take $\mathbb{E}[\cdot|\mathcal{F}_k]$ above and use the hereditary property $\mathbb{E}[\mathbb{E}[\cdot|\widehat{\mathcal{F}}_k]|\mathcal{F}_k] = \mathbb{E}[\cdot|\mathcal{F}_k]$ in order to get $\mathbb{E}[M_{k+1}(x^*) - M_k(x^*)|\mathcal{F}_k] = 0$ as claimed.

We now take the conditional expectation with respect to $\mathcal{F}_k$ in Lemma 3.11, obtaining

$$
(44) \qquad \mathbb{E}[\|x^{k+1} - x^*\|^2 | \mathcal{F}_k] \leq \|x^k - x^*\|^2 - \frac{\rho_k}{2} r_{\alpha_k}(x^k)^2 + \mathbb{E}[A_{k+1} - A_k | \mathcal{F}_k],
$$

using the facts that $x^k \in \widehat{\mathcal{F}}_k$ and that $\{M_{k+1}(x^*) - M_k(x^*), \mathcal{F}_k\}$ is a martingale difference sequence. We have that

$$
32\left(1 + L\alpha_k + \mathsf{H}_{k,2}(x^*)\right)^2 + 2(8 + \rho_k) > 32\mathsf{H}_{k,2}(x^*)^2 + 16 + 2(8 + \rho_k).
$$

Hence, under Assumption 3.8(i)–(ii), the bound of $\{A_{k+1} - A_k\}$ given in Proposition 3.14 implies that

$$(45) \qquad |A_{k+1} - A_k|\mathcal{F}_k|_q \le \mathsf{C}_k(x^*)\frac{\mathcal{I}\|x^k - x^*\|^2 + 1}{\mathcal{N}_k'},$$

for all $k \ge 0$, with $\mathcal{I} = 1$, $\mathcal{N}_k' = \mathcal{N}_k$, and the definition of $\mathsf{C}_k(x^*)$.

Under Assumption 3.8(iii), Proposition 3.14 implies (45), with $\mathcal{I} = 1$ and $\mathcal{N}_k' = N_{k,\min}$ and the definition of C_k. The claimed relation follows from (44) and (45) for $q = 1$. □

Remark 3.16. Under Assumption 3.8(i), the inequality of Proposition 3.15 holds for a given $x^* \in X^*$, as described in Assumption 3.8(i). Under Assumption 3.8(ii) or (iii), the inequality of Proposition 3.15 holds for every $x^* \in X^*$.

Remark 3.17 (bounds of $A_{k+1} - A_k$).   Recall the definition of $\mathsf{C}_k(x^*)$ in the previous proposition. Under Assumption 3.8(i)–(ii), the upper bound on

$$(46) \qquad \mathsf{C}(x^*) := \sup_k \mathsf{C}_k(x^*),$$

depends only on p, L, the sampling rate $\mathcal{N}_k$, $\sigma(x^*)^2$, and $\hat{\alpha}$ as defined in Assumption 3.7. From (26) and (42), under Assumption 3.8(i)–(ii) there exists a constant $c > 1$ such that

$$(47) \quad \frac{\mathsf{C}_k(x^*)}{\mathcal{N}_k} \le c\mathsf{H}_{k,2}(x^*)^2 \left(1 + \mathsf{H}_{k,2}(x^*)^2\right) \le c\hat{\alpha}^2 C_2^2 \frac{\mathsf{A}\sigma(x^*)^2}{\mathcal{N}_k}\left(1 + \hat{\alpha}^2 C_2^2 \frac{\mathsf{A}\sigma(x^*)^2}{\mathcal{N}_k}\right),$$

that is, $\mathsf{C}(x^*) \lesssim \sigma(x^*)^4$. But since at least $\mathcal{N}_k \ge N_{k,\min} l \approx \Theta k^{1+a}(\ln k)^{1+b}$, for some $\Theta > 0$, $a > 0$, $b \ge -1$ or $a = 0$, $b > 0$, the following nonasymptotic bound holds:

$$(48) \qquad \mathsf{C}_k(x^*) \lesssim \sigma(x^*)^2 \left(1 + \frac{\sigma(x^*)^2}{\Theta k^{1+a}(\ln k)^{1+b}}\right),$$

which is $\approx \sigma(x^*)^2$ for an iteration index k large enough as compared to $\sigma(x^*)^2$.[8] Under Assumption 3.8(iii), the following *uniform* bound holds on X^*: $\mathsf{C}_k \lesssim \sigma^2$.

We finish this section with the asymptotic convergence result.

THEOREM 3.18 (asymptotic convergence).   *Under Assumptions 3.1–3.8, a.s. the sequence $\{x^k\}$ generated by (11)–(12) is bounded, $\lim_{k\to\infty} \mathrm{d}(x^k, X^*) = 0$, and $r_{\alpha_k}(x^k)$ converges to 0 and in L^2. In particular, a.s. every cluster point of $\{x^k\}$ belongs to X^*.*

Proof. The result in Proposition 3.15 may be rewritten as

$$(49) \quad \mathbb{E}\left[\|x^{k+1} - x^*\|^2|\mathcal{F}_k\right] \le \left(1 + \frac{\mathcal{I}\mathsf{C}(x^*)}{\mathcal{N}_k'}\right)\|x^k - x^*\|^2 - \frac{\rho_k}{2}r_{\alpha_k}(x^k)^2 + \frac{\mathsf{C}(x^*)}{\mathcal{N}_k'},$$

for all $k \ge 0$ and for some $x^* \in X^*$, as ensured by Assumption 3.8. Taking into account Assumption 3.5, i.e., $\sum_k \mathcal{N}_k^{-1} < \infty$, (49), and the fact that $x^k \in \mathcal{F}_k$, we apply Theorem 2.2 with $y_k := \|x^k - x^*\|^2$, $a_k = \mathcal{I} \cdot \mathsf{C}(x^*)/\mathcal{N}_k'$, $b_k = \mathsf{C}(x^*)/\mathcal{N}_k'$, and $u_k := \rho_k r_{\alpha_k}(x^k)^2/2$, in order to conclude that a.s. $\{\|x^k - x^*\|^2\}$ converges

---

[8]In terms of numerical constants, a sharper bound can be obtained by exploiting the first order term $H_{k,2}(x^*) \sim \sigma(x^*)\mathcal{N}_k^{-1/2}$ in the definition of $\mathsf{C}_k(x^*)$. We do not carry out this procedure here.

and $\sum_k \rho_k r_{\alpha_k}(x^k)^2 < \infty$. In particular, $\{x^k\}$ is bounded, and $\hat{\rho} \sum_k r_{\alpha_k}(x^k)^2 \leq \sum_k \rho_k r_{\alpha_k}(x^k)^2 < \infty$, where $\hat{\rho} := 1 - 6\hat{\alpha}^2 L^2 > 0$ and $\hat{\alpha} := \sup_k \alpha_k$ by Assumption 3.7. Hence, a.s.,

$$(50) \qquad 0 = \lim_{k \to \infty} r_{\alpha_k}(x^k)^2 = \lim_{k \to \infty} \left\| x^k - \Pi\left[x^k - \alpha_k T(x^k) \right] \right\|^2.$$

The boundedness of the stepsize sequence, (50), and the continuity of T (Assumption 3.3) and Π (Lemma 2.1(iii)) imply that a.s. every cluster point $\bar{x}$ of $\{x^k\}$ satisfies

$$0 = \bar{x} - \Pi\left[\bar{x} - \bar{\alpha} T(\bar{x}) \right],$$

for some $\bar{\alpha} > 0$, in view of Assumption 3.7, i.e., the fact that the stepsizes are bounded away from zero; from Lemma 2.1(iv) we have that $\bar{x} \in X^*$. Almost surely, the boundedness of $\{x^k\}$ and the fact that every cluster point of $\{x^k\}$ belongs to X^* imply that $\lim_{k \to \infty} d(x^k, X^*) = 0$ as claimed.

We now prove convergence of $r_{\alpha_k}(x^k)$ to 0 in L^2. We take total expectation in (49) and obtain for all $k \geq 0$

$$(51) \quad \mathbb{E}\left[\|x^{k+1} - x^*\|^2 \right] \leq \left(1 + \frac{\mathcal{I}\mathsf{C}(x^*)}{\mathcal{N}'_k} \right) \mathbb{E}\left[\|x^k - x^*\|^2 \right] - \frac{\rho_k}{2} \mathbb{E}\left[r_{\alpha_k}(x^k)^2 \right] + \frac{\mathsf{C}(x^*)}{\mathcal{N}'_k}.$$

Taking into account Assumption 3.5, i.e., $\sum_k \mathcal{N}_k^{-1} < \infty$, (51), we apply Theorem 2.2 with $y_k := \mathbb{E}[\|x^k - x^*\|^2]$, $a_k = \mathcal{I} \cdot \mathsf{C}(x^*)/\mathcal{N}'_k$, $b_k = \mathsf{C}(x^*)/\mathcal{N}'_k$, and $u_k := \rho_k \mathbb{E}[r_{\alpha_k}(x^k)^2]/2$, in order to conclude that $\sum_k \rho_k \mathbb{E}[r_{\alpha_k}(x^k)^2] < \infty$. In particular, $\hat{\rho} \sum_k \mathbb{E}[r_{\alpha_k}(x^k)^2] \leq \sum_k \rho_k \mathbb{E}[r_{\alpha_k}(x^k)^2] < \infty$, which implies that $\lim_{k \to \infty} \mathbb{E}[r_{\alpha_k}(x^k)^2] = 0$ as claimed. □

3.4. Convergence rate and complexity analysis. We now study the convergence rate and the oracle complexity of our algorithm. Besides the relation in Proposition 3.15 for $p = 2$, we can also obtain a recursive relation for higher order moments, assuming that $p \geq 4$. This recursion, derived as a consequence of Propositions 3.14 and 3.19(i), will give an explicit upper bound on the p-norm of the generated sequence (see Proposition 3.20). The explicit bound of the 2-norm of the sequence will be used for giving explicit estimates on the convergence rate and complexity under Assumption 3.8(i)–(ii), i.e., when X and T are unbounded, in Theorem 3.22. In this setting, we will also obtain sharper estimates of the constants assuming uniform variance over the *solution set* (see Propositions 3.19(ii), Proposition 3.20(ii), and Theorem 3.25). Important cases satisfying these assumptions include the cases in which X^* is a singleton or a compact set (which can occur even when the feasible set X is unbounded).[9] Under the stronger Assumption 3.8(iii), that is, uniform variance over the feasible set, even sharper bounds on the estimates will be presented (see Propositions 3.19(iii) and 3.20(iii) and Theorem 3.25).

PROPOSITION 3.19 (improved stochastic quasi-Fejér properties).
(i) *If Assumption 3.8(i) holds for $p \geq 4$ and some $x^* \in X^*$, then for all k_0, k such that $0 \leq k_0 < k$, it holds that*

_____

[9] This occurs when the solution set is a singleton in the case of a strictly or strongly pseudomonotone operator. See Theorems 2.3.5 and 2.3.16 in [11] for general conditions of compactness of the solution set of a pseudomonotone VI. An example is the so-called strictly feasible complementarity problem over a cone.

$$\left|\|x^k - x^*\|^2\right|_q \le \left|\|x^{k_0} - x^*\|^2\right|_q$$
$$+ C_q \sqrt{\sum_{i=k_0+1}^{k} |M_i(x^*) - M_{i-1}(x^*)|_q^2} + \sum_{i=k_0+1}^{k} |A_i - A_{i-1}|_q.$$

(ii) *If Assumption 3.8(ii) holds for $p \ge 2$, then $\mathsf{C} := \sup_k \mathsf{C}_k : X^* \to \mathbb{R}_+$, as defined in (42) and (46), is a locally bounded measurable function, and for all $k \ge 0$,*

$$\mathbb{E}\left[d(x^{k+1}, X^*)^2 | \mathcal{F}_k\right] \le d(x^k, X^*)^2 - \frac{\rho_k}{2} r_{\alpha_k}(x^k)^2 + \mathsf{C}_k\left(\Pi_{X^*}(x^k)\right) \frac{d(x^k, X^*)^2 + 1}{\mathcal{N}_k}.$$

(iii) *If Assumption 3.8(iii) holds, then for all $k \ge 0$,*

(52) $$\mathbb{E}\left[d(x^{k+1}, X^*)^2 | \mathcal{F}_k\right] \le d(x^k, X^*)^2 - \frac{\rho_k}{2} r_{\alpha_k}(x^k)^2 + \frac{17 C_2^2 \hat{\alpha}^2 \sigma^2}{N_{k,\min}}.$$

Proof.
 (i) Define for simplicity $d_k := \|x^k - x^*\|^2$. Summing the relation in Lemma 3.11 from k_0 to $k - 1$ we obtain $0 \le d_k \le d_{k_0} + M_k(x^*) - M_{k_0}(x^*) + A_k - A_{k_0}$. This relation implies

(53) $$0 \le d_k \le d_{k_0} + [M_k(x^*) - M_{k_0}(x^*) + A_k - A_{k_0}]_+,$$

since $a \le b \Rightarrow [a]_+ \le [b]_+$ for any $a, b \in \mathbb{R}$. We take the q-norm in (53), getting

$$|d_k|_q \le |d_{k_0}|_q + |[M_k(x^*) - M_{k_0}(x^*) + A_k - A_{k_0}]_+|_q$$
$$\le |d_{k_0}|_q + |M_k(x^*) - M_{k_0}(x^*) + A_k - A_{k_0}|_q$$
(54) $$\le |d_{k_0}|_q + |M_k(x^*) - M_{k_0}(x^*)|_q + \sum_{i=k_0+1}^{k} |A_i - A_{i-1}|_q,$$

using Minkowski's inequality in the first and last inequalities and the fact that $|U_+|_q \le |U|_q$ for any random variable U in the second inequality.
Since $q \ge 2$ ($p \ge 4$), the norm of the martingale term above may be estimated via the Burkholder–Davis–Gundy (BDG) inequality (8) applied to the martingale $\tilde{M}_i := M_{k_0+i}(x^*) - M_{k_0}(x^*)$. This gives

(55) $$|M_k(x^*) - M_{k_0}(x^*)|_q \le C_q \sqrt{\sum_{i=k_0+1}^{k} |M_i(x^*) - M_{i-1}(x^*)|_q^2}.$$

Plugging (55) into (54) completes the proof of item (i).
(ii) Under Assumption 3.8(ii), we define $\bar{x}^k := \Pi_{X^*}(x^k)$, recalling Assumption 3.1, and obtain from Proposition 3.15

$$\mathbb{E}\left[d(x^{k+1}, X^*)^2 | \mathcal{F}_k\right] \le \mathbb{E}\left[\|x^{k+1} - \bar{x}^k\|^2 | \mathcal{F}_k\right]$$
$$\le \|x^k - \bar{x}^k\|^2 - \frac{\rho_k}{2} r_{\alpha_k}(x^k)^2 + \mathsf{C}_k(\bar{x}^k) \frac{\|x^k - \bar{x}^k\|^2 + 1}{\mathcal{N}_k}$$
$$= d(x^k, X^*)^2 - \frac{\rho_k}{2} r_{\alpha_k}(x^k)^2 + \mathsf{C}_k\left(\Pi_{X^*}(x^k)\right) \frac{d(x^k, X^*)^2 + 1}{\mathcal{N}_k},$$

using the fact that $\bar{x}^k \in X^*$ in the first inequality, the facts that $\mathsf{C}_k(\bar{x}^k) \in \mathcal{F}_k$ (which holds because $x^k \in \mathcal{F}_k$, Π_{X^*} is continuous and C_k is measurable), and $\bar{x}^k \in X^*$ (cf. Proposition 3.15) in the second inequality, and the fact that $\mathrm{d}(x^k, X^*) = \|x^k - \bar{x}^k\|$ in the equality. Note that the function $\mathsf{C} : X^* \to \mathbb{R}_+$ is measurable and locally bounded by Assumption 3.8(ii) and the definition of $\mathsf{C}_k(x^*)$.

(iii) We use a proof line analogous to the one in item (ii), with Assumption 3.8(iii) and Proposition 3.15. □

The following result gives explicit bounds on the p-norm of the sequence in the unbounded setting. In order to make the presentation easier, we introduce some definitions. Recall the constant c defined in Remark 3.17. We set

$$(56) \qquad \mathsf{D}_p(x^*) := 2c\hat{\alpha}^2 C_p^2 \sigma(x^*)^2,$$

with $\mathsf{D}(x^*) := \mathsf{D}_2(x^*)$. Define also $\mathsf{B}_2(x^*) := 0$ and for $p \geq 4$, set $\tilde{\mathsf{G}}_p(x^*) := C_p \hat{\alpha} \sigma(x^*)$ and

$$(57) \qquad \mathsf{B}_p(x^*) := \sqrt{3\mathsf{B}} C_q \tilde{\mathsf{G}}_p(x^*) \left[(1 + L\hat{\alpha})^2 + (3 + 2L\hat{\alpha})\sqrt{\mathsf{A}}\tilde{\mathsf{G}}_p(x^*) + 2\mathsf{A}\tilde{\mathsf{G}}_p(x^*)^2 \right].$$

PROPOSITION 3.20 (uniform boundedness in L^p).
(i) *Let Assumptions 3.1–3.8(i) hold for some $x^* \in X^*$ and $p \in \{2\} \cup [4, \infty)$. Choose $k_0 := k_0(x^*) \in \mathbb{N}$ and $\gamma := \gamma(x^*) > 0$ such that*

$$(58) \qquad \beta(x^*) := \mathsf{B}_p(x^*)\sqrt{\gamma} + \mathsf{D}_p(x^*)\gamma + \mathsf{D}_p(x^*)^2\gamma^2 < 1, \quad \sum_{k \geq k_0} \frac{1}{\mathcal{N}_k} < \gamma.$$

Then

$$\sup_{k \geq k_0} \left\| \|x^k - x^*\| \right\|_p^2 \leq \mathsf{c}_p(x^*) \left[1 + \left\| \|x^{k_0} - x^*\| \right\|_p^2 \right],$$

with $\mathsf{c}_2(x^*) = [1 - \beta(x^*)]^{-1}$ and $\mathsf{c}_p(x^*) = 4[1 - \beta(x^*)]^{-2}$ for $p \geq 4$.
(ii) *Let Assumptions 3.1–3.8(ii) hold and suppose there exists $\sigma > 0$ such that $\sigma(x^*) \leq \sigma$ for all $x^* \in X^*$. Let $\phi \in (0, \frac{\sqrt{5}-1}{2})$. Choose $k_0 \in \mathbb{N}$ such that $\sum_{k \geq k_0} \frac{1}{\mathcal{N}_k} \leq \frac{\phi}{2c\hat{\alpha}^2 C_2^2 \sigma^2}$. Then*

$$\sup_{k \geq k_0} \mathbb{E}\left[\mathrm{d}(x^k, X^*)^2 \right] \leq \frac{1 + \mathbb{E}\left[\mathrm{d}(x^{k_0}, X^*)^2 \right]}{1 - \phi - \phi^2}.$$

(iii) *If Assumptions 3.1–3.8(iii) hold, then*

$$\sup_{k \geq 0} \mathbb{E}\left[\mathrm{d}(x^k, X^*)^2 \right] \leq \mathrm{d}(x^0, X^*)^2 + \sum_{k=0}^{\infty} \frac{17 C_2^2 \hat{\alpha}^2 \sigma^2}{N_{k,\min}}.$$

Proof.
(i) Denote $d^k := \|x^k - x^*\|$. We first unify the Fejér-type relations obtained so far under Assumption 3.8(i)–(ii) as, for all $k > k_0$,

$$|d_k|_p^2 \leq |d_{k_0}|_p^2 + \mathsf{B}_p(x^*)\sqrt{\sum_{i=k_0}^{k-1} \frac{1 + |d_i|_p^2 + |d_i|_p^4}{\mathcal{N}_i}}$$

$$(59) \qquad + \mathsf{D}_p(x^*) \sum_{i=k_0}^{k-1} \frac{1 + |d_i|_p^2}{\mathcal{N}_i} + \mathsf{D}_p(x^*)^2 \sum_{i=k_0}^{k-1} \frac{1 + |d_i|_p^2}{\mathcal{N}_i^2}.$$

Indeed, for $p = 2$, we have $\mathsf{B}_2(x^*) = 0$ so that (59) results by summing the relation in Proposition 3.15 from k_0 to $k - 1$ and using the estimate in (47), the fact that $\mathsf{A} \leq 2$, $c > 1$, and the definition of $\mathsf{D}_2(x^*)$ as stated before the proposition. For $p \geq 4$, we recall the bounds of increments of $\{M_k(x^*)\}$ in Proposition 3.14. The common factor is bounded by $\sqrt{\mathsf{B}/\mathsf{A}} \cdot \mathsf{H}_{k,p}(x^*) \leq \sqrt{\mathsf{B}}\tilde{\mathsf{G}}_p(x^*)/\sqrt{\mathcal{N}_k}$. Using the definition of $\mathsf{H}_{k,p}(x^*)$ in (27), $\hat{\alpha}$ in Assumption 3.7, $\tilde{\mathsf{G}}_p(x^*)$, and $\mathcal{N}_k \geq 1$, it is easy to see that, in the bound of $M_{k+1}(x^*) - M_k(x^*)$ in Proposition 3.14, the sum of terms multiplying $\sqrt{\mathsf{B}/\mathsf{A}} \cdot \mathsf{H}_{k,p}(x^*)$ is at most $(1 + L\hat{\alpha})^2 + (3 + 2L\hat{\alpha})\sqrt{\mathsf{A}}\tilde{\mathsf{G}}_p + 2\mathsf{A}\tilde{\mathsf{G}}_p^2$. We use these bounds, the facts that $(|d_i|_p^2 + |d_i|_p + 1)^2 \leq 3(|d_i|_p^4 + |d_i|_p^2 + 1)$, and the definition (57) in order to obtain, for all $i \in \mathbb{N}_0$,

$$(60) \qquad |M_{i+1}(x^*) - M_i(x^*)|_q^2 \leq \mathsf{B}_p(x^*)^2 \frac{1 + |d_i|_p^2 + |d_i|_p^4}{\mathcal{N}_i}.$$

The proof of (59) for $p \geq 4$ follows from (45), (47) with $\mathsf{A} \leq 2$, $c > 1$ and the definition of $\mathsf{D}_p(x^*)$ as well as (60) and Proposition 3.19(i).

By Assumption 3.5, we can choose $k_0 \in \mathbb{N}_0$ and $\gamma > 0$ as in (58). In particular, $\sum_{i \geq k_0} \mathcal{N}_i^{-2} < \gamma^2$. Given an arbitrary $a > |d_{k_0}|_p$, define $\tau_a := \inf\{k > k_0 : |d_k|_p \geq a\}$. Suppose first that $\tau_a < \infty$ for all $a > |d_{k_0}|_p$. By (58), (59), and the definition of τ_a, we have

$$(61) \qquad a^2 \leq |d_{\tau_a}|_p^2 \leq |d_{k_0}|_p^2 + \mathsf{B}_p(x^*)\sqrt{\sum_{i=k_0}^{\tau_a-1} \frac{1 + a^2 + a^4}{\mathcal{N}_i}}$$

$$+ \mathsf{D}_p(x^*) \sum_{i=k_0}^{\tau_a-1} \frac{a^2 + 1}{\mathcal{N}_i} + \mathsf{D}_p(x^*)^2 \sum_{i=k_0}^{\tau_a-1} \frac{a^2 + 1}{\mathcal{N}_i^2}$$

$$\leq |d_{k_0}|_p^2 + \mathsf{B}_p(x^*)\sqrt{\gamma}\left(1 + a + a^2\right)$$

$$+ \mathsf{D}_p(x^*)\gamma\left(1 + a^2\right) + \mathsf{D}_p(x^*)^2\gamma^2\left(1 + a^2\right).$$

For $p = 2$, $\mathsf{B}_2(x^*) = 0$. By (61), taking $\beta := \beta(x^*) \in (0, 1)$ in (58), we get

$$(62) \qquad a^2 \leq \frac{|d_{k_0}|_p^2 + 1}{1 - \beta}.$$

For $p \geq 4$, by (61), taking $\beta := \beta(x^*)$ in (58), we obtain $\lambda a^2 \leq |d_{k_0}|_p^2 + a + 1$ with $\lambda := 1 - \beta$. It follows that

$$\left(a - \frac{1}{2\lambda}\right)^2 \leq \frac{4\lambda|d_{k_0}|_p^2 + 4\lambda + 1}{4\lambda^2} \implies a \leq \frac{2|d_{k_0}|_p + \sqrt{5} + 1}{2\lambda} \leq \frac{|d_{k_0}|_p + 2}{\lambda},$$

and finally

$$(63) \qquad a^2 \leq 4\frac{|d_{k_0}|_p^2 + 1}{(1 - \beta)^2}.$$

Since (62)–(63) hold for an arbitrary $a > |d_{k_0}|_p$ and $\beta \in (0, 1)$, it follows that $\sup_{k \geq k_0} |d_k|_p^2 \leq \mathsf{c}_p(x^*)[1 + |d_{k_0}|_p^2]$, with $\mathsf{c}_p(x^*)$ as in the statement of the proposition. This contradicts the initial assumption that $\tau_a < \infty$ for all $a > |d_{k_0}|_p$. Hence there exists $\bar{a} > |d_{k_0}|_p$ such that $\hat{a} := \sup_{k \geq k_0} |d_k|_p \leq \bar{a} < \infty$ by the definition of $\tau_{\bar{a}}$. For any $k > k_0$, we use the fact that $|d_i|_p \leq \hat{a}$ for $k_0 \leq i < k$ in (59), obtaining

$$|d_k|_p^2 \leq |d_{k_0}|_p^2 + \mathsf{B}_p(x^*)\sqrt{\gamma}\left(1 + \hat{a} + \hat{a}^2\right) + \mathsf{D}_p(x^*)\gamma\left(1 + \hat{a}^2\right)$$
$$(64) \qquad + \mathsf{D}_p(x^*)^2\gamma^2\left(1 + \hat{a}^2\right).$$

The inequality (64) holds trivially for $k := k_0$. Thus, after taking the supremum over $k \geq k_0$ in (64), we proceed as we did immediately after inequality (61), obtaining (62) and (63), respectively, for $p = 2$ and $p \geq 4$, but with $\hat{a}$ substituting for a. Using the definition of $\mathsf{c}_p(x^*)$, the claim follows.

(ii) The proof line is the same as for the case $p = 2$ in item (i), but summing (41) with the estimate (47), which gives the following uniform estimate: for all $k \geq 0$, $\mathsf{C}_k\left(\Pi\left(\bar{x}^k\right)\right)\mathcal{N}_k^{-1} \leq 2c\hat{\alpha}^2 C_2^2\sigma^2\mathcal{N}_k^{-1}\left(1 + 2\hat{\alpha}^2 C_2^2\sigma^2\mathcal{N}_k^{-1}\right)$. We remark that we may replace $\beta(x^*)$ in (58) and (62) by $\beta := 2c\hat{\alpha}^2 C_2^2\sigma^2 + 4c\hat{\alpha}^4 C_2^4\sigma^4$. In this case, the definition of ϕ and k_0 imply that $0 < 1 - \phi - \phi^2 \leq 1 - \beta$.

(iii) Given $k \in \mathbb{N}$, we take total expectation in (52) and sum from 0 to k, obtaining

$$\mathbb{E}\left[\mathrm{d}(x^{k+1}, X^*)^2\right] \leq \mathrm{d}(x^0, X^*)^2 + \sum_{i=0}^{k}\frac{17C_2^2\hat{\alpha}^2\sigma^2}{N_{i,\min}} \leq \mathrm{d}(x^0, X^*)^2 + \sum_{i=0}^{\infty}\frac{17C_2^2\hat{\alpha}^2\sigma^2}{N_{i,\min}},$$

and the claim follows. □

Remark 3.21. In the statement of Proposition 3.20(i), for $p \geq 2$, it is sufficient to set $\phi \in (0, \frac{\sqrt{5}-1}{2})$ and $k_0 \in \mathbb{N}_0$ such that $\sum_{k \geq k_0}\mathcal{N}_k^{-1} \leq \phi\mathsf{D}(x^*)^{-1}$ to obtain $\sup_{k \geq k_0}\mathbb{E}[\|x^k - x^*\|^2] \leq \frac{1 + \mathbb{E}[\|x^{k_0} - x^*\|^2]}{1 - \phi - \phi^2}$.

We now give explicit estimates on the convergence rate and oracle complexity. In what follows we assume that the stepsize sequence is constant. Proposition 10.3.6 in [11] states that $\{r_a : a > 0\}$ is a family of equivalent merit functions of $\mathrm{VI}(T, X)$. Hence, the convergence rate analysis can be deduced for varying stepsizes satisfying Assumption 3.7, and constant stepsizes are assumed just for simplicity. Recall the definition of $\mathsf{D}(x^*)$ in (56). Define, for $k, \ell \in \mathbb{N}_0 \cup \{\infty\}$, $\phi \in \mathbb{R}$, $x^* \in X^*$, and $\alpha \in \left(0, 1/\sqrt{6}L\right)$,

$$(65) \qquad \rho := 1 - 6\alpha^2 L^2, \qquad \mathsf{a}_0^k := \sum_{i=0}^{k}\frac{1}{\mathcal{N}_i}, \qquad \mathsf{b}_0^k := \sum_{i=0}^{k}\frac{1}{\mathcal{N}_i^2},$$

$$(66) \qquad \mathsf{J}(x^*, k, \phi) := \frac{1 + \max_{0 \leq i \leq k}\mathbb{E}[\|x^i - x^*\|^2]}{1 - \phi - \phi^2},$$

$$(67) \quad \mathsf{Q}_k(x^*, \ell, \phi) := \frac{2}{\rho}\left\{\|x^0 - x^*\|^2 + [1 + \mathsf{J}(x^*, \ell, \phi)]\left[\mathsf{D}(x^*)\mathsf{a}_0^k + \mathsf{D}(x^*)^2\mathsf{b}_0^k\right]\right\}.$$

THEOREM 3.22 (convergence rate: nonuniform variance). *Consider Assumptions 3.1–3.8(i) for some $x^* \in X^*$ and take $\alpha_k \equiv \alpha \in (0, 1/\sqrt{6}L)$. Choose $\phi \in (0, \frac{\sqrt{5}-1}{2})$ and $k_0 := k_0(x^*) \in \mathbb{N}$ such that*

$$(68) \qquad \sum_{k \geq k_0}\frac{1}{\mathcal{N}_k} \leq \frac{\phi}{\mathsf{D}(x^*)},$$

where $\mathsf{D}(x^*)$ is defined in (56). Then for all $\epsilon > 0$ there exists $K_\epsilon \in \mathbb{N}$ such that

$$\mathbb{E}[r_\alpha(x^{K_\epsilon})^2] \leq \epsilon \leq \frac{\mathsf{Q}_\infty(x^*, k_0(x^*), \phi)}{K_\epsilon}.$$

Additionally, if Assumption 3.8(ii) holds, then K_ϵ is independent of $x^* \in X^*$.

Proof. First note that finiteness of $\mathsf{a}_0^\infty, \mathsf{b}_0^\infty$ as defined before the statement of the theorem follows from Assumption 3.5, which also ensures existence of $k_0(x^*)$ satisfying (68), because $\sum_{i>k} \mathcal{N}_i^{-1} \to 0$ as $k \to \infty$. Set ϕ and $k_0 := k_0(x^*)$ such that (68) holds. Proposition 3.20(i) for $p = 2$ and Remark 3.21 imply

$$
\sup_{k \geq k_0} \mathbb{E}[\|x^k - x^*\|^2] \leq \frac{1 + \mathbb{E}[\|x^{k_0} - x^*\|^2]}{1 - \phi - \phi^2}
$$
$$
\leq \frac{1 + \max_{0 \leq k \leq k_0} \mathbb{E}[\|x^k - x^*\|^2]}{1 - \phi - \phi^2} = \mathsf{J}(x^*, k_0, \phi).
$$

From the above inequality and the fact that $1 - \phi - \phi^2 \in (0, 1)$ we get the following uniform bound:

$$
(69) \qquad\qquad \sup_{k \geq 0} \mathbb{E}[\|x^k - x^*\|^2] \leq \mathsf{J}(x^*, k_0(x^*), \phi).
$$

We now invoke Proposition 3.15. Given $0 \leq i \leq k$, we take total expectation in (41), with the estimate (47) using $\mathsf{A} \leq 2$, $c > 1$ defined in Remark 3.17, the definition of $\hat{\alpha}$ in Assumption 3.7, and the definition of $\mathsf{D}(x^*)$ in (56). We then sum iteratively with i running from 0 to k, obtaining

(70)

$$
\frac{\rho}{2} \sum_{i=0}^{k} \mathbb{E}[r_\alpha(x^i)^2]
$$

$$
\leq \|x^0 - x^*\|^2 + \mathsf{D}(x^*) \sum_{i=0}^{k} \frac{1 + \mathbb{E}\left[\|x^i - x^*\|^2\right]}{\mathcal{N}_i} + \mathsf{D}(x^*)^2 \sum_{i=0}^{k} \frac{1 + \mathbb{E}\left[\|x^i - x^*\|^2\right]}{\mathcal{N}_i^2}
$$

$$
\leq \|x^0 - x^*\|^2 + \left(1 + \sup_{0 \leq i \leq k} \mathbb{E}[\|x^i - x^*\|^2]\right) \left(\mathsf{D}(x^*) \sum_{i=0}^{k} \frac{1}{\mathcal{N}_i} + \mathsf{D}(x^*)^2 \sum_{i=0}^{k} \frac{1}{\mathcal{N}_i^2}\right)
$$

$$
\leq \|x^0 - x^*\|^2 + [1 + \mathsf{J}(x^*, k_0(x^*), \phi)] \left[\mathsf{D}(x^*)\mathsf{a}_0^k + \mathsf{D}(x^*)^2 \mathsf{b}_0^k\right] = \frac{\rho}{2} \mathsf{Q}_k(x^*, k_0(x^*), \phi),
$$

using (69) in the last inequality.

Given $\epsilon > 0$, define $K = K_\epsilon := \inf\{k \in \mathbb{N}_0 : \mathbb{E}[r_\alpha(x^k)^2] \leq \epsilon\}$. For every $k < K$ we have

$$
(71) \qquad\qquad \frac{\rho}{2}\epsilon(k+1) < \frac{\rho}{2} \sum_{i=0}^{k} \mathbb{E}[r_\alpha(x^i)^2],
$$

using the fact that $\mathbb{E}[r_\alpha(x^i)^2] > \epsilon$ for all $0 \leq i \leq k$, which follows from the definition of K.

We claim that K is finite. Indeed, if $K = \infty$, then (70) and (71) hold for all $k \in \mathbb{N}$. Hence, we arrive at a contradiction by letting $k \to \infty$ and using the facts that $\mathsf{a}_0^\infty < \infty$ and $\mathsf{b}_0^\infty < \infty$, which hold by Assumption 3.5. Since K is finite, we have $\mathbb{E}[r_\alpha(x^K)^2] \leq \epsilon$ by definition. Setting $k := M - 1$ in (70)–(71), we get $K \leq \frac{\mathsf{Q}_{K-1}(x^*, k_0, \phi)}{\epsilon} \leq \frac{\mathsf{Q}_\infty(x^*, k_0, \phi)}{\epsilon}$, using the definition of $\mathsf{Q}_k(x^*, k_0, \phi)$. We have thus proved the claim. Under Assumption 3.8(ii), the proof is valid for any $x^* \in X^*$ and, hence, K is independent of $x^* \in X^*$. $\qquad\square$

In the previous theorem, given $x^* \in X^*$, the constant $\mathsf{Q}_\infty(x^*, k_0(x^*), \phi)$ in the convergence rate depends on the variance $\sigma(x^*)^2$ and on the distance of the $k_0(x^*)$ initial iterates to x^*, where $k_0(x^*)$ and ϕ are chosen such that (68) is satisfied. Under Assumption 3.8(ii), since K_ϵ does not depend on $x^* \in X^*$, we get indeed the uniform estimate

$$(72) \qquad \sup_{\epsilon > 0} \epsilon K_\epsilon \leq \inf_{x^* \in X^*} \mathsf{Q}_\infty(x^*, k_0(x^*), \phi).$$

In view of (72), the performance of method (11)–(12) under nonuniform variance depends on the $x^* \in X^*$ such that $\mathsf{Q}_\infty(x^*, k_0(x^*), \phi)$ is minimal.

PROPOSITION 3.23 (rate and oracle complexity for $m = 1$: nonuniform variance). *Consider Assumptions* 3.1–3.8(i) *for some* $x^* \in X^*$ *and take* $\alpha_k \equiv \alpha \in (0, 1/\sqrt{6}L)$. *Define* $\mathcal{N}_k$ *as*

$$(73) \qquad \mathcal{N}_k = \left\lceil \theta(k + \mu)(\ln(k + \mu))^{1+b} \right\rceil,$$

for any $\theta > 0$, $b > 0$, $\epsilon > 0$, *and* $2 < \mu \leq \epsilon^{-1}$. *Choose* $\phi \in (0, \frac{\sqrt{5}-1}{2})$ *and let* $k_0(x^*)$ *be the minimum natural number satisfying*

$$(74) \qquad k_0(x^*) \geq \exp\left[\left(\frac{2cC_2^2 \hat\alpha^2 \sigma(x^*)^2}{\phi b \theta} \right)^{1/b} \right] - \mu + 1.$$

Then Theorem 3.18 *holds and there are nonnegative constants* $\overline{\mathsf{Q}}(x^*)$, $\mathsf{P}(x^*)$, *and* $\mathsf{I}(x^*)$ *depending on* x^*, $k_0(x^*)$, *and* ϕ *such that for all* $\epsilon > 0$, *there exists* $K := K_\epsilon \in \mathbb{N}$ *such that*

$$(75) \quad \mathbb{E}[r_\alpha(x^K)^2] \leq \epsilon \leq \frac{\max\{1, \theta^{-2}\}\overline{\mathsf{Q}}(x^*)}{K},$$

$$(76) \qquad \sum_{k=1}^{K} 2\mathcal{N}_k \leq \frac{\max\{1, \theta^{-4}\}\max\{1, \theta\}\mathsf{I}(x^*)\left\{ \left[\ln\left(\mathsf{P}(x^*)\epsilon^{-1}\right)\right]^{1+b} + \frac{1}{\mu} \right\}}{\epsilon^2}.$$

Proof. For $\phi \in (0, \frac{\sqrt{5}-1}{2})$, we want $k_0 := k_0(x^*)$ to satisfy (68) of Theorem 3.22. We have

$$\sum_{k \geq k_0} \frac{1}{\mathcal{N}_k} \leq \theta^{-1} \sum_{k \geq k_0} \frac{1}{(k + \mu)(\ln(k + \mu))^{1+b}}$$

$$\leq \theta^{-1} \int_{k_0 - 1}^{\infty} \frac{dt}{(t + \mu)(\ln(t + \mu))^{1+b}}$$

$$(77) \qquad = \frac{\theta^{-1}}{b(\ln(k_0 - 1 + \mu))^b}.$$

From (77) and (68), it is enough to choose k_0 as the minimum natural number such that the RHS of (77) is less than $\phi/\mathsf{D}(x^*)$. Using the definition of $\mathsf{D}(x^*)$ in (56), it is enough to choose k_0 as in (74).

We now give an estimate of $\mathsf{Q}_\infty(x^*, k_0, \phi)$ as defined in (67). Set $\lambda := 2c\hat\alpha^2 C_2^2$ with c as defined in Remark 3.17. From the definitions (56) and (65), we have the bound

$$(78) \qquad \mathsf{D}(x^*)\mathsf{a}_0^\infty + \mathsf{D}(x^*)^2\mathsf{b}_0^\infty \leq \int_{-1}^{\infty} \frac{\lambda\theta^{-1}\sigma(x^*)^2 dt}{(t + \mu)(\ln(t + \mu))^{1+b}}$$

$$+ \int_{-1}^{\infty} \frac{\lambda^2\theta^{-2}\sigma(x^*)^4 dt}{(t + \mu)^2(\ln(t + \mu))^{2+2b}} \leq \frac{\lambda\theta^{-1}\sigma(x^*)^2}{b(\ln(\mu - 1))^b} + \frac{\lambda^2\theta^{-2}\sigma(x^*)^4}{(\mu - 1)(1 + 2b)[\ln(\mu - 1)]^{1+2b}}.$$

Using Theorem 3.22, (78), and the definition of $\mathsf{Q}_\infty(x^*, k_0, \phi)$ as defined by (65)–(67), we prove (75), noting that $\overline{\mathsf{Q}}(x^*)$ is specified as in Remark 3.24.

We now prove (76). Denoting $\mathsf{Q}_\infty(x^*) := \mathsf{Q}_\infty(x^*, k_0, \phi)$ and using $K := K_\epsilon \leq \mathsf{Q}_\infty(x^*)/\epsilon$, $\mu\epsilon \leq 1$, and $\mathcal{N}_k \leq \theta(k + \mu)(\ln(k + \mu))^{1+b} + 1$, we have

$$\sum_{k=1}^{K} 2\mathcal{N}_k \leq \max\{\theta, 1\} \sum_{k=1}^{K} 2\left[(k + \mu)(\ln(k + \mu))^{1+b} + 1\right]$$

$$\leq \max\{\theta, 1\} K(K + 2\mu) \left[(\ln(K + \mu))^{1+b} + \frac{2}{K + 2\mu}\right]$$

$$(79) \qquad \leq \max\{\theta, 1\} \frac{\left\{\left[\ln\left(\mathsf{Q}_\infty(x^*)\epsilon^{-1} + \epsilon^{-1}\right)\right]^{1+b} + \mu^{-1}\right\} \mathsf{Q}_\infty(x^*)\left(\mathsf{Q}_\infty(x^*) + 2\right)}{\epsilon^2}.$$

We now use (79) with $\mathsf{Q}_\infty(x^*)(\mathsf{Q}_\infty(x^*) + 2) \leq (\mathsf{Q}_\infty(x^*) + 2)^2$, the definition of $\mathsf{Q}_\infty(x^*, k_0, \phi)$ as in (65)–(67), the bound (78), and the fact that $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ in order to prove (76), where $\mathsf{I}(x^*)$ and $\mathsf{P}(x^*)$ are given in Remark 3.24. $\qquad\square$

Remark 3.24 (constants). We make use of the following definitions for the sake of clarity:

$$(80) \qquad \mathcal{A}_{\mu,b} := \frac{2c\hat{\alpha}^2 C_2^2}{b[\ln(\mu - 1)]^b}, \quad \mathcal{B}_{\mu,b} := \frac{(2c\hat{\alpha}^2 C_2^2)^2}{(\mu - 1)(1 + 2b)[\ln(\mu - 1)]^{1+2b}},$$

$$(81) \qquad \overline{\mathsf{Q}}(d, A, J) := 2\rho^{-1}d^2 + 2\rho^{-1}A(1 + J),$$

$$(82) \qquad \mathsf{I}(d, A, J) := 12\rho^{-2}d^4 + 12\rho^{-2}A^2(1 + J)^2 + 1,$$

using the definition of c in Remark 3.17 and (65). Using the definitions (56), (65)–(67), and (80)–(82), the constants in the statement of Proposition 3.23 are given by

$$\overline{\mathsf{Q}}(x^*) := \overline{\mathsf{Q}}(\|x^0 - x^*\|, \sigma(x^*)^2 \mathcal{A}_{\mu,b} + \sigma(x^*)^4 \mathcal{B}_{\mu,b}, \mathsf{J}(x^*, k_0(x^*), \phi)),$$
$$\mathsf{P}(x^*) := \mathsf{Q}_\infty(x^*, k_0(x^*), \phi) + 1,$$
$$\mathsf{I}(x^*) := \mathsf{I}(\|x^0 - x^*\|, \sigma(x^*)^2 \mathcal{A}_{\mu,b} + \sigma(x^*)^4 \mathcal{B}_{\mu,b}, \mathsf{J}(x^*, k_0(x^*), \phi)).$$

Given $\epsilon > 0$, we may use the definitions of $\overline{\mathsf{Q}}(x^*)$, $\mathsf{I}(x^*)$, and $\mathsf{P}(x^*)$ and optimize the estimates given in (75)–(76) over $(\hat{\alpha}, \theta)$, obtaining optimal constants in terms of L and $\sigma(x^*)^2$. For simplicity we do not carry this procedure here.

We give the next sharper estimates in the case the variance is *uniform* over X^* or X. We state them without proofs since they follow the same proof line of Theorem 3.22 and Proposition 3.23, but using Propositions 3.19(ii) and 3.20(ii) when the variance is uniform over X^* and Proposition 3.19(iii) when the variance is uniform over X. Define

$$(83) \qquad \mathsf{D}_\sigma := 2c\hat{\alpha}^2 C_2^2 \sigma^2,$$

$$(84) \qquad \mathsf{J}(\ell, \phi) := \frac{1 + \max_{0 \leq k \leq \ell} \mathbb{E}[\mathsf{d}(x^k, X^*)^2]}{1 - \phi - \phi^2},$$

$$(85) \qquad \mathsf{Q}_k(\sigma, \ell, \phi) := 2\rho^{-1}\left\{\mathsf{d}(x^0, X^*)^2 + (1 + \mathsf{J}(\ell, \phi))\left(\mathsf{D}_\sigma \mathsf{a}_0^k + \mathsf{D}_\sigma^2 \mathsf{b}_0^k\right)\right\},$$

$$(86) \qquad \widetilde{\mathsf{Q}}_k(\sigma) := 2\rho^{-1}\left\{\mathsf{d}(x^0, X^*)^2 + 17 C_2^2 \hat{\alpha}^2 \sigma^2 \sum_{i=0}^{k} \frac{1}{N_{i,\min}}\right\},$$

using c as defined in Remark 3.17 and the definitions in (65).

THEOREM 3.25 (convergence rate: uniform variance).   *Consider Assumptions* 3.1–3.8 *and take* $\alpha_k \equiv \alpha \in (0, 1/\sqrt{6}L)$.

Suppose first that Assumption 3.8(ii) *holds and* $\sup_{x^* \in X^*} \sigma(x^*) \le \sigma$ *for some* $\sigma > 0$. *Take* $\phi \in (0, \frac{\sqrt{5}-1}{2})$ *and* $k_0 := k_0(\sigma) \in \mathbb{N}$ *such that*

$$\sum_{k \ge k_0} \frac{1}{\mathcal{N}_k} \le \frac{\phi}{\mathsf{D}_\sigma}.$$

Then, for all $\epsilon > 0$, *there exists* $K_\epsilon \in \mathbb{N}$, *satisfying*

$$\mathbb{E}[r_\alpha \left(x^{K_\epsilon}\right)^2] \le \epsilon \le \frac{\mathsf{Q}_\infty(\sigma, k_0(\sigma), \phi)}{K_\epsilon}.$$

Suppose now that Assumption 3.8(iii) *holds for some* $\sigma > 0$. *Then, for all* $\epsilon > 0$, *there exists* $K_\epsilon \in \mathbb{N}$, *satisfying*

$$\mathbb{E}[r_\alpha \left(x^{K_\epsilon}\right)^2] \le \epsilon \le \frac{\widetilde{\mathsf{Q}}_\infty(\sigma)}{K_\epsilon}.$$

PROPOSITION 3.26 (rate and oracle complexity for $m = 1$: uniform variance). *Consider Assumptions* 3.1–3.8, *take* $\alpha_k \equiv \alpha \in (0, 1/\sqrt{6}L)$, *and suppose* $\sup_{x^* \in X^*} \sigma(x^*) \le \sigma$ *for some* $\sigma > 0$. *Define* $\mathcal{N}_k$ *as*

$$\mathcal{N}_k = \left\lceil \theta(k + \mu)(\ln(k + \mu))^{1+b} \right\rceil,$$

for any $\theta > 0$, $b > 0$, $\epsilon > 0$, *and* $2 < \mu \le \epsilon^{-1}$. *Then the following holds:*

   (i) *Suppose Assumption* 3.8(ii) *holds. Choose* $\phi \in (0, \frac{\sqrt{5}-1}{2})$ *and* $k_0 := k_0(\sigma) \in \mathbb{N}$ *such that*

$$k_0 \ge \exp\left[\left(\frac{2cC_2^2\hat{\alpha}^2\sigma^2}{\phi b\theta}\right)^{1/b}\right] - \mu + 1.$$

   Then Theorem 3.18 *holds and there exist nonnegative constants* $\overline{\mathsf{Q}}(\sigma)$, $\mathsf{P}(\sigma)$, *and* $\mathsf{I}(\sigma)$ *depending on* σ, $k_0(\sigma)$, *and* ϕ *such that for all* $\epsilon > 0$, *there exists* $K := K_\epsilon \in \mathbb{N}$ *such that*

$$\mathbb{E}[r_\alpha(x^K)^2] \le \epsilon \le \frac{\max\{1, \theta^{-2}\}\overline{\mathsf{Q}}(\sigma)}{K},$$

$$\sum_{k=1}^{K} 2\mathcal{N}_k \le \frac{\max\{1, \theta^{-4}\}\max\{1, \theta\}\mathsf{I}(\sigma)\left\{\left[\ln\left(\mathsf{P}(\sigma)\epsilon^{-1}\right)\right]^{1+b} + \frac{1}{\mu}\right\}}{\epsilon^2}.$$

   (ii) *Suppose that Assumption* 3.8(iii) *holds. Then Theorem* 3.18 *holds and there exist nonnegative constants* $\widetilde{\mathsf{Q}}(\sigma)$, $\widetilde{\mathsf{P}}(\sigma)$, *and* $\widetilde{\mathsf{I}}(\sigma)$ *depending on* σ *such that for all* $\epsilon > 0$, *there exists* $K := K_\epsilon \in \mathbb{N}$ *such that*

$$\mathbb{E}[r_\alpha(x^K)^2] \le \epsilon \le \frac{\max\{1, \theta^{-1}\}\widetilde{\mathsf{Q}}(\sigma)}{K_\epsilon},$$

$$\sum_{k=1}^{K} 2\mathcal{N}_k \le \frac{\max\{1, \theta^{-2}\}\max\{1, \theta\}\widetilde{\mathsf{I}}(\sigma)\left\{\left[\ln\left(\widetilde{\mathsf{P}}(\sigma)\epsilon^{-1}\right)\right]^{1+b} + \frac{1}{\mu}\right\}}{\epsilon^2}.$$

Remark 3.27 (constants).   We recall the definitions (65), (80)–(82), and (83)–(86). The constants in the statement of Proposition 3.26(i) are given by $\overline{\mathsf{Q}}(\sigma) := \overline{\mathsf{Q}}(\mathrm{d}(x^0, X^*), \sigma^2 \mathcal{A}_{\mu,b} + \sigma^4 \mathcal{B}_{\mu,b}, \mathsf{J}(k_0(\sigma), \phi))$, $\mathsf{P}(\sigma) := \mathsf{Q}_\infty(\sigma, k_0(\sigma), \phi) + 1$, and $\mathsf{I}(\sigma) := \mathsf{I}(\mathrm{d}(x^0, X^*), \sigma^2 \mathcal{A}_{\mu,b} + \sigma^4 \mathcal{B}_{\mu,b}, \mathsf{J}(k_0(\sigma), \phi))$. For item (ii) they are given by

$$\widetilde{\mathsf{Q}}(\sigma) := 2\rho^{-1}\, \mathrm{d}(x^0, X^*)^2 + 2\rho^{-1} \frac{17 C_2^2 \hat{\alpha}^2 \sigma^2}{b(\ln(\mu - 1))^b},$$

$$\widetilde{\mathsf{I}}(\sigma) := 12\rho^{-2}\, \mathrm{d}(x^0, X^*)^4 + 12\rho^{-2} \frac{17^2 C_2^4 \hat{\alpha}^4 \sigma^4}{b^2(\ln(\mu - 1))^{2b}} + 1,$$

and $\widetilde{\mathsf{P}}(\sigma) := \widetilde{\mathsf{Q}}_\infty(\sigma) + 1$.

We now turn our attention to the distributed solution of a Cartesian SVI for a large network ($m \gg 1$). If a *decentralized sampling* is used, then higher order factors of m appear in rate and complexity. The next result shows that if, in addition, a deterministic and decreasing sequence of exponents $\{b_i\}_{i=1}^m$ and an approximate estimate of the network dimension m is coordinated, then the oracle complexity is proportional to m (up to a scaling in the sampling rate).

PROPOSITION 3.28 (rate and oracle complexity for a network).   *Consider Assumptions* 3.1–3.8(i) *for some* $x^* \in X^*$ *and take* $\alpha_k \equiv \alpha \in (0, 1/\sqrt{6}L)$. *Under Assumptions* 3.1–3.8(i) *with Assumption* 3.6(i) *(centralized sampling), the results of Proposition* 3.23 *hold.*

Consider now Assumption 3.6(ii) *(decentralized sampling). Let* $\{b_i\}_{i=1}^m$ *be a positive sequence such that*

$$(87) \qquad N_{k,i} = \left\lceil \theta_i (k + \mu_i)^{1+a} (\ln(k + \mu_i))^{1+b_i} \right\rceil,$$

$$(88) \qquad b_1 \geq b_i + 2\ln(i+1) - \ln \mathsf{S},$$

for any $\theta_i > 0$, $a > 0$, $\mathsf{S} \geq 1$, $\epsilon > 0$, $2 < \mu_i \leq \epsilon^{-1}$. *Choose* $\phi \in (0, \frac{\sqrt{5}-1}{2})$ *and let* $k_0(x^*)$ *be the minimum natural number greater than* $e - \mu_{\min} + 1$ *such that*

$$(89) \qquad k_0(x^*) \geq \left[\frac{2c C_2^2 \hat{\alpha}^2 \sigma(x^*)^2}{\phi \theta_{\min} b_{\min}} \right]^{1/a} - \mu_{\min} + 1.$$

Then Theorem 3.18 *holds and there exist nonnegative constants* $\widehat{\mathsf{Q}}(x^*)$, $\widehat{\mathsf{P}}(x^*)$, *and* $\widehat{\mathsf{I}}(x^*)$ *depending on* x^*, $k_0(x^*)$, *and* ϕ *such that for all* $\epsilon > 0$, *there exists* $K := K_\epsilon \in \mathbb{N}$ *such that*

$$(90) \qquad \mathbb{E}[r_\alpha(x^K)^2] \leq \epsilon \leq \frac{\widehat{\mathsf{Q}}(x^*)}{K},$$

$$(91) \qquad \sum_{k=1}^{K} \sum_{i=1}^{m} 2N_{k,i} \leq \frac{\mathsf{S} \max\{\theta_{\max}, 1\} \left\{ \ln\left(\widehat{\mathsf{P}}(x^*) \epsilon^{-1} \right) \right\}^{1+b_1} \widehat{\mathsf{I}}(x^*)}{\epsilon^{2+a}}.$$

*(Above, the subscripts "*min*" and "*max*" refer, respectively, to the minimal and maximal terms of the corresponding sequences.)*

Proof. In what follows we will use the following estimate. For any $k \in \mathbb{N}_0$, $a > 0$, $0 < b < 1$, $\mu > 1$,

$$(92) \qquad \int_k^\infty \frac{dt}{(t+\mu)^{1+a}(\ln(t+\mu))^{1+b}} \le \max\left\{\frac{1}{a(k+\mu)^a}, \frac{1}{(k+\mu)^a b[\ln(k+\mu)]^b}\right\}.$$

For $\phi \in (0, \frac{\sqrt{5}-1}{2})$ we want $k_0 := k_0(x^*)$ to satisfy (68) of Theorem 3.22. Since $\mathcal{N}_k$ is the harmonic average of $\{N_{k,i}\}_{i=1}^m$ and $N_{k,i} \ge \theta_{\min}(k+\mu_{\min})^{1+a}[\ln(k+\mu_{\min})]^{1+b_{\min}}$ for all $i \in [m]$, we get from (92)

$$\sum_{k \ge k_0} \frac{1}{\mathcal{N}_k} \le \theta_{\min}^{-1} \sum_{k \ge k_0} \frac{1}{(k+\mu_{\min})^{1+a}[\ln(k+\mu_{\min})]^{1+b_{\min}}}$$

$$(93) \qquad \le \frac{\theta_{\min}^{-1}}{(k_0-1+\mu_{\min})^a b_{\min}[\ln(k_0-1+\mu_{\min})^{b_{\min}}]} \le \frac{\theta_{\min}^{-1}}{(k_0-1+\mu_{\min})^a b_{\min}},$$

if $k_0 \ge e - \mu_{\min} + 1$. From (93) and (68), it is enough to choose k_0 as the minimum natural number greater than $e - \mu_{\min} + 1$ such that the RHS of (93) is less than $\phi/\mathsf{D}(x^*)$. Using the definition of $\mathsf{D}(x^*)$ in (56), it is enough to choose k_0 as in (89).

Next we estimate the value of $\mathsf{Q}_\infty(x^*, k_0, \phi)$ as defined in (67). Recall $\frac{1}{\mathcal{N}_k} = \sum_{i=1}^m \frac{1}{N_{k,i}}$ and set $\lambda := 2c\hat{\alpha}^2 C_2^2$ with c as defined in Remark 3.17. Definitions (56) and (65) imply

$$\mathsf{D}(x^*)\mathsf{a}_0^\infty + \mathsf{D}(x^*)^2 \mathsf{b}_0^\infty \le \sigma(x^*)^2 \sum_{k \ge 0} \sum_{i=1}^m \frac{\lambda \theta_i^{-1}}{(k+\mu_i)^{1+a}(\ln(k+\mu_i))^{1+b_i}}$$

$$(94) \qquad\qquad + \sigma(x^*)^4 \sum_{k \ge 0} \left[\sum_{i=1}^m \frac{\lambda \theta_i^{-1}}{(k+\mu_i)^{1+a}(\ln(k+\mu_i))^{1+b_i}}\right]^2.$$

The first summation in (94) is bounded by

$$(95) \qquad \sum_{i=1}^m \sum_{k \ge 0} \frac{\lambda \theta_i^{-1}}{(k+\mu_i)^{1+a}} \le \sum_{i=1}^m \int_{-1}^\infty \frac{\lambda \theta_i^{-1} dt}{(t+\mu_i)^{1+a}} \le \sum_{i=1}^m \frac{\lambda}{\theta_i a(\mu_i-1)^a} =: \mathcal{A}_m.$$

Using estimate (92), the second summation in (94) is bounded by

$$\sum_{i=1}^m \sum_{j=1}^m \sum_{k \ge 0} \frac{\lambda^2 \theta_i^{-1} \theta_j^{-1}}{(k+\mu_{\min})^{2+2a}[\ln(k+\mu_{\min})]^{2+b_i+b_j}}$$

$$(96) \qquad \le \frac{1}{\vartheta}\left\{\sum_{i=1}^m \frac{\lambda}{\theta_i[\ln(\mu_{\min}-1)]^{b_i}}\right\}^2 =: \mathcal{B}_m,$$

where $\vartheta := (1+2b_{\min})(\mu_{\min}-1)^{1+2a}\ln(\mu_{\min}-1)$. From Theorem 3.22, (94)–(96), and the definition of $\mathsf{Q}_\infty(x^*, k_0, \phi)$ as specified in (65)–(67), we prove (90), noting that $\widehat{\mathsf{Q}}(x^*)$ is specified as in Remark 3.29.

We now prove the bound on the oracle complexity. Let $\mathsf{Q}_\infty(x^*) := \mathsf{Q}_\infty(x^*, k_0, \phi)$. Using the facts that $K \le \mathsf{Q}_\infty(x^*)/\epsilon$, $\mu_i \le \epsilon^{-1}$ and the definition of $N_{k,i}$, we have

$$\sum_{k=1}^{K}\sum_{i=1}^{m} 2N_{k,i} \leq \sum_{k=1}^{K}\sum_{i=1}^{m} 2\left[\theta_i(k+\mu_i)^{1+a}(\ln(k+\mu_i))^{1+b_i} + 1\right]$$

$$\leq 2\max\{\theta_{\max},1\}K\sum_{i=1}^{m}\left[(K+\mu_i)^{1+a}\left(\ln(K+\mu_i)\right)^{1+b_i} + 1\right]$$

$$\leq 4\max\{\theta_{\max},1\}K\sum_{i=1}^{m}\left[(K+\mu_i)^{1+a}\left(\ln(K+\mu_i)\right)^{1+b_i}\right]$$

(97) $$\leq 4\Phi\frac{(\mathsf{Q}_\infty(x^*)+1)^{2+a}}{\epsilon^{2+a}}\sum_{i=1}^{m}\left(\ln\left(\mathsf{Q}_\infty(x^*)\epsilon^{-1}+\epsilon^{-1}\right)\right)^{1+b_i},$$

using the fact that $1 \leq (K+\mu_i)^{1+a}\left(\ln(K+\mu_i)\right)^{1+b_i}$ for $i \in [m]$ in the third inequality, and defining $\Phi := \max\{\theta_{\max},1\}$ in the last inequality.

Set $h := \ln\left(\mathsf{Q}_\infty(x^*)\epsilon^{-1}+\epsilon^{-1}\right)$ with $h \geq e$ for sufficiently small $\epsilon > 0$. By definition of $\{b_i\}_{i=1}^{m}$ we have, for $i \in [m]$,

(98) $$b_1 \geq b_i + 2\ln(i+1) - \ln\mathsf{S} \geq b_i + \frac{2\ln(i+1) - \ln\mathsf{S}}{\ln h} \Rightarrow h^{b_i} \leq \frac{\mathsf{S}h^{b_1}}{(i+1)^2}.$$

From (98) we obtain that

(99) $$\sum_{i=1}^{m} h^{b_i} \leq \mathsf{S}h^{b_1}\sum_{i=1}^{m}\frac{1}{(i+1)^2} \leq \mathsf{S}h^{b_1}.$$

From the bounds (94)–(99), the definitions of h and $\mathsf{Q}_\infty(x^*,k_0,\phi)$, as specified in (65)–(67), and the fact that $(x+y+z)^{2+a} \leq 3^{1+a}(x^{2+a}+y^{2+a}+z^{2+a})$, we obtain the required bound on $\sum_{k=1}^{K}\sum_{i=1}^{m} 2N_{k,i}$, noting that $\widehat{\mathsf{I}}(x^*)$ and $\widehat{\mathsf{P}}(x^*)$ are specified as in Remark 3.29. □

Remark 3.29 (constants). Define

(100) $$\widehat{\mathsf{I}}(d, A, J, \nu) := 4 \cdot 3^{\nu-1}\left\{(2\rho^{-1})^\nu d^{2\nu} + (2\rho^{-1})^\nu A^\nu[1+J]^\nu + 1\right\},$$

using (65). In view of (56), (65)–(67), (81), (100) and (95)–(96), the constants in the statement of Proposition 3.28 are given by

$$\widehat{\mathsf{Q}}(x^*) := \overline{\mathsf{Q}}(\|x^0 - x^*\|, \sigma(x^*)^2\mathcal{A}_m + \sigma(x^*)^4\mathcal{B}_m, \mathsf{J}(x^*, k_0(x^*), \phi)),$$
$$\widehat{\mathsf{P}}(x^*) := \mathsf{Q}_\infty(x^*, k_0(x^*), \phi) + 1,$$
$$\widehat{\mathsf{I}}(x^*) := \widehat{\mathsf{I}}(\|x^0 - x^*\|, \sigma(x^*)^2\mathcal{A}_m + \sigma(x^*)^4\mathcal{B}_m, \mathsf{J}(x^*, k_0(x^*), \phi), 2 + a).$$

Remark 3.30 (oracle complexity of $\mathcal{O}(m)$). For the choice of parameters (87)–(88), if we have $\theta_i \sim \theta m$ for some $\theta > 0$, then $\mathcal{A}_m \lesssim \frac{\theta^{-1}}{a(\mu_{\min}-1)^a}$ and $\mathcal{B}_m \lesssim \frac{\theta^{-2}}{(\mu_{\min}-1)^{1+2a}}$, where $\mathcal{A}_m$ and $\mathcal{B}_m$ are defined in (95)–(96). Also, $b_{\min} \leq b_1 + \ln\mathsf{S} - 2\ln(m+1)$ so that it is enough to choose $b_1 > 2\ln(m+1) - \ln\mathsf{S}$, which is reasonably small in terms of m. Finally, the bound on the oracle complexity in Proposition 3.28 is of order $\max\{1, \theta^{-2(2+a)}\}\theta_{\max} \lesssim \max\{\theta, \theta^{-(3+2a)}\}m$. Moreover, the sampling is robust in the sense that the convergence rate is proportional to $\max\{1, \theta^{-2}\}$ and the oracle complexity is proportional to $\max\{\theta, \theta^{-(3+2a)}\}$. We remark that improvements can be achieved if a coordination $\mu_{\min} \sim \epsilon^{-1}$ is possible (given a prescribed tolerance $\epsilon > 0$). For simplicity we do not present the analogous results of Proposition 3.26

for the case $m \gg 1$ under Assumption 3.8(iii). In that case, the estimates depend on $\mathrm{d}(x^0, X^*)$, the rate is proportional to $\max\{1, \theta^{-1}\}$, and the oracle complexity is proportional to $\max\{\theta, \theta^{-(2+a)}\}m$.

3.4.1. Comparison of complexity estimates. Next, we briefly compare our complexity results in terms of the quadratic natural residual, given in this section, with related results presented in previous work in terms of other merit functions for the SVI. As commented in subsection 1.1, the quadratic natural residual and the D-gap function are equivalent merit functions. An immediate result is that the previous complexity analysis, given in Theorems 3.22–3.25 and Propositions 3.23–3.28 in terms of the quadratic natural residual, are also valid in terms of the D-gap function. In this sense, our rate of convergence of $\mathcal{O}(1/K)$ in terms of the D-gap function improves the rate $\mathcal{O}(1/\sqrt{K})$ in terms of the dual gap-functions analyzed in [19, 7, 44, 45].

By Proposition 3.23 and Remark 3.24, if Assumption 3.8(ii) holds, then the algorithm performance, in terms of the convergence rate and oracle complexity, depends on some $x^* \in X^*$ such that $\sigma(x^*)^4 \cdot \max_{0 \le k \le k_0(x^*)} \mathbb{E}[\|x^k - x^*\|^2]$ is minimal, that is to say, we have a trade-off between variance of the oracle error and distance to initial iterates. We also remark that the sampling rate $\mathcal{N}_k$ possesses a *robust property*: a scaling in the sampling rate by a factor θ keeps the algorithm running with a proportional scaling of $\max\{1, \theta^{-2}\}$ in the rate and $\max\{\theta, \theta^{-3}\}$ in the oracle complexity (see [32] for discussion on robust algorithms). By Proposition 3.26, when the variance is bounded by σ^2 over X^*, the estimates depend on $\sigma^4 \max_{0 \le k \le k_0(\sigma)} \mathbb{E}[\mathrm{d}(x^k, X^*)^2]$ and $k_0(\sigma)$ is independent of any $x^* \in X^*$. When the variance is uniform over X, the estimates depend only on $\mathrm{d}(x^0, X^*)$ and a scaling factor θ in the sampling rate implies a factor of $\max\{1, \theta^{-1}\}$ in the rate and of $\max\{\theta, \theta^{-1}\}$ in the oracle complexity. In the estimates of Propositions 3.23–3.26, we may obtain optimal constants in terms of L, the variance and distance to initial iterates by optimizing over $(\hat{\alpha}, \theta)$. Interestingly, in the case of a compact feasible set, the estimates do not depend on $\mathrm{diam}(X)$, as in [19, 7], but rather on the distance of the initial iterates to X^*, which is a sharper result. In the case of networks the same conclusions hold, except that the dependence in the dimension is higher if a decentralized sampling is used. From Proposition 3.28, if a distributed sampling is used and a coordination of a rapid decreasing sequence of positive numbers is offered (in any order), then the oracle complexity depends linearly on the size of the network (up to a scaling factor in the sampling rate).

We briefly compare our convergence rate and complexity bounds presented in Propositions 3.23 and 3.26 with those in [7, Corollaries 3.2 and 3.4]. In [7], for a compact X with uniform variance over X, the convergence rate obtained in terms of the dual gap function is of order $L \, \mathrm{diam}(X)^2 K^{-1} + \sigma \, \mathrm{diam}(X) K^{-1/2}$, and the oracle complexity is of order $L \, \mathrm{diam}(X)^2 \epsilon^{-1} + \sigma^2 \, \mathrm{diam}(X)^2 \epsilon^{-2}$. For an unbounded X with uniform variance over X, the convergence rate in terms of the relaxed dual gap function $\tilde{G}(x, v)$ described in subsection 1.1 is of order $L\|x^0 - x^*\|^2 K^{-1} + \sigma\|x^0 - x^*\|^2 K^{-1/2}$, while the oracle complexity is of order $L\|x^0 - x^*\|^2 \epsilon^{-1} + \sigma^2\|x^0 - x^*\|^4 \epsilon^{-2}$. Note that the optimal constants in terms of L and σ in these bounds require tuning the stepsize to L and σ. In the estimates given in Propositions 3.23–3.26, the "coercivity" modulus ρ^{-1} introduced by the extragradient step behaves qualitatively as L. We improve on the rate of convergence to $\mathcal{O}(1/K)$ with respect to the stochastic term $\mathcal{O}(\sigma/\sqrt{K})$ by reducing iteratively the variance. Differently from [7], our analysis is the same for a compact or unbounded X, in the sense that the same merit function is used. For the case of a compact X, our bounds depend on $\mathrm{d}(x^0, X^*)$ rather than $\mathrm{diam}(X)$

as in [7], which is a sharper result. When the variance is uniform over an unbounded X, our bounds depend on $d(x^0, X^*)$ instead of $\|x^0 - x^*\|$ for a given $x^* \in X^*$ as in [7], which is also a sharper bound. We analyze the new case of nonuniform variance, which has a similar performance, except that the estimates depend on a point $x^* \in X^*$ with a minimum trade-off between variance and distances to a few initial iterates. Moreover, we include asymptotic convergence, which it is not reported in [7].

4. Concluding remarks. In this work we propose an extragradient method for pseudomonotone stochastic variational inequalities that combines the SA methodology alongside an iterative variance reduction procedure. We obtain asymptotic convergence, nonasymptotic convergence rates, and oracle complexity estimates and prove that the generated sequence is uniformly bounded in L^p. In order to achieve these properties, we require the operator to be just pseudomonotone and Lipschitz continuous. Our results give an accelerated rate with near optimal oracle complexity, coping with unbounded feasible sets and an oracle with nonuniform variance. The method admits a robust sampling rate. We also include the analysis for the distributed solution of Cartesian SVIs.

A potential direction of future research is to obtain sharp complexity estimates for *exponential convergence* of method (11)–(12). In previous works [19, 7, 13], exponential convergence is proved, assuming a *uniform* tail bound for the oracle error, that is, that there exists $\sigma > 0$ such that $\mathbb{E}[\exp\{\frac{\|F(\xi,x)-T(x)\|^2}{\sigma^2}\}] \leq \exp\{1\}$ for all $x \in X$. This assumption is not satisfied in general for unbounded feasible sets and, even for compact ones, σ^2 may be a conservative upper bound of the oracle variance at points of the trajectory of the method. Moreover, based on Section 3.4.1, in the case of a compact feasible set or uniform tail bound, we wish to study sharp complexity estimates with respect to the distance of the initial iterate to the solution set. We intend to make this analysis in a second paper assuming a nonuniform tail bound in the spirit of Assumption 3.8(i)–(ii). Motivated by this work, another interesting line of research we intend to pursue is to verify if (extra)gradient methods with *robust stepsizes* can achieve accelerated convergence rates with respect to the stochastic error.

Finally, we discuss error bounds on the solution set. It is well known that important classes of variational inequalities admit the natural residual as an error bound for the solution set, i.e., for all $\alpha > 0$, there exists $\delta > 0$ such that $d(x, X^*) \lesssim r_\alpha(x)$ for all $x \in \mathbb{R}^n$ with $r_\alpha(x) \leq \delta$. This property holds, for example, for (i) semistable VIs, (ii) composite strongly monotone VIs such that X is a polyhedron, (iii) VIs such that T is linear and X is a cone (see [11]). Item (ii) includes affine VIs and strongly monotone VIs. Item (iii) includes linear homogeneous complementarity problems and a linear system of equations. When such a property holds, the results of Theorems 3.22–3.25 and Propositions 3.23–3.28 provide other classes of SVIs for which convergence of $\mathcal{O}(1/K)$ holds in terms of the mean-squared distance to the solution set. In the previous literature, such a property was shown only for strongly pseudomonotone or weak-sharp SVIs on a compact set. In an upcoming paper, we intend to refine the complexity analysis for the case in which this error bound on the solution set is valid.

Appendix. Proof of lemmas.

Proof of Lemma 3.11. Let $x^* \in X^*$. In order to simplify the notation, in what follows we say $\widehat{F}(\epsilon_2^k, z^k) := T(z^k) + \epsilon_2^k$ and $y^k := x^k - \alpha_k \widehat{F}(\epsilon_2^k, z^k)$, so that, $x^{k+1} = \Pi(y^k)$. For every $x \in X$, we have

$$\|x^{k+1} - x\|^2 = \|\Pi(y^k) - x\|^2$$
$$\leq \|y^k - x\|^2 - \|y^k - \Pi(y^k)\|^2$$
$$= \|(x^k - x) - \alpha_k \widehat{F}(\epsilon_2^k, z^k)\|^2 - \|(x^k - x^{k+1}) - \alpha_k \widehat{F}(\epsilon_2^k, z^k)\|^2$$
$$= \|x^k - x\|^2 - \|x^k - x^{k+1}\|^2 + 2\langle x - x^{k+1}, \alpha_k \widehat{F}(\epsilon_2^k, z^k)\rangle$$
$$= \|x^k - x\|^2 - \|x^k - x^{k+1}\|^2 + 2\langle x - z^k, \alpha_k \widehat{F}(\epsilon_2^k, z^k)\rangle$$
(101)
$$+ \; 2\langle z^k - x^{k+1}, \alpha_k \widehat{F}(\epsilon_2^k, z^k)\rangle$$
$$= \|x^k - x\|^2 - \|(x^k - z^k) + (z^k - x^{k+1})\|^2$$
$$+ \; 2\langle z^k - x^{k+1}, \alpha_k \widehat{F}(\epsilon_2^k, z^k)\rangle + 2\langle x - z^k, \alpha_k \widehat{F}(\epsilon_2^k, z^k)\rangle$$
$$= \|x^k - x\|^2 - \|x^k - z^k\|^2 - \|z^k - x^{k+1}\|^2$$
$$- \; 2\langle x^k - z^k, z^k - x^{k+1}\rangle + 2\langle z^k - x^{k+1}, \alpha_k \widehat{F}(\epsilon_2^k, z^k)\rangle$$
$$+ \; 2\langle x - z^k, \alpha_k \widehat{F}(\epsilon_2^k, z^k)\rangle$$
$$= \|x^k - x\|^2 - \|x^k - z^k\|^2 - \|z^k - x^{k+1}\|^2$$
$$+ \; 2\langle x^{k+1} - z^k, x^k - \alpha_k \widehat{F}(\epsilon_2^k, z^k) - z^k\rangle + 2\langle x - z^k, \alpha_k \widehat{F}(\epsilon_2^k, z^k)\rangle,$$

using Lemma 2.1(ii) in the inequality and simple algebra in the equalities.

Looking at the fourth term $\mathsf{I} := 2\langle x^{k+1} - z^k, x^k - \alpha_k \widehat{F}(\epsilon_2^k, z^k) - z^k\rangle$ in the RHS of the last equality of (101), we take into account (15) and the fact that $\widehat{F}(\epsilon_2^k, z^k) = T(z^k) + \epsilon_2^k$, and apply Lemma 2.1(i) with $C = X$, $x = x^k - \alpha_k(T(x^k) + \epsilon_1^k)$ and $y = x^{k+1} \in X$, obtaining

$$\mathsf{I} = 2\langle x^{k+1} - z^k, x^k - \alpha_k(T(x^k) + \epsilon_1^k) - z^k\rangle$$
(102)
$$+ \; 2\langle x^{k+1} - z^k, \alpha_k \left[(T(x^k) + \epsilon_1^k) - (T(z^k) + \epsilon_2^k) \right]\rangle$$
$$\leq 2\alpha_k \|x^{k+1} - z^k\| \|(T(z^k) + \epsilon_2^k) - (T(x^k) + \epsilon_1^k)\|,$$

using the Cauchy–Schwarz inequality. Next we apply Lemma 2.1(iii) to (15)–(16), obtaining

$$\|x^{k+1} - z^k\| = \|\Pi[x^k - \alpha_k(T(z^k) + \epsilon_2^k)] - \Pi[x^k - \alpha_k(T(x^k) + \epsilon_1^k)]\|$$
(103)
$$\leq \alpha_k \|(T(z^k) + \epsilon_2^k) - (T(x^k) + \epsilon_1^k)\|.$$

Combining (102) and (103) we get

$$\mathsf{I} \leq 2\alpha_k^2 \|(T(z^k) + \epsilon_2^k) - (T(x^k) + \epsilon_1^k)\|^2$$
$$\leq 2\alpha_k^2 \left(\|T(z^k) - T(x^k)\| + \|\epsilon_1^k\| + \|\epsilon_2^k\| \right)^2$$
(104)
$$\leq 2\alpha_k^2 \left(L\|z^k - x^k\| + \|\epsilon_1^k\| + \|\epsilon_2^k\| \right)^2$$
$$\leq 6L^2 \alpha_k^2 \|z^k - x^k\|^2 + 6\alpha_k^2 \|\epsilon_1^k\|^2 + 6\alpha_k^2 \|\epsilon_2^k\|^2,$$

using the triangle inequality in the second inequality, Lipschitz continuity of T in the third inequality, and the fact that $(a + b + c)^2 \leq 3a^2 + 3b^2 + 3c^2$ in the last inequality. We set $x := x^*$ in (101). Looking now at the last term in the last equality in (101), we get

$$2\langle x^* - z^k, \alpha_k \widehat{F}(\epsilon_2^k, z^k)\rangle = 2\langle x^* - z^k, \alpha_k(T(z^k) + \epsilon_2^k)\rangle$$
(105)
$$= 2\langle x^* - z^k, \alpha_k T(z^k)\rangle + 2\langle x^* - z^k, \alpha_k \epsilon_2^k\rangle$$
$$\leq 2\langle x^* - z^k, \alpha_k \epsilon_2^k\rangle =: \mathsf{J}_k$$

using, in the last inequality, the fact that $\langle x^* - z^k, \alpha_k T(z^k) \rangle \leq 0$, which follows from Assumption 3.4, the fact that $\alpha_k > 0$, $x^* \in X^*$, and $z^k \in X$. Combining (101), (104), and (105), we get

$$
\begin{aligned}
\|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 - \|z^k - x^k\|^2 - \|z^k - x^{k+1}\|^2 \\
&\quad + 6L^2\alpha_k^2\|z^k - x^k\|^2 + 6\alpha_k^2\left(\|\epsilon_1^k\|^2 + \|\epsilon_2^k\|^2\right) + \mathsf{J}_k \\
(106) \qquad &\leq \|x^k - x^*\|^2 - \rho_k\|z^k - x^k\|^2 + 6\alpha_k^2(\|\epsilon_2^k\|^2 + \|\epsilon_1^k\|^2) + \mathsf{J}_k,
\end{aligned}
$$

using the fact that $\rho_k = 1 - 6L^2\alpha_k^2$.

Recalling that $z^k = \Pi[x^k - \alpha_k(T(x^k) + \epsilon_1^k)]$, we note that

$$
\begin{aligned}
r_{\alpha_k}(x^k)^2 &= \|x^k - \Pi[x^k - \alpha_k T(x^k)]\|^2 \\
&\leq 2\|x^k - z^k\|^2 + 2\|\Pi[x^k - \alpha_k(T(x^k) + \epsilon_1^k)] - \Pi[x^k - \alpha_k T(x^k)]\|^2 \\
(107) \qquad &\leq 2\|x^k - z^k\|^2 + 2\alpha_k^2\|\epsilon_1^k\|^2,
\end{aligned}
$$

using Lemma 2.1(iii) in the second inequality. From (106), (107), and the definitions (24)–(25) and $\mathsf{J}_k = M_{k+1}(x^*) - M_k(x^*)$, we get the claimed relation. $\qquad\square$

We now give the proof of Lemma 3.12.

Proof of Lemma 3.12. We first prove the result under Assumption 3.8(i)–(ii). Consider item 1. Assume first that $m > 1$ and take $i \in [m]$. For $1 \leq t \leq N_i$, define $U_i^t \in \mathbb{R}^{n_i}$ by

$$
U_i^t := \sum_{j=1}^t \frac{F_i(\xi_{j,i}, x) - T_i(x)}{N_i}.
$$

Defining $U_i^0 = 0$, $\mathcal{G}_0 := \sigma(U_i^0)$ and the natural filtration $\mathcal{G}_t := \sigma(\xi_{1,i}, \ldots, \xi_{t,i})$ for $1 \leq t \leq N_i$, $\{U_i^t, \mathcal{G}_t\}_{t=0}^{N_i}$ defines a vector-valued martingale (since it is a sum of N_i independent mean-zero vector random variables) whose increments satisfy

$$
\begin{aligned}
\left|\|U_i^t - U_i^{t-1}\|\right|_p &= \left|\frac{\|F_i(\xi, x) - T_i(x)\|}{N_i}\right|_p \\
&\leq \frac{\|\|F(\xi, x) - T(x)\|\|_p}{N_i} \leq \frac{\sigma(x^*)(1 + \|x - x^*\|)}{N_i},
\end{aligned}
$$

by Assumption 3.8, using the same notation $\|\cdot\|$ for the Euclidean norm in $\mathbb{R}^{n_i}$ and in $\mathbb{R}^n$. Hence,

$$
(108) \qquad \left|\|U_i^{N_i}\|\right|_p \leq \frac{C_p\,\sigma(x^*)(1 + \|x - x^*\|)}{\sqrt{N_i}},
$$

which follows from the BDG inequality (8). For each $i \in [m]$, $\epsilon_i(x) = U_i^{N_i}$. Hence, since $q \geq 1$, from Minkowski's inequality and (108) we get

$$
(109) \quad \||\epsilon(x)|\|_p^2 = \left|\|\epsilon(x)\|^2\right|_q \leq \sum_{i=1}^m \left|\|U_i^{N_i}\|^2\right|_q \leq C_p^2\left(\sum_{i=1}^m \frac{2}{N_i}\right)\sigma(x^*)^2(1 + \|x - x^*\|^2),
$$

using $(a+b)^2 \leq 2a^2 + 2b^2$. The first claim follows from (109) with $\mathsf{A} = 2$. If $m = 1$, the same proof line holds with $\mathsf{A} = 1$, since relation $(a+b)^2 \leq 2a^2 + 2b^2$ is not required.

We now prove item 2. Suppose that $m > 1$ and that $\{\xi_{j,i} : 1 \leq i \leq m, 1 \leq j \leq N_i\}$ is i.i.d. We have

(110) $$|\langle v, \epsilon(x)\rangle|_p \le \|v\| \|\|\epsilon(x)\|\|_p,$$

by the Cauchy–Schwarz inequality. The claim follows from (109) and (110) with $\mathsf{B} = 2$.

Finally, we prove item 3. Suppose that $m = 1$, or $m > 1$ with $N_i \equiv N$, $\xi_{j,i} \equiv \xi_j$ for all $i \in [m]$. Define $U^0 := 0$ and $U^t := (U_1^t, \ldots, U_m^t)$ for $t \ge 1$ and $W_t := \langle v, \cdot U^t \rangle$. Observe that $\{(W_t, \mathcal{G}_t)\}_{k=0}^N$ defines a real valued martingale with the filtration given by $\mathcal{G}_0 := \sigma(U^0)$ and $\mathcal{G}_t := \sigma(\xi_1, \ldots, \xi_t)$ for $t \ge 1$, since it is a sum of N i.i.d. random variables. Its increments $|W_t - W_{t-1}|_p$ are equal to

$$\left|\left\langle v, \frac{F(\xi_t, x) - T(x)}{N} \right\rangle\right|_p \le \frac{\|v\| \|\|F(\xi, x) - T(x)\|\|_p}{N}$$

(111) $$\le \frac{\|v\|\sigma(x^*)(1 + \|x - x^*\|)}{N},$$

using the Cauchy–Schwarz inequality in the first inequality and Assumption 3.8 in the last one. Hence, from (111) and the BDG inequality (8), we get the claim with $\mathsf{B} = 1$ (in this case $\mathcal{N} = N$).

The proof of the bounds under the stronger Assumption 3.8(iii) is essentially the same with sharper bounds on the increments, and so we omit it.   ☐

REFERENCES

[1] F. Bach and E. Moulines, *Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning*, in Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2011.

[2] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.

[3] R. S. Burachik, A. N. Iusem, and B. F. Svaiter, *Enlargement of monotone operators with applications to variational inequalities*, Set-Valued Anal., 5 (1998), pp. 159–180.

[4] D. L. Burkholder, B. Davis, and R. F. Gundy, *Integral inequalities for convex functions of operators on martingales*, in Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, vol. 2, 1972, pp. 223–240.

[5] R. H. Byrd, G. M. Chin, J. Nocedal, and Y. Wu, *Sample size selection in optimization methods for machine learning*, Math. Program., 134 (2012), pp. 127–155.

[6] X. Chen, R. J. B. Wets, and Y. Zhang, *Stochastic variational inequalities: Residual minimization smoothing sample average approximations*, SIAM J. Optim., 22 (2012), pp. 649–673.

[7] Y. Chen, G. Lan, and Y. Ouyang, *Accelerated Schemes for a Class of Variational Inequalities*, preprint, arXiv:1403.4164, 2014.

[8] G. Deng and M. C. Ferris, *Variable-number sample-path optimization*, Math. Program., 117 (2009), pp. 81–109.

[9] J. C. Duchi, P. L. Bartlett, and M. J. Wainwright, *Randomized smoothing for stochastic optimization*, SIAM J. Optim., 22 (2012), pp. 674–701.

[10] R. Durret, *Probability: Theory and Examples*, Cambridge University Press, Cambridge, UK, 2010.

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

[12] M. C. Ferris and J. S. Pang, *Engineering and economic applications of complementarity problems*, SIAM Rev., 39 (1997), pp. 669–713.

[13] M. P. Friedlander and G. Goh, *Tail Bounds for Stochastic Approximation*, preprint, arXiv:1304.5586., 2013.

[14] G. Gürkan, A. Y. Özge, and S. M. Robinson, *Sample-path solution of stochastic variational inequalities*, Math. Program., 84 (1999), pp. 313–333.

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

[16] T. Homem-de-Mello, *Variable-sample methods for stochastic optimization*, ACM Trans. Model. Comput. Simul., 13 (2003), pp. 108–133.

[17] A. Iusem, A. Jofré, and P. Thompson, *Incremental constraint projection methods for monotone stochastic variational inequalities*, submitted.

[18] H. Jiang and H. Xu, *Stochastic approximation approaches to the stochastic variational inequality problem*, IEEE Trans. Automat. Control, 53 (2008), pp. 1462–1475.

[19] A. Juditsky, A. Nemirovski, and C. Tauvel, *Solving variational inequalities with stochastic mirror-prox algorithm*, Stoch. Systems, 1 (2011), pp. 17–58.

[20] A. Kannan and U. V. Shanbhag, *Distributed computation of equilibria in monotone Nash games via iterative regularization techniques*, SIAM J. Optim., 22 (2012), pp. 1177–1205.

[21] A. Kannan and U. V. Shanbhag, *The pseudomonotone stochastic variational inequality problem: Analytical statements and stochastic extragradient schemes*, presented at the American Control Conference, Portland, OR, 2014.

[22] A. Kannan and U. V. Shanbhag, *The Pseudomonotone Stochastic Variational Inequality Problem: Analysis and Optimal Stochastic Approximation Schemes*, preprint, arXiv:1410.1628, 2014.

[23] A. J. King and R. T. Rockafellar, *Asymptotic theory for solutions in statistical estimation and stochastic programming*, Math. Oper. Res., 18 (1993), pp. 148–162.

[24] I. V. Konnov, *Equilibrium Models and Variational Inequalities*, Elsevier, New York, 2007.

[25] G. M. Korpelevich, *The extragradient method for finding saddle points and other problems*, Ekon. Mat. Metody, 12 (1976), pp. 747–756.

[26] J. Koshal, A. Nedić, and U. V. Shanbhag, *Regularized iterative stochastic approximation methods for stochastic variational inequality problems*, IEEE Trans. Automat. Control, 58 (2013), pp. 594–609.

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

[28] C. Marinelli and M. Röckner, *On the maximal inequalities of Burkholder, Davis and Gundy*, Expo. Math., 34 (2016), pp. 1–26.

[29] R. D. Monteiro and B. F. Svaiter, *On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean*, SIAM J. Optim., 20 (2010), pp. 275–287.

[30] R. D. Monteiro and B. F. Svaiter, *Complexity of variants of Tseng's modified F-B splitting and Korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems*, SIAM J. Optim., 21 (2011), pp. 1688–1720.

[31] A. Nemirovski, *Prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems*, SIAM J. Optim., 15 (2004), pp. 229–251.

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

[33] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Math. Program., 120 (2009), pp. 261–283.

[34] U. Ravat and U. V. Shanbhag, *On the Existence of Solutions to Stochastic Quasi-Variational Inequality and Complementarity Problems*, preprint, arXiv:1306.0586, 2013.

[35] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Stat., 22 (1951), pp. 400–407.

[36] H. Robbins and D. O. Siegmund, *A convergence theorem for non negative almost supermartingales and some applications*, in Optimizing Methods in Statistics (Proceedings of a Symposium at Ohio State University, Columbus, Ohio), J. S. Rustagi, ed., Academic Press, New York, 1971, pp. 233–257.

[37] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on Stochastic Programming: Modeling and Theory*, MOS-SIAM Ser. Optim. SIAM, Philadelphia, 2009.

[38] U. V. Shanbhag and J. Blanchet, *Budget constrained stochastic approximation*, in Proceedings of the Winter Simulation Conference, IEEE, 2015.

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

[40] H. Xu and D. Zhang, *Stochastic Nash equilibrium problems: Sample average approximation and applications*, Comput. Optim. Appl., 55 (2013) pp. 597–645.

[41] H. Xu, *Sample average approximation methods for a class of stochastic variational inequality problems*, Asia-Pac. J. Oper. Res., 27 (2010), pp. 103–119.

[42] H. Xu, *Uniform exponential convergence of sample average random functions under general sampling with applications in stochastic programming*, J. Math. Anal. Appl., 368 (2010), pp. 692–710.

[43] F. YOUSEFIAN, A. NEDIĆ, AND U. V. SHANBHAG, *Self-tuned stochastic approximation schemes for non-Lipschitzian stochastic multi-user optimization and Nash games*, IEEE Trans. Automat. Control, 61 (2016), pp. 1753–1766.

[44] F. YOUSEFIAN, A. NEDIĆ, AND U. V. SHANBHAG, *Optimal robust smoothing extragradient algorithms for stochastic variational inequality problems*, presented at IEEE Conference on Decision and Control, 2014.

[45] F. YOUSEFIAN, A. NEDIĆ AND U. V. SHANBHAG, *On Smoothing, Regularization and Averaging in Stochastic Approximation Methods for Stochastic Variational Inequalities*, preprint. arXiv:1411.0209.