Noname manuscript No.
(will be inserted by the editor)

Variance-based stochastic extragradient methods with linear search for stochastic variational inequalities

A. N. Iusem · A. Jofré · R. Oliveira · P. Thompson

the date of receipt and acceptance should be inserted later

Abstract We propose stochastic extragradient methods for stochastic variational inequalities with a linear search requiring only pseudo-monotonicity of the operator and no knowledge of the Lipschitz constant L. We provide convergence and complexity analysis, allowing for an unbounded feasible set, unbounded operator, non-uniform variance of the oracle and we do not require any regularization. We also prove the generated sequence is bounded in L^p. Alongside the stochastic approximation procedure, we iteratively reduce the variance of the stochastic error. Our methods cope with stepsizes bounded away from zero and attain the near-optimal oracle complexity $O(\log_{1/\theta} L)\cdot\epsilon^{-2}\cdot[\ln(\epsilon^{-1})]^{1+b}$ and an accelerated rate $O(1/K)$ in terms of the mean (quadratic) natural residual and the mean D-gap function, where K is the number of iterations required for a given tolerance $\epsilon>0$ for arbitrary $\theta\in(0,1)$ and $b>0$. Explicit estimates for the convergence rate, oracle complexity and the p-moments are given depending on problem parameters and the distance of initial iterates to the solution set.

Keywords

Mathematics Subject Classification (2000) 65K15, 90C33, 90C15, 62L20.

1 Introduction

The standard (deterministic) variational inequality problem, which we will denote as $\mathrm{VI}(T,X)$ 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 $\mathrm{VI}(T,X)$. 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 *equilibrium problems*. The complementarity problem and systems of equations are important classes of problems where the feasible set is unbounded.

A. N. Iusem
Instituto Nacional de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Rio de Janeiro, RJ, 22460-320, Brazil,
E-mail: iusp@impa.br

A. Jofré
Centro de Modelamiento Matemático & Departamento de Ingeniería Matemática, Universidad de Chile, Beauchef 851, Edificio Norte, Piso 7, Santiago, Chile,
E-mail: ajofre@dim.uchile.cl

R. I. Oliveira
Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Rio de Janeiro, RJ, 22460-320, Brazil,
E-mail: rimfo@impa.br

P. Thompson
Centro de Modelamiento Matemático, Universidad de Chile, Beauchef 851, Edificio Norte, Piso 7, Santiago, Chile &
Instituto Nacional de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Rio de Janeiro, RJ, 22460-320, Brazil,
E-mail: philipthomp@gmail.com

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 $\mathrm{VI}(T, X)$ as defined by (1), where $T : \mathbb{R}^n \to \mathbb{R}^n$ is the expected value of $F(\xi, \cdot)$, i.e.,

$$T(x) = \mathbb{E}[F(\xi, x)], \quad \forall x \in \mathbb{R}^n. \tag{2}$$

Precisely, the definition of the *stochastic variational inequality* problem (SVI) is the following:

Definition 1 (SVI) Under the setting of (2), find a random variable $x^* : \Omega \to X$, such that $\langle T(x^*(\xi)), x - x^*(\xi)\rangle \geq 0$, for all $x \in X$ and almost every $\xi \in \Xi$.

Such formulation of SVI is also called *expected value* formulation. It was first proposed in [8], as a natural generalization of stochastic optimization problems (SP). Recently, a more general definition of stochastic variational inequality was considered in [4] where the feasible set is also affected by randomness, that is, $X : \Xi \rightrightarrows \mathbb{R}^n$ is a random set-valued function. This setting appears, e.g., in economical or traffic equilibrium problems where an uncertain demand is present in the constraints.

Methods for the deterministic $\mathrm{VI}(T, X)$ have been extensively studied (see [6]). If T is fully available then SVI can be solved by these methods. As in the case of SP, the SVI in Definition 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 close 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 "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 and SVI, see e.g., [8,29] and references therein.

The SA methodology has a long tradition in probability, statistics and optimization, initiated by the seminal work of Robbins and Monro in [26]. In this paper they consider $X = \mathbb{R}^n$ and $T = \nabla f$ in Definition 1 for a smooth strongly convex function f under specific conditions. Thus, the problem they analyse is: under (2), almost surely find $x^*(\xi) \in \mathbb{R}^n$ such that $T(x^*(\xi)) = 0$. The SA methodology has been applied to SVI in [13], [14], [34], [21], [32], [10], [11], [5], [35], [16], [17], [36]. SA-typed methods for 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 an stochastic error $F(\xi, x) - T(x)$ for $x \in X$ in the trajectory of the method. See also [22], [2] for other problems where the stochastic approximation procedure is relevant (such as machine learning, online optimization, repeated games, queueing theory, signal processing and control theory).

1.1 Related work on SA

The first SA method for SVI was analyzed in [13]. Its iteration is given by:

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

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 [13], the almost sure (a.s.) convergence of $\{x^k\}$ 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 $0 < \alpha_k < 2\rho/L^2$ in case T is ρ-strong monotone) and an unbiased oracle with uniform variance, i.e., there exists $\sigma > 0$ such that for all $x \in X$,

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

After the above mentioned work, more recent research on SA methods for SVI have been developed in [14], [34], [21], [32], [10], [11], [5], [35], [16], [17], [36]. 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 [20] and the mirror-prox algorithm of Nemiroviski [23] were used in [14], [5], [35], [16], [17], [11] and iterative

Tykhonov and proximal regularization procedures (see [15], [19]), were used in [34], [21], [10], [36]. Other objectives in some of these papers were the use of incremental constraint projections in the case of difficulties accessing the feasible set [32], [10], the convergence analysis in the absence of the Lipschitz constant [34], [35], [36], and the distributed solution of Cartesian variational inequalities [34], [21], [10], [15].

1.2 **Two new extragradient methods for SVI**

We will need now some notation. We shall denote by $\epsilon(\xi, x) := F(\xi, x) - T(x)$ the oracle error at the point $x \in \mathbb{R}^n$. Given a sample $\xi^N := \{\xi_j\}_{j=1}^N$ of ξ, we denote the associated empirical average of the random operator at a point $x \in \mathbb{R}^n$ by

$$\widehat{F}\left(\xi^N, x\right) := \frac{1}{N}\sum_{j=1}^{N} F(\xi_j, x),$$

and the empirical mean error by

$$\widehat{\epsilon}(\xi^N, x) := \widehat{F}\left(\xi^N, x\right) - T(x) := \frac{1}{N}\sum_{j=1}^{N} F(\xi_j, x) - T(x).$$

In [11], the following stochastic extragradient method was proposed:

$$z^k = \Pi\left[x^k - \alpha_k \widehat{F}(\xi^k, x^k)\right], \tag{5}$$

$$x^{k+1} = \Pi\left[x^k - \alpha_k \widehat{F}(\eta^k, z^k)\right], \tag{6}$$

where $\{N_k\} \subset \mathbb{N}$ is a non-decreasing sequence (termed *sample rate*) and $\xi^k := \{\xi_j^k : k \in \mathbb{N}, j = 1, \dots, N_k\}$ and $\eta^k := \{\eta_j^k : k \in \mathbb{N}, j = 1, \dots, N_k\}$ are independent identically distributed (i.i.d.) samples of ξ. Method (5)-(6) is proved to convergence under pseudo-monotonicity of the operator[1], i.e., for all $z, x \in \mathbb{R}^n$, $\langle T(x), z - x\rangle \geq 0 \implies \langle T(z), z - x\rangle \geq 0$, Lipschitz-continuity of the random operator, i.e., for any $x, y \in \mathbb{R}^n$,

$$\|F(\xi, x) - F(\xi, y)\| \leq L(\xi)\|x - y\|, \tag{7}$$

for some measurable function $L : \Xi \to \mathbb{R}_+$ with finite variance, an oracle with finite variance over X, i.e., for all $x \in X$,

$$\mathbb{E}\left[\|F(\xi, x) - T(x)\|^2\right] < \infty, \tag{8}$$

and assuming a stepsize bounded away from zero, i.e., $0 < \inf_k \alpha_k \leq \sup_k \alpha_k < (2L)^{-1}$ where $L > 0$ is the Lipschitz constant of T and a sample rate satisfying $\sum_k (N_k)^{-1} < \infty$, which is typically satisfied by $N_k = O\left(k(\ln k)^{1+b}\right)$ for some $b > 0$.

As it will be discussed more precisely in the sequel, method (5)-(6) has important improvements with respect to previous SA methods for SVI (the main advantages being of accelerating the convergence rate with respect to the noise error and coping efficiently with unboundedness and variance of the oracle). One drawback, however, is that it requires the knowledge of the Lipschitz constant L. The main purpose of this paper is the introduction of an extragradient method with a linear search for determining the stepsizes, as was done by Khobotov [18] and by Iusem and Svaiter [12] for the deterministic case. See also [30], [33], [1] for other deterministic projection methods with linear search. Importantly, we are able to maintain the good properties of method (5)-(6) in the absence of L. The introduction of such a linear search has two goals. First, it allows the method to deal with problems where the Lipschitz constant of the operator T is inexistent, unknown, or too large, in which case the stepsizes become too small with a detrimental effect on the convergence. It also improves over the alternative of “small” exogenous stepsizes, (i.e., a summable sequence $\{\alpha_k\}$), which has also a very detrimental effect on the convergence. The intuition is that a linear search provides a procedure which uses the information available at iteration k in order to determine the largest possible value of the stepsize α_k for which the convergence properties of the algorithms can be ensured. The prototype of the linear search is the Armijo search applied to the steepest descent method for unconstrained optimization problems, adapted to the VI problem in [18] and [12]. It is widely recognized that the Armijo search substantially enhances the numerical performance of

[1] Pseudo-monotonicity is a weaker condition than monotonicity: for all $z, x \in \mathbb{R}^n$, $\langle T(z) - T(x), z - x\rangle \geq 0$.

the steepest descent method, compared with the variants which use exogenous stepsizes, be it summable ones, or dependent on the Lipschitz constant. All these nice properties make the extragradient methods with linear search we propose more implementable.

We thus propose two stochastic extragradient methods with linear search. The first variant, which we call the *stochastic hyperplane projection method* (SHP), is the stochastic variant of the hyperplane projection method proposed in [12] by Iusem and Svaiter. In such method, the linear search is based on the geometric interpretation of separating the current iterate and the solution set by a hyperplane. The SHP method takes the form: choose, $\hat{\alpha} \in (0,1]$, $\theta \in (0,1)$, $0 < \hat{\beta} \leq \tilde{\beta}$, $\{\beta_k\} \subset [\hat{\beta}, \tilde{\beta}]$, $\{N_k\} \subset \mathbb{N}$ and $\lambda > 0$; given iterate x^k, generate a sample $\xi^k := \{\xi_j^k : j = 1, \dots, N_k\}$ of ξ and take α_k as the maximum $\alpha \in \{\theta^j \hat{\alpha} : j \in \mathbb{N}_0\}$ such that

$$\left\langle \widehat{F}(\xi^k, \bar{z}^k(\alpha)), x^k - \Pi(g^k) \right\rangle \geq \frac{\lambda}{\beta_k} \|x^k - \Pi(g^k)\|^2, \tag{9}$$

where $g^k := x^k - \beta_k \widehat{F}(\xi^k, x^k)$ and for all $\alpha > 0$, $\bar{z}^k(\alpha) := \alpha \Pi(g^k) + (1-\alpha)x^k$. Then set $z^k := \bar{z}^k(\alpha_k)$ and $x^{k+1} := \Pi\left[x^k - \gamma_k \widehat{F}(\xi^k, z^k)\right]$, where

$$\gamma_k := \left\langle \widehat{F}(\xi^k, z^k), x^k - z^k \right\rangle \cdot \|\widehat{F}(\xi^k, z^k)\|^{-2}.$$

It is not difficult to see that $x^{k+1} = \Pi\left[\Pi_{H_k}(x^k)\right]$ where Π_{H_k} is Euclidean projection onto the hyperplane

$$H_k := \left\{ x \in \mathbb{R}^n : \langle \widehat{F}(\xi^k, z^k), x - z^k \rangle = 0 \right\}.$$

The second variant we propose, which we call *stochastic extragradient method with linear search* (SELS), is the stochastic variant of the method proposed by Khobotov in [18]. The SELS method takes the form: choose $\hat{\alpha} > 0$, $\theta \in (0,1)$, $\{N_k\} \subset \mathbb{N}$ and $\lambda > 0$; given iterate x^k, generate samples $\xi^k := \{\xi_j^k : j = 1, \dots, N_k\}$ and $\eta^k := \{\eta_j^k : j = 1, \dots, N_k\}$ of the random variable ξ and choose α_k as the maximum $\alpha \in \{\theta^j \hat{\alpha} : j \in \mathbb{N}_0\}$ such that

$$\alpha \left\| \widehat{F}(\xi^k, z^k(\alpha)) - \widehat{F}(\xi^k, x^k) \right\| \leq \lambda \|z^k(\alpha) - x^k\|, \tag{10}$$

where for all $\alpha > 0$, $z^k(\alpha) := \Pi\left[x^k - \alpha \widehat{F}(\xi^k, x^k)\right]$. Then set z^k and x^{k+1} as in (5)-(6) for α_k as chosen in (10).

We now introduce some additional notation required in the next subsection. For any $a > 0$ we consider the *natural residual function* r_a, defined, for any $x \in \mathbb{R}^n$, by $r_a(x) := \|x - \Pi(x - aT(x))\|$ and the *regularized gap-function* g_a, defined, for any $x \in \mathbb{R}^n$, by $g_a(x) := \sup_{y \in X}\{\langle T(x), x - y\rangle - \frac{a}{2}\|x - y\|^2\}$. For fixed $b > a > 0$, the D-*gap function* $g_{a,b}$ is defined as, for any $x \in \mathbb{R}^n$, by $g_{a,b}(x) := g_a(x) - g_b(x)$. It is known that the D-gap function and the natural residual are continuous unrestricted merit functions of $\mathrm{VI}(T, X)$, i.e., $X^* = g_{a,b}^{-1}(0) = r_a^{-1}(0)$ for any $b > a > 0$. 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 $x \in \mathbb{R}^n$, $C_1 r_{b^{-1}}(x)^2 \leq g_{a,b}(x) \leq C_2 r_{a^{-1}}(x)^2$ (see [6], Theorems 10.2.3, 10.3.3). For fixed $\alpha > 0$ and given $\epsilon > 0$, we consider an iteration index $K = K_\epsilon$, such that $\mathbb{E}[r_\alpha(x^K)^2] < \epsilon$, and we look at $\mathbb{E}[r_\alpha(x^K)^2]$ as a non-asymptotic convergence rate. In particular, we will have an $O(1/K)$ convergence rate if $\mathbb{E}[r_\alpha(x^K)^2] \leq Q/K$ for some constant $Q > 0$. The *oracle complexity* will be defined as the total number of oracle calls needed for $\mathbb{E}[r_\alpha(x^K)^2] < \epsilon$ to hold. As an example, for method (5)-(6), the oracle complexity is $\sum_{k=1}^K 2N_k$.

1.3 Comparison with previous works

To the best of our knowledge, methods SHP and SELS are the first extragradient methods with linear search for SVIs. We remark that, as will be presented in this paper, the SELS method maintains the good properties of the extragradient method (5)-(6) up to a factor of $O(\log_{1/\theta} L)$ in the oracle complexity and the number of projections per iteration, since it does not require knowledge of the Lipschitz constant L. Importantly, $O(\log_{1/\theta} L)$ is a small factor for practical purposes (see Remark 4). In summary, SELS and SHP have the following characteristics:

i) SELS and SHP require only (7) and pseudo-monotonicity of T without any regularization.

ii) SELS requires a stepsize bounded away from zero and hence has an accelerated rate of $O(1/K)$ in terms of the mean quadratic natural residual or the mean D-gap function with a near-optimal oracle complexity of $O\left(\log_{1/\theta} L\right) \cdot \epsilon^{-2} \cdot \left[\ln(\epsilon^{-1})\right]^{1+b}$ for any $b > 0$ and $2 \cdot O\left(\log_{1/\theta} L\right)$ projections per iteration (see Proposition 7 and Remark 4),

iii) SELS and SHP only require a stochastic oracle with a non-uniform variance, i.e., satisfying (8). Importantly, the performance of SELS and SHP depends on a *minimal trade-off* between the variance on X^* and the distance of initial iterates to X^*, i.e., the performance depends on the factor

$$\widehat{\mathsf{Q}} := \inf_{x^* \in X^*} \left\{ \sigma(x^*) \cdot \max_{0 \leq k \leq k_0(x^*)} \mathbb{E}[\|x^k - x^*\|^2] \right\}, \tag{11}$$

where, for $x^* \in X^*$, $k_0(x^*) \in \mathbb{N}$ is explicitly determined (see Propositions 5 and 7). This result also improves in the case in which (4) *does* holds but $\sigma(x^*)^2 \ll \sigma^2$ or in the case in which X is compact but $\max_{0 \leq k \leq k_0(x^*)} \mathbb{E}[\|x^k - x^*\|^2] \ll \operatorname{diam}(X)^2$. In this sense, SELS depends only on the variance of the oracle error at points of the *solution set* and *the trajectory of the method.*[2]

iv) SELS and SHP allow an unbounded feasible set or operator, keeping asymptotic convergence of the generated sequence and results of items ii)-iii). We also prove that the generated sequence is bounded in L^p.

v) Under conditions of items i)-iv), the sample rate $\{N_k\}$ is *robust* in the sense that a scaling factor $\Theta > 0$ on the sampling rate maintains the progress of the algorithm with proportional scaling in the convergence rate and oracle complexity (see Propositions 5 and 7. See also [24] for robust methods).

Before method (5)-(6) and SELS, previous works required: (i) specific classes of pseudo-monotonicity, bounded monotone operators or a regularization procedure (which requires additional coordination of parameters and a sub-optimal rate), (ii) small stepsizes, i.e. satisfying $\sum_k \alpha_k^2 < \infty$, with a slower convergence rate of $O(1/\sqrt{K})$ and oracle complexity $O(\epsilon^{-2})$ in terms of a mean gap-function [3] for bounded monotone operators, (iii) for the general monotone case without regularization, previous methods required a uniform variance, i.e., satisfying (4), which is much more demanding than (8) and excludes, e.g., affine monotone stochastic complementarity problems, (iv) for unbounded X or T, asymptotic convergence was obtained under demanding monotonicity properties or regularization, and convergence rates were only given for strongly monotone operators or for monotone operators requiring a uniform variance with the slower rate $O(1/\sqrt{K})$ in terms of a mean gap-function [4] and with no guaranteed asymptotic convergence (see [11], Example 1).

It should be noticed that methods which avoid the use of the Lipschitz constant or Lipschitz continuity, were proposed in [35,36], but by means of a very different procedure. Instead of linear searches they use a random smoothing technique by means of sampling an auxiliary random variable. It is an interesting idea, but it requires compactness of the feasible set, uniformly bounded variance of the oracle for monotone operators, and achieves the slower rate $O(1/\sqrt{K})$, while we can cope with unbounded sets, non-uniform variance for pseudo-monotone operators and achieve the rate $O(1/K)$.

We make some final comments on the results of methods SHP and SELS. In the deterministic case, the hyperplane projection method in [12] requires only continuity. The SHP method requires Hölder continuity of the random operator in order to control the variance of the oracle error. This variant also uses two projections per iteration with convergence rate $O(1/\sqrt{K})$, sample rate $N_k \sim k^2$ (up to logarithm terms) and oracle complexity $O(\epsilon^{-6})$ (up to logarithm terms). The SELS method requires Lipschitz continuity of the random operator, $2 \cdot O(\log_{1/\theta} L)$ projections per iteration, sample rate $N_k \sim k$ (up to logarithmic terms) and oracle complexity of $O(\log_{1/\theta} L) \cdot \epsilon^{-2}$ (up to logarithmic terms). Hence, the choice between these two linear search variants depends on a trade-off between computational and oracle complexity (which might depend on the application of interest). If oracle complexity is expensive, our results tend to suggest SELS (if Lipschitz continuity is available).

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 proposed algorithms and their convergence analysis. In Subsection 3.1 the assumptions required for the analysis are discussed. Subsection 3.2 presents

[2] A typical example includes monotone linear SVIs where $F(\xi, x) = A(\xi)x$ for some random matrix $A(\xi)$ such that $A = \mathbb{E}[A(\xi)]$ is a semi-definite positive matrix. In this case, we have $\mathbb{V}[\epsilon(\xi, x)] = O(\|x\|^2)$ so that (4) is not satisfied for unbounded X. Note that for a compact X such that $0 \in X$, (4) does hold but $0 \in X^*$ and $\mathbb{V}[\epsilon(\xi, 0)] = 0$. For such case, the performance of methods in [14] and [5] depends on $\sigma^2 > 0$ which is very conservative compared to (11).

[3] Precisely, the *dual-gap function* defined as $G(x) := \sup_{y \in X} \langle T(y), x - y \rangle$.

[4] Precisely, the relaxed dual gap-function of Monteiro-Svaiter defined as $\tilde{G}(x, v) := \sup_{y \in X} \langle T(y) - v, x - y \rangle$ for $x \in X$ and $v \in \mathbb{R}^n$.

the convergence analysis while Subsection 3.3 focus on convergence rates and complexity results. We present the analysis of SPH and SELS separately. The Appendix provide proofs of essential lemmas.

2 Preliminaries

2.1 **Projection operator and notation**

For $x, y \in \mathbb{R}^n$, we denote by $\langle x, y\rangle$ the standard inner product, and by $\|x\| = \sqrt{\langle x, x\rangle}$ 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 $\Pi_C(x) := \mathrm{argmin}_{y\in C} \|y - x\|^2$ for $x \in \mathbb{R}^n$. Given $H : \mathbb{R}^n \to \mathbb{R}^n$, $\mathrm{S}(H, C)$ denotes the solution set of $\mathrm{VI}(H, C)$. The following properties of the projection operator are well known.

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

i) Given $x \in \mathbb{R}^n$, $\Pi_C(x)$ is the unique point of C satisfying the 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 $H : \mathbb{R}^n \to \mathbb{R}^n$, $\mathrm{S}(H, C) = \{x \in \mathbb{R}^n : x = \Pi_C[x - H(x)]\}$.
v) For all $x \in C, y \in \mathbb{R}^n$, $\langle x - y, x - \Pi_C(y)\rangle \geq \|x - \Pi_C(y)\|^2$.

In the case of the feasible set X as in (1), we shall use the notation $\Pi := \Pi_X$. Given an operator $H : \mathbb{R}^n \to \mathbb{R}^n$, for any $x \in \mathbb{R}^n$ and $\alpha > 0$, we denote the natural residual function associated to $\mathrm{VI}(H, X)$ by

$$r_\alpha(H; x) := \|x - \Pi\left[x - \alpha H(x)\right]\|.$$

In the case of the operator T as in (1), we use the notation $r_\alpha := r_\alpha(T, \cdot)$. For the unit stepsize $\alpha = 1$, we use the notation $r(H; \cdot) := r_1(H; \cdot)$ and $r := r_1$.

We shall also use the following useful lemma (see [6], Proposition 10.3.6).

Lemma 2 *Given $x \in \mathbb{R}^n$, the function $(0, \infty) \ni \alpha \mapsto \frac{r_\alpha(H,x)}{\alpha}$ is non-increasing.*

We use the abbreviation “RHS” for “right hand side”. Given sequences $\{x^k\}$ and $\{y^k\}$, we use the notation $x^k = O_p(y^k)$ or $\|x^k\| \lesssim_p \|y^k\|$ to mean that there exists a constant $C_p > 0$ (depending only on p) such that $\|x^k\| \leq C_p\|y^k\|$ for all k (we omit the reference to p if no confusion arises or if there is no such dependence). 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. 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 $|\xi\,|\mathcal{F}|_p := \sqrt[p]{\mathbb{E}\left[|\xi|^p\,|\mathcal{F}\right]}$ is the L_p-norm of ξ conditional to the σ-algebra $\mathcal{F}$. Given $x \in \mathbb{R}$, we denote $\lceil x\rceil$ the smallest integer greater than x. For a matrix $B \in \mathbb{R}^{n\times n}$, B^T denotes its transpose, $\|B\|$ denotes its spectral norm and $\mathrm{tr}(B)$ denotes its trace. For $m \in \mathbb{N}$, we use the notation $[m] = \{1, \ldots, m\}$.

2.2 **Probabilistic tools**

As in other stochastic approximation methods, a fundamental tool to be used is the following Convergence Theorem of Robbins and Siegmund [27], which can be seen as the stochastic version of the properties of quasi-Fejér convergent sequences.

Theorem 1 *Let $\{y_k\}, \{u_k\}, \{a_k\}, \{b_k\}$ be sequences of non-negative random variables, adapted to the filtration $\{\mathcal{F}_k\}$, such that a.s. $\sum a_k < \infty$, $\sum b_k < \infty$ and for all $k \in \mathbb{N}$, $\mathbb{E}\big[y_{k+1}\big|\mathcal{F}_k\big] \leq (1+a_k)y_k - u_k + b_k$. 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 will require the following result, proved in [3]:

Theorem 2 *Let $X_j := (X_{j,t})_{t\in\mathcal{T}}$ for $j \in [N]$ denote independent random vectors indexed by $\mathcal{T}$ such that $\mathbb{E}[X_{j,t}] = 0$ for all $t \in \mathcal{T}$ and $j \in [N]$. Let*

$$\begin{aligned} Z &:= \sup_{t\in\mathcal{T}} \left| \sum_{j=1}^{N} X_{j,t} \right|, \\ M &:= \max_{j\in[N]} \sup_{t\in\mathcal{T}} |X_{j,t}|, \\ \widehat{\sigma}^2 &:= \sup_{t\in\mathcal{T}} \mathbb{E}\left[\sum_{j=1}^{N} X_{j,t}^2 \right]. \end{aligned}$$

Then, there exists constant $\kappa > 0$, such that

$$|Z|_p \le 2\mathbb{E}[Z] + 2\sqrt{2\kappa p}\widehat{\sigma} + 20\kappa p |M|_p + 4\sqrt{\kappa p}|M|_2.$$

A random vector Y taking values in $\mathbb{R}^n$ is called *sub-Gaussian* with variance parameter $\sigma^2 > 0$ if for all $v \in \mathbb{R}^n$,

$$\mathbb{E}\left[e^{\langle v,Y\rangle} \right] \le e^{\frac{\sigma^2\|v\|^2}{2}}.$$

We will use the following result, proved in [9]:

Theorem 3 (Quadratic forms of sub-Gaussian vectors) *Let $A \in \mathbb{R}^{n\times n}$ be a matrix, and let $S := A^T A$. Suppose that Y is a zero-mean sub-Gaussian random vector with variance parameter σ^2 taking values in $\mathbb{R}^n$. Then for all $0 \le s < \frac{1}{2\sigma^2\|S\|}$,*

$$\mathbb{E}\left[\exp\left\{ s\|AY\|^2 \right\} \right] \le \exp\left\{ \sigma^2 \operatorname{tr}(S)s + \frac{\sigma^4 \operatorname{tr}(S^2)s^2}{1 - 2\sigma^2\|S\|s} \right\}.$$

We shall also use the following characterization of sub-Gaussian random variables found in Theorem 2.1 of [3].

Theorem 4 *Let Y a random variable taking values in $\mathbb{R}$ with $\mathbb{E}[Y] = 0$. If for some $\sigma^2 > 0$, for all $s > 0$,*

$$\mathbb{P}(|Y| \ge s) \le 2\exp\left\{ -\frac{s^2}{2\sigma^2} \right\}$$

then Y is a sub-Gaussian random variable with variance parameter $4\sigma^2$.

A random variable Y taking values in $\mathbb{R}$ is called *sub-Gamma* with parameters $\sigma^2 > 0$ and $c > 0$ if for all $0 < s < \frac{1}{c}$,

$$\mathbb{E}\left[e^{sY} \right] \le e^{\frac{\sigma^2 s^2}{2(1-cs)}}.$$

We will need the following result established in Corollary 2.6 of [3].

Lemma 3 *Let $\{Z_i\}_{i=1}^N$ be real-valued sub-Gamma random variables with parameters $\sigma^2 > 0$ and $c > 0$. Then*

$$\mathbb{E}\left[\max_{i=1,\dots,N} Z_i \right] \le \sqrt{2\sigma^2 \ln N} + c\ln N.$$

We will need the following result implied by Theorem 1 of [25].

Theorem 5 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(\Xi, \mathcal{G})$ be a measurable space, $\{\xi : \Omega \to \Xi\}_{j=1}^N$ be independent random variables and a measurable function $f : \Xi \to \mathbb{R}$. Let also $\{\eta_j\}_{j=1}^N$ be a collection of independent random variables which are independent of $\{\xi_j\}_{j=1}^N$ and $\mathcal{H}_N := \sigma(\xi_j : j \in [N])$. Set*

$$V := \mathbb{E}\left[\sum_{j=1}^{N} (f(\xi_j)) - f(\eta_j))^2 \middle| \mathcal{H}_N \right].$$

Then there exists constant $C > 0$ such that for all $\lambda > 0$,

$$\mathbb{P}\left\{ \left| \sum_{j=1}^{N} f(\xi_j) \right| \ge C\sqrt{V(1+\lambda)} \right\} \le e^{-\lambda}.$$

Finally, we will use the following result in Lemma 13.11 of [3]. We give first the definition of metric entropy of a pseudo-metric space. A subset $\mathcal{V} \subset \mathcal{T}$ of a countable pseudo-metric space $(\mathcal{T}, d)$ is called a δ-net for $\mathcal{T}$ if for every $t \in \mathcal{T}$, there exists $v \in \mathcal{V}$ such that $d(t,v) \leq \delta$. If, additionally, $\mathcal{V}$ is finite, with cardinality $N(\delta,\mathcal{T})$ of minimum size, then the δ-entropy number is $H(\delta,\mathcal{T}) := \ln N(\delta,\mathcal{T})$.

Lemma 4 *We consider $\mathbb{R}^n$ with the ℓ_q metric and denote by B_q the correspondent unit ball. For all $q \geq 1$ and all $u \in (0,1]$, the metric entropy $H(u,B_q)$ satisfies*

$$H(u,B_q) \leq n \ln\left(1+\frac{1}{u}\right).$$

For further details on the probabilistic tools mentioned above, we refer the reader to the book [3].

3 Stochastic extragradient methods with linear search

In this section we present formally the two extragradient methods with linear search we propose. We start with the *stochastic hyperplane projection method.* The idea is to adapt the deterministic hyperplane projection method by replacing the mean operator with the empirical average of the random operator associated to a progressively increasing sample rate $\{N_k\}$.

Algorithm 1 (The stochastic hyperplane projection method)

1. **Initialization:** Choose the initial iterate $x^0 \in \mathbb{R}^n$, parameters $\tilde{\beta} \geq \hat{\beta} > 0$, $\lambda \in (0,1)$, $\hat{\alpha} \in (0,1]$ and $\theta \in (0,1)$, the step sequence $\{\beta_k\} \subset [\hat{\beta},\tilde{\beta}]$, the sample rate $\{N_k\}$ and initial samples $\xi^0 := \{\xi_j^0\}_{j=1}^{N_0}$ of the random variable ξ.
2. **Iterative step:** Given x^k, generate samples $\xi^k := \{\xi_j^k\}_{j=1}^{N_k}$ of ξ.
   If $x^k = \Pi\left[x^k - \beta_k \widehat{F}(\xi^k, x^k)\right]$ stop. Otherwise:
   Linear search rule: Find the maximum $\alpha \in \{\theta^j \hat{\alpha} : j \in \mathbb{N}_0\}$ such that

$$\left\langle \widehat{F}\left(\xi^k, \bar{z}^k(\alpha)\right), x^k - \Pi(g^k)\right\rangle \geq \frac{\lambda}{\beta_k}\|x^k - \Pi(g^k)\|^2, \tag{12}$$

   where $g^k := x^k - \beta_k \widehat{F}(\xi^k, x^k)$ and for all $\alpha > 0$, $\bar{z}^k(\alpha) := \alpha\Pi(g^k) + (1-\alpha)x^k$.
   Denoting by $\alpha_k > 0$ the above maximum value, set

$$z^k := \bar{z}^k(\alpha_k) \; = \; \alpha_k \Pi\left[x^k - \beta_k \widehat{F}(\xi^k, x^k)\right] + (1-\alpha_k)x^k, \tag{13}$$

$$x^{k+1} := \Pi\left[x^k - \gamma_k \widehat{F}(\xi_k, z^k)\right], \tag{14}$$

   where $\gamma_k := \left\langle \widehat{F}(\xi^k, z^k), x^k - z^k\right\rangle \cdot \|\widehat{F}(\xi^k, z^k)\|^{-2}$.

Set $y^k := x^k - \gamma_k \widehat{F}(\xi_k, z^k)$. We remark that, as in the deterministic hyperplane projection method of Iusem-Svaiter [12], x^{k+1} is the projection of x^k onto the hyperplane $H_k := \{x \in \mathbb{R}^n : \langle \widehat{F}(\xi^k, z^k), x - z^k\rangle = 0\}$, or alternatively onto the halfspace $L_k := \{x \in \mathbb{R}^n : \langle \widehat{F}(\xi^k, z^k), x - z^k\rangle \leq 0\}$. In the deterministic case, the monotonicity of the operator implies a crucial fact used in the convergence analysis: if the method does not stop in finitely many iterations then $x^k \notin L^k$ and H^k *strictly separates the solution set X^* from the iterate x^k*, which entails a strict Fejér relation. In Algorithm 1, we still have $x^k \notin L^k$, but the separation property is no longer valid, since a solution $x^* \in X^*$ may fail to belong to L^k if the angle $\langle \epsilon(\xi^k, z^k), x^* - z^k\rangle$ is positive. Nevertheless, a recursive relation can be obtained in order to control this infeasibility of the solution to L^k in terms of $\langle \epsilon(\xi^k, z^k), x^* - z^k\rangle$ (see Lemma 8).

Concerning Algorithm 1, we define the oracle errors:

$$\bar{\epsilon}_1^k := \widehat{F}(\xi^k, x^k) - T(x^k), \tag{15}$$

$$\bar{\epsilon}_2^k := \widehat{F}(\xi^k, z^k) - T(z^k), \tag{16}$$

$$\bar{\epsilon}_3^k := \widehat{F}(\xi^k, \widehat{z}^k) - T(z^k), \tag{17}$$

where $\widehat{z}^k := \bar{z}^k(\theta\alpha_k)$ (see linear search (12) for the definition of $\bar{z}^k(\alpha)$).

We now present the *stochastic extragradient with linear search* which differs from Algorithm 1 by the use of a different linear search.

Algorithm 2 (The stochastic extragradient method with linear search)

1. **Initialization:** Choose the initial iterate $x^0 \in \mathbb{R}^n$, parameters $\hat{\alpha} > 0$, $\lambda \in (0, 1/\sqrt{6})$ and $\theta \in (0,1)$, the sample rate $\{N_k\}$ and initial samples $\xi^0 := \{\xi_j^0\}_{j=1}^{N_0}$ and $\eta^0 := \{\eta_j^0\}_{j=1}^{N_0}$ of the random variable ξ.
2. **Iterative step:** Given iterate x^k, generate samples $\xi^k := \{\xi_j^k\}_{j=1}^{N_k}$ and $\eta^k := \{\eta_j^k\}_{j=1}^{N_k}$ of ξ.
   If $x^k = \Pi\left[x^k - \hat{\alpha}\widehat{F}(\xi^k, x^k)\right]$ stop. Otherwise:
   Linear search rule: define α_k as the maximum $\alpha \in \{\theta^j \hat{\alpha} : j \in \mathbb{N}_0\}$ such that
   $$\alpha \left\| \widehat{F}\left(\xi^k, z^k(\alpha)\right) - \widehat{F}\left(\xi^k, x^k\right) \right\| \leq \lambda \|z^k(\alpha) - x^k\|, \tag{18}$$
   where $z^k(\alpha) := \Pi\left[x^k - \alpha\widehat{F}(\xi^k, x^k)\right]$ for all $\alpha > 0$. Set
   $$z^k = \Pi\left[x^k - \alpha_k\widehat{F}(\xi^k, x^k)\right], \tag{19}$$
   $$x^{k+1} = \Pi\left[x^k - \alpha_k\widehat{F}(\eta^k, z^k)\right]. \tag{20}$$

Note that if T is Lipschitz continuous with constant L, Algorithm 2 recovers Algorithm (5)-(6) of [11] if we set $0 < \inf_k \alpha_k \leq \sup_k \alpha_k = \hat{\alpha} < 1/2L$ (i.e., the linear search rule (18) is satisfied in the first iteration with $\alpha_k := \hat{\alpha}$).

Concerning Algorithm 2, we define the oracle errors:

$$\epsilon_1^k := \widehat{F}(\xi^k, x^k) - T(x^k), \tag{21}$$
$$\epsilon_2^k := \widehat{F}(\eta^k, z^k) - T(z^k), \tag{22}$$
$$\epsilon_3^k := \widehat{F}(\xi^k, z^k) - T(z^k). \tag{23}$$

3.1 Discussion of the assumptions

Concerning Algorithm 1, we shall study the the stochastic process $\{x^k\}$ with respect to the filtration

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

Concerning Algorithm 2, we shall study the stochastic process $\{x^k\}$ with respect to the filtrations

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

A significant difference between Algorithms 1, 2 and Algorithm (5)-(6) in [11] is that the *adaptative* stepsize α_k obtained in the linear search is a random variable which *depends on the sample* ξ^k. The inevitable consequence is that the errors $\{\bar{\epsilon}_2^k, \bar{\epsilon}_3^k\}$ in Algorithm 1 and ϵ_3^k in Algorithm 2 do *not* induce martingales. This complicates considerably the convergence analysis requiring other statistical tools (see Lemma 6).

We state next the assumptions needed for the convergence analysis of these algorithms.

Assumption 1 (Consistency) *The solution set* $X^* := \mathrm{S}(T, X)$ *is non-empty.*

Assumption 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,* [5] $\xi : \Omega \to \Xi$ *is a random variable defined on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *and* $\mathbb{E}[\|F(\xi, x)\|] < \infty$ *for all* $x \in X$.

Remark 1 We observe that by Example 14.29 of [28], if F is a Carathéodory map and $\xi^N = \{\xi_j\}_{j=1}^N$ is a sample of ξ, then $(\omega, x) \mapsto \|\widehat{\epsilon}(\xi^N(\omega), x)\|$ is a *normal integrand*, that is,

$$\omega \mapsto \operatorname{epi} \|\widehat{\epsilon}(\xi^N(\omega), \cdot)\| := \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} : \|F(\xi, \alpha)\| \leq \alpha\}$$

is a measurable function. Hence, for any $x^* \in \mathbb{R}^n$ and measurable positive function $R : \Omega \to \mathbb{R}_+$, by Theorem 14.37 in [28], we have that

$$\omega \mapsto \sup_{x \in B[x^*, R(\omega)]} \|\widehat{\epsilon}(\xi^N(\omega), x)\|$$

is a measurable function.

[5] 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.

Assumption 3 (Pseudo-monotonicity) *We assume that $T:\mathbb{R}^n\to\mathbb{R}^n$ as defined in (2) is pseudo-monotone, i.e., for all $z,x\in\mathbb{R}^n$, $\langle T(x),z-x\rangle\geq 0\Longrightarrow\langle T(z),z-x\rangle\geq 0$.*

Assumption 4 (Hölder-continuity) *There exist $\delta\in(0,1]$ and measurable function $L:\Xi\to\mathbb{R}_+$ with finite $2p$-moment for some $p\geq 2$ such that for all $x,y\in X$,*

$$\|F(\xi,x)-F(\xi,y)\|\leq L(\xi)\|x-y\|^{\delta}.$$

We will not require knowledge of L or δ in our methods. With respect to Algorithm 2, we shall always assume that $\delta=1$ in Assumption 4, that is, Lipschitz-continuity. With respect to Algorithm 1, we perform the asymptotic analysis for any $\delta\in(0,1]$, while in the convergence rate estimates and the complexity analysis we assume that $\delta=1$.

Assumption 5 (Sample rate) *In Algorithms 1 and 2, $\{\xi_j\}_{j=1}^{N_k}$ and $\{\eta_j\}_{j=1}^{N_k}$ are i.i.d. samples of ξ. In the case of Algorithm 2, $\{\xi_j\}_{j=1}^{N_k}$ and $\{\eta_j\}_{j=1}^{N_k}$ are independent of each other.*

Moreover, concerning the sample rate $\{N_k\}$ we assume:

i) In Algorithm 1, $\sum_{k=0}^{\infty}\frac{1}{\sqrt{N_k}}<\infty$.
ii) In Algorithm 2, $\sum_{k=0}^{\infty}\frac{1}{N_k}<\infty$.

Assumption 6 (Variance control) *There exists* [6] *$p\geq 2$ and a locally bounded and measurable function $\sigma:X^*\to\mathbb{R}_+$ such that for all $x^*\in X^*$, $x\in X$,*

$$|\|F(\xi,x)-T(x)\||_{2p}\leq\sigma(x^*)\left(1+\|x-x^*\|\right).$$

Denote $q:=p/2$. We remark that Assumption 6 is implied by the non-uniform variance (8) over X^* and the Hölder-continuity (4).

3.2 Convergence analysis

In this section, we provide, in two subsections, the convergence analysis of Algorithms 1 and 2. We first state two lemmas whose proofs can be found in the Appendix.

The following lemma is Lemma 4 in [11], applied to the oracle error $\bar{\epsilon}_1^k$ in Algorithm 1 and to the oracle errors $\{\epsilon_1^k,\epsilon_2^k\}$ in Algorithm 2.

Lemma 5 *Consider Assumptions 1-6. Let $\xi^N:=\{\xi_j:j\in[N]\}$ be an i.i.d. sample of ξ.*

Then for any $\mathfrak{p}\in[p,2p]$ and for all $x^*\in X^*$, $x\in X,v\in\mathbb{R}^n$,

$$\left|\left\|\widehat{\epsilon}\left(\xi^N,x\right)\right\|\right|_{\mathfrak{p}}\leq C_{\mathfrak{p}}\sqrt{\frac{n}{N}}\sigma(x^*)(1+\|x-x^*\|),$$
$$\left|\langle v,\widehat{\epsilon}\left(\xi^N,x\right)\rangle\right|_{\mathfrak{p}}\leq C_{\mathfrak{p}}\frac{\|v\|}{\sqrt{N}}\sigma(x^*)(1+\|x-x^*\|).$$

As commented above, errors $\{\bar{\epsilon}_2^k,\bar{\epsilon}_3^k\}$ in Algorithm 1 and ϵ_3^k in Algorithm 2 do not induce martingales. For this reason, in order to bound such stochastic errors we will need some additional statistical theory. A difficult additional feature is that X may be unbounded. This is the subject of the following lemma, proved in the Appendix.

Lemma 6 *Suppose that Assumptions 1-6 hold. Let $\xi^N:=\{\xi_j:j\in[N]\}$ be an i.i.d. sample of ξ and let $\mathcal{F}_N$ be a σ-algebra independent of ξ^N. Given $\beta>0$, a random vector $x^N\in\mathcal{F}_N$ in X, and a random variable $\alpha_N\in(0,1]$, set*

$$z^N:=\alpha_N\Pi\left[x^N-\beta\widehat{F}(\xi^N,x^N)\right]+(1-\alpha_N)x^N.$$

Then there exists $C_{p,L}>0$ (depending on $L(\xi)$ and p), such that for any $x^*\in X^*$,

$$\left|\left\|\widehat{\epsilon}\left(\xi^N,z^N\right)\right\|\,\middle|\mathcal{F}_N\right|_p\leq C_{L,p}\sqrt{\frac{n}{N}}\sigma(x^*)(\|x^N-x^*\|+1).$$

[6] In Assumptions 4 and 6, we ask the Lipschitz modulus $L(\xi)$ and the oracle error $\epsilon(\xi,x)$ to have finite $2p$-moments for some $p\geq 2$. This is slightly more than asking finite p-moment for some $p\geq 2$ as in Algorithm (5)-(6) in [11]. This is a technical assumption for facilitating the analysis, and sufficient for practical purposes. For instance, if $L(\xi)$ is bounded we could ask the oracle error to have finite p-moment. Moreover, assuming that $L(\xi)$ has all finite moments, we could ask the oracle error to have finite p-moment for an arbitrary $p>2$.

Finally, we remark that although the proofs of Lemmas 5 and 6 are given for the Lipschitz-continuous case in Assumption 4, this entails no loss of generality since, for $\delta \in (0,1)$, we always have the bound

$$\|x - x^*\|^\delta \leq \max\{1, \|x - x^*\|\} \leq 1 + \|x - x^*\|,$$

for any $x \in \mathbb{R}^n$ and $x^* \in X^*$.

3.2.1 ***The stochastic hyperplane projection method***

We now present the convergence analysis of Algorithm 1. We start by showing the linear search (12) in Algorithm 1 is well defined.

Lemma 7 (Good definition of the linear search) *Consider Assumption 2. Then*

i) The linear search (12) *in Algorithm 1 terminates after a finite number of iterations.*
ii) If the method does not stop at iteration $k+1$, then $\langle \widehat{F}(\xi^k, z^k), x^k - z^k\rangle > 0$. In particular, $\gamma_k > 0$ in (14).

Proof Item (ii) is a direct consequence of (i). We prove next item (i). Assume by contradiction that for every $j \in \mathbb{N}_0$,

$$\left\langle \beta_k \widehat{F}\Big(\xi^k, z^k\left(\theta^{-j}\widehat{\alpha}\right)\Big), x^k - \Pi(g^k)\right\rangle < \lambda\|x^k - \Pi(g^k)\|^2.$$

We let $j \to \infty$ above and by continuity of $\widehat{F}(\xi^k, \cdot)$, resulting from Assumption 2, we obtain

$$\lambda\|x^k - \Pi(g^k)\|^2 \geq \langle x^k - g^k, x^k - \Pi(g^k)\rangle \geq \|x^k - \Pi(g^k)\|^2,$$

using Lemma 1(v) in the last inequality. Since we have $x^k \neq \Pi(g^k)$ by the definition of the method, we obtain that $\lambda \geq 1$, a contradiction.

The following Lemma is also proved in the Appendix.

Lemma 8 *Consider Assumptions 1-3. Suppose that the method does not stop at iteration $k+1$. Then, for all $x^* \in X^*$,*

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \|y^k - x^k\|^2 + 2\gamma_k\langle \bar{\epsilon}_2^k, x^* - z^k\rangle.$$

We now aim at controlling the error term $\gamma_k\langle \bar{\epsilon}_2^k, x - z^k\rangle$, which perturbs the hyperplane separation property. This term is not a martingale, since z^k depends on ξ^k. Hence, differently than Algorithm 2, we must analyze the first order moments of stochastic errors (see Assumption 5(i)). We shall need the following simple lemma for establishing a useful endogenous bound for γ_k, in terms of the deterministic bounded stepsize β_k.

Lemma 9 *Suppose that the method does not stop at iteration $k+1$. Then*

$$0 < \gamma_k < \frac{\alpha_k\beta_k}{\lambda} \leq \frac{\widehat{\alpha}\beta_k}{\lambda}. \tag{24}$$

Proof We only need to prove the second inequality. The linear search (12) and the fact that $x^k - z^k = \alpha_k(x^k - \Pi(g^k))$ imply that

$$\langle \widehat{F}(\xi^k, z^k), x^k - z^k\rangle \geq \frac{\lambda}{\alpha_k\beta_k}\|x^k - z^k\|^2. \tag{25}$$

From (25) and the definition of γ_k we get

$$\gamma_k = \frac{\langle \widehat{F}(\xi^k, z^k), x^k - z^k\rangle}{\|\widehat{F}(\xi^k, z^k)\|^2} > \frac{\lambda}{\alpha_k\beta_k}\frac{\|x^k - z^k\|^2}{\|\widehat{F}(\xi^k, z^k)\|^2}, \tag{26}$$

while the definition of γ_k gives

$$\gamma_k = \frac{\langle \widehat{F}(\xi^k, z^k), x^k - z^k\rangle}{\|\widehat{F}(\xi^k, z^k)\|^2} \leq \frac{\|\widehat{F}(\xi^k, z^k)\|\|x^k - z^k\|}{\|\widehat{F}(\xi^k, z^k)\|^2} = \frac{\|x^k - z^k\|}{\|\widehat{F}(\xi^k, z^k)\|}, \tag{27}$$

using the Cauchy-Schwartz inequality. Inequalities (26)-(27) imply the claim.

For $x^* \in X^*$, $k \in \mathbb{N}_0$, we define

$$\widehat{\mathsf{H}}_k(x^*) := \max\{C_p, C_{p,L}\}\,\frac{\tilde\beta}{\lambda}\sqrt{\frac{n}{N_k}}\sigma(x^*),$$

where C_p and $C_{p,L}$ are the constants defined in Lemmas 5 and 6.

Lemma 10 (Error decay) *Consider Assumptions 1-6. Suppose that Algorithm 1 does not stop at iteration $k+1$. Then for all $x^* \in X^*$,*

$$\left|\gamma_k\langle\bar\epsilon_2^k, x^*-z^k\rangle\middle|\mathcal{F}_k\right|_q \lesssim \tilde\beta L\widehat{\mathsf{H}}_k(x^*)\left(1+\|x^k-x^*\|^2\right) + \widehat{\mathsf{H}}_k(x^*)^2\left(1+\|x^k-x^*\|^2\right).$$

Proof We denote $\widetilde{z}^k := \Pi(g^k)$, so that

$$x^*-z^k = \alpha_k(x^*-\widetilde{z}^k) + (1-\alpha_k)(x^*-x^k), \tag{28}$$

using the fact that $x^* = \alpha_k x^* + (1-\alpha_k)x^*$. In view of (28), we have

$$\begin{aligned}\gamma_k\langle\bar\epsilon_2^k, x^*-z^k\rangle &= \gamma_k\alpha_k\langle\epsilon_2^k, x^*-\widetilde{z}^k\rangle + \gamma_k(1-\alpha_k)\langle\bar\epsilon_2^k, x^*-x^k\rangle \\ &\leq \frac{\tilde\beta}{\lambda}\|\bar\epsilon_2^k\|\left(\|x^*-\widetilde{z}^k\| + \|x^*-x^k\|\right),\end{aligned} \tag{29}$$

using the Cauchy-Schwarz inequality, Lemma 9, and the facts that $0<\alpha_k\leq\hat\alpha\leq 1$ and $0<\beta_k\leq\tilde\beta$.

Since $x^*\in X^*$, by Lemma 1(iv), we use the fact that $x = \Pi[x-\beta_k T(x)]$ and the definitions of $\tilde{z}^k$, g^k and $\bar\epsilon_1^k$ in order to obtain

$$\begin{aligned}\|\tilde z^k - x^*\| &= \|\Pi[x^k-\beta_k(T(x^k)+\bar\epsilon_1^k)] - \Pi[x^*-\beta_k T(x^*)]\| \\ &\leq \|x^k - x^* + \beta_k(T(x^*)-T(x^k)) - \beta_k\bar\epsilon_1^k\| \\ &\leq \|x^k-x^*\| + \tilde\beta L\|x^k-x^*\|^\delta + \tilde\beta\|\bar\epsilon_1^k\|,\end{aligned} \tag{30}$$

using Lemma 1(iii) in the first inequality, and the fact that $0<\beta_k\leq\tilde\beta$ together with Assumption 4 in the last inequality.

Using (29)-(30) and the fact that $\|x^k-x^*\|^\delta \leq 1+\|x^k-x^*\|$, we consider $|\cdot|\mathcal{F}_k|_q$ and get

$$\begin{aligned}\left|\gamma_k\langle\bar\epsilon_2^k, x^*-z^k\rangle\middle|\mathcal{F}_k\right|_q \leq{}& \left[\tilde\beta L + (2+\tilde\beta L)\|x^k-x^*\|\right]\frac{\tilde\beta}{\lambda}\left|\|\bar\epsilon_2^k\|\middle|\mathcal{F}_k\right|_q \\ &+\frac{\tilde\beta^2}{\lambda}\left|\|\bar\epsilon_1^k\|\|\bar\epsilon_2^k\|\middle|\mathcal{F}_k\right|_q,\end{aligned} \tag{31}$$

using the fact that $x^k\in\mathcal{F}_k$. By Lemma 5, $x^k\in\mathcal{F}_k$ and the independence between ξ^k and $\mathcal{F}_k$, we get

$$\left|\|\bar\epsilon_1^k\|\middle|\mathcal{F}_k\right|_p \leq C_p\sqrt{\frac{n}{N_k}}\sigma(x^*)(\|x^k-x_*\|+1). \tag{32}$$

By Lemma 6, the definitions of z^k, $x^k\in\mathcal{F}_k$ and the fact that $\alpha_k\in(0,1]$, we get

$$\left|\|\bar\epsilon_2^k\|\middle|\mathcal{F}_k\right|_p \leq C_{L,p}\sqrt{\frac{n}{N_k}}\sigma(x^*)(\|x^k-x_*\|+1). \tag{33}$$

Invoking Hölder's inequality, we also get

$$\left|\|\bar\epsilon_1^k\|\|\bar\epsilon_2^k\|\middle|\mathcal{F}_k\right|_q \leq \left|\|\epsilon_1^k\|\middle|\mathcal{F}_k\right|_p\cdot\left|\|\epsilon_2^k\|\middle|\mathcal{F}_k\right|_p \tag{34}$$

Relations (31)-(34), the definition of $\widehat{\mathsf{H}}_k(x^*)$ and the convexity of $t\mapsto t^2$ entail the claim.

Consider $\widehat{\mathsf{H}}_k(x^*)$ as in Lemma 10. Define $\widehat{\mathsf{C}}(x^*) := \sqrt{N_k}\widehat{\mathsf{H}}_k(x^*)$.

Proposition 1 (Stochastic quasi-Fejér property) *Consider Assumptions 1-6 and assume that Algorithm 1 generates an infinite sequence $\{x^k\}$. Then*

(i) There exists $\widehat{c}\geq 1$ such that, for all $k\in\mathbb{N}$ and $x^*\in X^*$,

$$\begin{aligned}\mathbb{E}\left[\|x^{k+1}-x^*\|^2\middle|\mathcal{F}_k\right] \leq{}& \|x^k-x^*\|^2 - \mathbb{E}\left[\|y^k-x^k\|^2\middle|\mathcal{F}_k\right] \\ &+\widehat{c}\left[\frac{\tilde\beta L\widehat{\mathsf{C}}(x^*)}{\sqrt{N_k}} + \frac{\widehat{\mathsf{C}}(x^*)^2}{N_k}\right]\left(1+\|x^k-x^*\|^2\right).\end{aligned}$$

(ii) A.s. $\{\|x^k - x^*\|\}$ *and* $\{\mathrm{d}(x^k, X^*)\}$ *converge for all* $x^* \in X^*$*. In particular,* $\{x^k\}$ *is a.s.-bounded.*
(iii) A.s. if a cluster point of $\{x^k\}$ *belongs to* X^* *then* $\lim_{k\to\infty} \mathrm{d}(x^k, X^*) = 0$.

Proof i) It is an immediate consequence of Lemmas 8, 10 and the fact that $x^k \in \mathcal{F}_k$, after taking $\mathbb{E}[\cdot|\mathcal{F}_k]$ in Lemma 8.

ii) Set

$$\mathsf{c}_k(x^*) := \widehat{c}\left[\frac{\tilde{\beta}L\widehat{\mathsf{C}}(x^*)}{\sqrt{N}_k} + \frac{\widehat{\mathsf{C}}(x^*)^2}{N_k}\right].$$

From (i), for all $k \in \mathbb{N}_0$,

$$\mathbb{E}\big[\|x^{k+1} - x^*\|^2\big|\mathcal{F}_k\big] \le [1 + \mathsf{c}_k(x^*)]\,\|x^k - x^*\|^2 + \mathsf{c}_k(x^*). \tag{35}$$

By Assumption 5(i), we have $\sum_k \mathsf{c}_k(x^*) < \infty$. Hence, from (35) and Theorem 1 we conclude that a.s. $\{\|x^k - x^*\|\}$ converges and, in particular, $\{x^k\}$ is bounded.

Set $\bar{x}^k := \Pi_{X^*}(x^k)$. Relation (35) and the fact that $x^k \in \mathcal{F}_k$ imply

$$\mathbb{E}\big[\,\mathrm{d}(x^{k+1}, X^*)^2\big|\mathcal{F}_k\big] \le \big[1 + \mathsf{c}_k(\bar{x}^k)\big]\,\mathrm{d}(x^k, X^*)^2 + \mathsf{c}_k(\bar{x}^k). \tag{36}$$

The boundedness of $\{\bar{x}^k\}$ and Assumption 5(i) imply that $\sum_k \mathsf{c}_k(\bar{x}^k) < \infty$ a.s. Hence, Theorem 1 and (36) imply that $\{\mathrm{d}(x^k, X^*)\}$ a.s.-converges.

iii) Suppose that there exists $\bar{x} \in X^*$ and a subsequence $\{k_\ell\}$ such that $\lim_{\ell\to\infty} \|x^{k_\ell} - \bar{x}\| = 0$ a.s. Clearly, $\mathrm{d}(x^k, X^*) \le \|x^{k_\ell} - \bar{x}\|$ almost surely, and therefore it follows that $\lim_{\ell\to\infty} \mathrm{d}(x^{k_\ell}, X^*) = 0$. By (ii), $\{\mathrm{d}(x^k, X^*)\}$ a.s.-converges and hence $\lim_{k\to\infty} \mathrm{d}(x^k, X^*) = 0$.

We now prove asymptotic convergence of Algorithm 1. Consider $\widehat{\mathsf{C}}(x^*)$ as defined in Proposition 1.

Theorem 6 (Asymptotic convergence) *Under Assumptions 1-6, either Algorithm 1 stops at iteration* $k+1$*, in which case* x^k *is a solution of VI(*T, X*), or it generates an infinite sequence* $\{x^k\}$ *that a.s. is bounded and such that* $\lim_{k\to\infty} \mathrm{d}(x^k, X^*) = 0$*. In particular, a.s. every cluster point of* $\{x^k\}$ *belongs to* X^**.*

Proof If Algorithm 1 stops at iteration k, then $x^k = \Pi[x^k - \beta_k \widehat{F}(\xi^k, x^k)]$. From this fact and Lemma 1(iv) we get, for all $x \in X$,

$$\langle \widehat{F}(\xi^k, x^k), x - x^k\rangle \ge 0. \tag{37}$$

From the fact that $x^k \in \mathcal{F}_k$ and the independence of ξ^k and $\mathcal{F}_k$, we get $\mathbb{E}[\widehat{F}(\xi^k, x^k)|\mathcal{F}_k] = T(x^k)$. Using this equality and the fact that $x^k \in \mathcal{F}_k$, we consider $\mathbb{E}[\cdot|\mathcal{F}_k]$ in (37) and obtain, $\langle T(x^k), x - x^k\rangle \ge 0$ for all $x \in X$. Hence $x^k \in X^*$.

We now suppose that the sequence $\{x^k\}$ is infinite. By Proposition 1(iii), it is sufficient to show that a.s. the bounded sequence $\{x^k\}$ has a cluster point in X^*.

Choose any $x^* \in X^*$. As in Proposition 1, set

$$\mathsf{c}_k(x^*) := \widehat{c}\left[\frac{\tilde{\beta}L\widehat{\mathsf{C}}(x^*)}{\sqrt{N}_k} + \frac{\widehat{\mathsf{C}}(x^*)^2}{N_k}\right].$$

Using the property that $\mathbb{E}[\mathbb{E}[\cdot|\mathcal{F}_k]] = \mathbb{E}[\cdot]$, we take the expectation in Proposition 1(i), and get, for all $k \in \mathbb{N}_0$,

$$\mathbb{E}\big[\|x^{k+1} - x^*\|^2\big|\mathcal{F}_k\big] \le [1 + \mathsf{c}_k(x^*)]\,\mathbb{E}\left[\|x^k - x^*\|^2\right] - \mathbb{E}\left[\|y^k - x^k\|^2\right] + \mathsf{c}_k(x^*). \tag{38}$$

From the fact that $\sum_k \mathsf{c}_k(x^*) < \infty$, (38) and Theorem 1 we conclude that

$$\sum_{k=0}^{\infty} \mathbb{E}\left[\|y^k - x^k\|^2\right] < \infty, \tag{39}$$

and that $\left\{\mathbb{E}\left[\|x^k - x^*\|^2\right]\right\}$ converges. In particular, $\left\{\mathbb{E}\left[\|x^k - x^*\|^2\right]\right\}$ is a bounded sequence.

By the definition of Algorithm 1, we have that $\|y^k - x^k\|^2 = \langle T(z^k) + \bar{\epsilon}_2^k, x^k - z^k\rangle^2 \|T(z^k) + \bar{\epsilon}_2^k\|^{-2}$. Hence, from (39) we get

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{\langle T(z^k) + \bar{\epsilon}_2^k, x^k - z^k\rangle^2}{\|T(z^k) + \bar{\epsilon}_2^k\|^2}\right] = 0. \tag{40}$$

From the definitions of $\{\bar\epsilon_1^k, \bar\epsilon_2^k, \bar\epsilon_3^k\}$, Lemmas 5 and 6, the property that $\mathbb{E}[\mathbb{E}[\cdot|\mathcal{F}_k]] = \mathbb{E}[\cdot]$ and the boundedness of $\left\{\mathbb{E}\left[\|x^k - x^*\|^2\right]\right\}$, we get

$$\mathbb{E}\left[\|\bar\epsilon_s^k\|^2\right] \lesssim \frac{\sup_{k\in\mathbb{N}_0} \mathbb{E}\left[\|x^k - x^*\|^2\right] + 1}{\sqrt{N_k}}$$

for $s \in \{1,2,3\}$ and all $k \in \mathbb{N}_0$. Since $\lim_{k\to\infty} \sqrt{N_k} = 0$ (Assumption 5(i)), we have in particular that, for $s \in \{1,2,3\}$,

$$\lim_{k\to\infty} \mathbb{E}[\|\bar\epsilon_s^k\|^2] = 0. \tag{41}$$

Since L^2-convergence implies a.s.-convergence along a subsequence, from (40)-(41), we may take a (deterministic) subsequence $\{k_\ell\}_{\ell=1}^\infty$ such that a.s. for $s \in \{1,2,3\}$,

$$\lim_{\ell\to\infty} \frac{\alpha_{k_\ell}\langle T(z^{k_\ell}) + \bar\epsilon_2^{k_\ell}, x^{k_\ell} - \Pi(g^{k_\ell})\rangle}{\|T(z^{k_\ell}) + \bar\epsilon_2^{k_\ell}\|} = 0, \tag{42}$$

$$\lim_{\ell\to\infty} \bar\epsilon_s^{k_\ell} = 0, \tag{43}$$

using the fact that $x^k - z^k = \alpha_k[x^k - \Pi(g^k)]$. Since $\beta_k \in [\hat\beta, \tilde\beta]$ with $\hat\beta > 0$, we may refine $\{k_\ell\}$ if necessary so that, for some $\beta > 0$,

$$\lim_{\ell\to\infty} \beta_{k_\ell} = \beta. \tag{44}$$

From Proposition 1(ii), the a.s.-boundedness of the sequence $\{x^{k_\ell}\}$ implies that, on a set Ω_1 of total probability, there exists a (random) subsequence $\mathfrak{N} \subset \{k_\ell\}_{\ell=1}^\infty$ such that

$$\lim_{k\in\mathfrak{N}} x^k = x^*, \tag{45}$$

for some (random) $x^* \in \mathbb{R}^n$. Using the fact that $g^k = x^k - \beta_k[T(x^k) + \bar\epsilon_1^k]$, (43)-(45) and the continuity of T and Π, for the event Ω_1, we have

$$g^* := \lim_{k\in\mathfrak{N}} g^k = x^* - \beta T(x^*). \tag{46}$$

Also, for the event Ω_1, from the definition of z^k in (13), the fact that $\alpha_k \in (0,1]$, (43) and (45)-(46), we get that $\{T(z^k) + \bar\epsilon_2^k\}_{k\in\mathfrak{N}}$ is bounded so that, since (42), we obtain

$$\lim_{k\in\mathfrak{N}} \alpha_k\langle T(z^k) + \bar\epsilon_2^k, x^k - \Pi(g^k)\rangle = 0. \tag{47}$$

We now consider two cases for the event Ω_1.

Case (i): $\lim_{k\in\mathfrak{N}} \alpha_k \neq 0$. In this case, we may refine $\mathfrak{N}$ if necessary, and find some (random) $\bar\alpha > 0$ such that $\alpha_k \geq \bar\alpha$ for all $k \in \mathfrak{N}$. It follows from (47) that on Ω_1,

$$\lim_{k\in\mathfrak{N}} \langle T(z^k) + \bar\epsilon_2^k, x^k - \Pi(g^k)\rangle = 0. \tag{48}$$

From (12)-(13), we get

$$\langle T(z^k) + \bar\epsilon_2^k, x^k - \Pi(g^k)\rangle \geq \frac{\tau}{\beta_k}\|x^k - \Pi(g^k)\|^2 \geq \frac{\tau}{\tilde\beta}\|x^k - \Pi(g^k)\|^2 \tag{49}$$

for all k. Relations (48)-(49) imply that, for Ω_1,

$$0 = \lim_{k\in\mathfrak{N}} \|x^k - \Pi(g^k)\|. \tag{50}$$

From (45)-(46), we take limits in (50) and obtain

$$0 = \|x^* - \Pi[x^* - \beta T(x^*)]\|.$$

Therefore, $x^* = \Pi[x^* - \beta T(x^*)]$, so that $x^* \in X^*$ by Lemma 1(iv).

Case (ii): $\lim_{k\in\mathfrak{N}} \alpha_k = 0$. In this case we have

$$\lim_{k\in\mathfrak{N}} \theta\alpha_k = 0. \tag{51}$$

Since $\widehat{z}^k := \theta\alpha_k \Pi(g^k) + (1-\theta\alpha_k)x^k$ and $\{g^k\}_{k\in\mathfrak{N}}$ is bounded, we get from (51)

$$\lim_{k\in\mathfrak{N}} \hat{z}^k = x^*. \tag{52}$$

Observe that, by the definition of the linear search rule (12), we have

$$\langle T(\widehat{z}^k) + \bar{\epsilon}_3^k, x^k - \Pi(g^k)\rangle < \frac{\tau}{\beta_k}\|x^k - \Pi(g^k)\|^2 \tag{53}$$

for all $k \in \mathbb{N}_0$. We take limit in (53) along $\mathfrak{N}$, and we get, using the continuity of T and Π and relations (43)-(46) and (52) that

$$\langle T(x^*), x^* - \Pi(g^*)\rangle \leq \frac{\tau}{\beta}\|x^* - \Pi(g^*)\|^2. \tag{54}$$

Since the sequence $\{x^k\}$ is feasible and X is closed, the limit point x^* belongs to X. Thus, from (54) and Lemma 1(v), we get that, for Ω_1,

$$\tau\|x^* - \Pi(g^*)\|^2 \geq \beta\langle T(x^*), x^* - \Pi(g^*)\rangle = \langle x^* - g^*, x^* - \Pi(g^*)\rangle \geq \|x^* - \Pi(g^*)\|^2. \tag{55}$$

Since $\tau \in (0,1)$, (55) implies that $\|x^* - \Pi(g^*)\| = 0$. Hence, in view of (46), we have $x^* = \Pi(x^* - \beta T(x^*))$. By Lemma 1(iv), we conclude that $x^* \in X^*$.

We have proved that in the event Ω_1 of total probability, both in case (i) and in case (ii), $\{x^k\}$ has a cluster point which solves VI(T,X). The claim follows from Proposition 1(iii).

3.2.2 ***The stochastic extragradient method with linear search***

We now present the convergence analysis of Algorithm 2. We first show that the linear search (18) in Algorithm 2 is well defined.

Lemma 11 (Good definition of the linear search) *Consider Assumption 2. The linear search* (18) *terminates after a finite number of iterations.*

Proof Set $\gamma_j := \theta^{-j}\hat{\alpha}$ and $H_k := \widehat{F}(\xi^k,\cdot)$. Assuming by contradiction that the linear search (18) does not terminate after a finite number of iterations, for every $j \in \mathbb{N}_0$,

$$\left\|\widehat{F}\left(\xi^k, z^k(\gamma_j)\right) - \widehat{F}\left(\xi^k, x^k\right)\right\| > \lambda\frac{r_{\gamma_j}(H_k; x^k)}{\gamma_j} \geq \lambda \cdot r(H_k; x^k), \tag{56}$$

using the fact that $\gamma_j \in (0,1]$ and Lemma 2 in the last inequality. The contradiction follows by letting $j \to \infty$ in (56) and invoking the continuity of $\widehat{F}(\xi^k,\cdot)$, resulting from Assumption 2, the fact that $\lim_{j\to\infty} z^k(\gamma_j) = x^k$, which follows from the continuity of Π, and the fact that $r(H_k; x^k) > 0$, which follows from the definition of Algorithm 2.

Define recursively, for $k \in \mathbb{N}_0$, $A_0 := 0$,

$$A_{k+1} := A_k + (1-6\lambda^2)\hat{\alpha}^2\|\epsilon_1^k\|^2 + 6\hat{\alpha}^2\|\epsilon_2^k\|^2 + 6\hat{\alpha}^2\|\epsilon_3^k\|^2, \tag{57}$$

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

$$M_{k+1}(x^*) := M_k(x^*) + 2\alpha_k\langle x^* - z^k, \epsilon_2^k\rangle. \tag{58}$$

Using the notation $\Delta M_k(x^*) := M_{k+1}(x^*) - M_k(x^*)$ and $\Delta A_k := A_{k+1} - A_k$, the following recursive relation is proved in the Appendix. It generalizes Lemma 3 of [11] to the method with the linear search (18).

Lemma 12 *Consider Assumptions 1-3. If the method does not stop at iteration* $k{+}1$ *then for all* $x^* \in X^*$,

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \left(\frac{1-6\lambda^2}{2}\right) r_{\alpha_k}(x^k)^2 + \Delta M_k(x^*) + \Delta A_k.$$

We now need upper bounds on the moments of $\mathit{\Delta} A_k$ and $\mathit{\Delta} M_k(x^*)$ in terms of $\|x^k-x^*\|^2$ for any $x^*\in X^*$. With respect to $\mathit{\Delta} M_k(x^*)$ in (58), a minor difference is that the adaptive stepsize α_k is a random variable which depends on previous filtration information. Nevertheless, $\{\mathit{\Delta} M_k(x^*),\widehat{\mathcal{F}}_k\}$ is still a martingale, since $\alpha_k,z^k\in\widehat{\mathcal{F}}_k$ and η^k is independent of $\widehat{\mathcal{F}}_k$. With respect to $\mathit{\Delta} A_k$, a major difference with Algorithm (5)-(6) in [11] is the presence of the error ϵ_3^k in definition (57). As commented before, in order to bound this stochastic error we use in Lemma 6 statistical techniques different from than martingale methods, used in Lemma 5. By Lemma 6, we have

$$\left|\|\epsilon_3^k\|\middle|\mathcal{F}_k\right|_p\le C_{L,p}\sqrt{\frac{n}{N_k}}\sigma(x^*)(\|x^k-x^*\|+1). \tag{59}$$

Using (59), the proof of the following proposition follows a similar proofline as that of Proposition 2 in [11], so that we state it without proof.

Define

$$\mathsf{H}_k(x^*):=C_p\hat{\alpha}\sqrt{\frac{n}{N_k}}\sigma(x^*),\qquad \widetilde{\mathsf{H}}_k(x^*):=\max\{C_p,C_{p,L}\}\,\hat{\alpha}\sqrt{\frac{n}{N_k}}\sigma(x^*).$$

Proposition 2 (Bounds on increments) *Consider Assumptions 1-6. If the method does not stop at iteration $k+1$ then for all $x^*\in X^*$,*

$$|A_{k+1}-A_k|\mathcal{F}_k|_q\le\left[24\left(1+L\hat{\alpha}+\mathsf{H}_k(x^*)\right)^2+12+2(1-6\lambda^2)\right]\widetilde{\mathsf{H}}_k(x^*)^2\|x^k-x^*\|^2$$

$$+\left[24\mathsf{H}_k(x^*)^2+24+2(1-6\lambda^2)\right]\widetilde{\mathsf{H}}_k(x^*)^2,$$

$$|M_{k+1}(x^*)-M_k(x^*)|\mathcal{F}_k|_q\le\frac{\mathsf{H}_k(x^*)}{\sqrt{n}}\left[1+L\hat{\alpha}+\mathsf{H}_k(x^*)\right]^2\|x^k-x^*\|^2$$

$$+\frac{\mathsf{H}_k(x^*)}{\sqrt{n}}\left[1+L\hat{\alpha}+(3+2L\hat{\alpha})\mathsf{H}_k(x^*)+2\mathsf{H}_k(x^*)^2\right]\|x^k-x^*\|$$

$$+\frac{\mathsf{H}_k(x^*)}{\sqrt{n}}\left[\mathsf{H}_k(x^*)+\mathsf{H}_k(x^*)^2\right].$$

In the following proposition we summarize Lemma 12 and Proposition 2. The proof is omitted since it follows a proofline similar to that of Proposition 3 in [11]. Note that $\alpha_k\notin\mathcal{F}_k$. We define:

$$\mathsf{C}_k(x^*):=\max\{C_p,C_{p,L}\}\,\hat{\alpha}\sqrt{n}\sigma(x^*)\left\{24\left[1+L\hat{\alpha}+\mathsf{H}_k(x^*)\right]^2+14\right\}.$$

Proposition 3 (Stochastic quasi-Fejér property) *Consider Assumptions 1-6. If the method does not stop at iteration $k+1$, then for all $x^*\in X^*$,*

$$\mathbb{E}\left[\|x^{k+1}-x^*\|^2|\mathcal{F}_k\right]\le\|x^k-x^*\|^2-\frac{(1-6\lambda^2)}{2}\mathbb{E}\left[r_{\alpha_k}(x^k)^2|\mathcal{F}_k\right]+\mathsf{C}_k(x^*)\frac{\|x^k-x^*\|^2+1}{N_k}.$$

We now proceed to establish the asymptotic convergence of Algorithm 2. We shall need the following lemma:

Lemma 13 (Lower bound on stepsize) *Consider Assumptions 2 and 4 with $\delta=1$. Define $\widetilde{L}_k:=\frac{1}{N_k}\sum_{j=1}^{N_k}L(\xi_j^k)$. If the method does not stop at iteration $k+1$, then*

$$\lambda\theta\le\alpha_k\widetilde{L}_k. \tag{60}$$

Moreover, $\lambda\theta\le\left|\alpha_k\middle|\mathcal{F}_k\right|_2\cdot|L(\xi)|_2$.

Proof By the definition of the linear search (18), we have

$$\theta^{-1}\alpha_k\left\|\widehat{F}\left(\xi^k,z^k(\theta^{-1}\alpha_k)\right)-\widehat{F}(\xi^k,x^k)\right\|>\lambda\left\|z^k\left(\theta^{-1}\alpha_k\right)-x^k\right\|. \tag{61}$$

The inequality in (60) follows easily from (61) and the inequality

$$\left\|\widehat{F}\left(\xi^k,z^k(\theta^{-1}\alpha_k)\right)-\widehat{F}(\xi^k,x^k)\right\|\le\widetilde{L}_k\left\|z^k\left(\theta^{-1}\alpha_k\right)-x^k\right\|,$$

which results from Assumption 4 and the fact that $z^k\left(\theta^{-1}\alpha_k\right) \neq x^k$. For the second statement we take $\mathbb{E}[\cdot|\mathcal{F}_k]$ and get

$$
\begin{aligned}
\lambda\theta &\leq \mathbb{E}\left[\alpha_k \widetilde{L}_k \middle| \mathcal{F}_k\right] \\
&\leq \left|\alpha_k|\mathcal{F}_k\right|_2 \cdot \left|\widetilde{L}_k\middle|\mathcal{F}_k\right|_2 \\
&= \left|\alpha_k|\mathcal{F}_k\right|_2 \sqrt{\mathbb{E}\left[\left(\frac{1}{N_k}\sum_{j=1}^{N_k} L(\xi_j^k)\right)^2 \middle|\mathcal{F}_k\right]} \\
&\leq \left|\alpha_k|\mathcal{F}_k\right|_2 \sqrt{\frac{1}{N_k}\sum_{j=1}^{N_k}\mathbb{E}\left[L(\xi_j^k)^2\middle|\mathcal{F}_k\right]} = \left|\alpha_k|\mathcal{F}_k\right|_2 \cdot |L(\xi)|_2,
\end{aligned}
$$

using Hölder's inequality in the second inequality, the convexity of $t \mapsto t^2$ in the third inequality and the fact that $\xi^k = \{\xi_j^k\}_{j=1}^{N_k}$ is an i.i.d sample of ξ independent of $\mathcal{F}_k$ in the last equality.

Theorem 7 (Asymptotic convergence) *Under Assumptions 1-6, either Algorithm 2 stops at iteration $k+1$, in which case x^k is a solution of $\mathrm{VI}(T,X)$, or it generates an infinite sequence $\{x^k\}$ such that a.s. it is bounded, $\lim_{k\to\infty} \mathrm{d}(x^k, X^*) = 0$, and $r(x^k)$ converges to 0 almost surely and in L^2. In particular, a.s. every cluster point of $\{x^k\}$ belongs to X^*.*

Proof If Algorithm 2 stops at iteration k, then $x^k = \Pi[x^k - \hat{\alpha}\widehat{F}(\xi^k, x^k)]$. From this fact and Lemma 1(iv) we get, for all $x \in X$,

$$
\langle \widehat{F}(\xi^k, x^k), x - x^k\rangle \geq 0. \tag{62}
$$

From the fact that $x^k \in \mathcal{F}_k$ and the independence of ξ^k and $\mathcal{F}_k$, we have that $\mathbb{E}[\widehat{F}(\xi^k, x^k)|\mathcal{F}_k] = T(x^k)$. Using this result and the fact that $x^k \in \mathcal{F}_k$, we take $\mathbb{E}[\cdot|\mathcal{F}_k]$ in (62) and obtain, for all $x \in X$, $\langle T(x^k), x - x^k\rangle \geq 0$. Hence $x^k \in X^*$.

Suppose now that Algorithm 2 generates an infinite sequence. Take some $x^* \in X^*$. The result in Proposition 3 may be rewritten as: for all $k \in \mathbb{N}_0$,

$$
\mathbb{E}\left[\|x^{k+1} - x^*\|^2|\mathcal{F}_k\right] \leq \left(1 + \frac{\mathsf{C}(x^*)}{N_k}\right)\|x^k - x^*\|^2 - \frac{(1-6\lambda^2)}{2}\mathbb{E}\left[\alpha_k^2|\mathcal{F}_k\right] r(x^k)^2 + \frac{\mathsf{C}(x^*)}{N_k}, \tag{63}
$$

using the facts $x^k \in \mathcal{F}_k$ and $r_{\alpha_k}(x^k) \geq \alpha_k r(x^k)$ which follows from Lemma 2 and the fact that $\alpha_k \in (0,1]$.

Taking into account Assumption 5(ii), i.e., $\sum_k N_k^{-1} < \infty$, (63) and the fact that $x^k \in \mathcal{F}_k$, we apply Theorem 1 with $y_k := \|x^k - x^*\|^2$, $a_k = b_k = \mathsf{C}(x^*)/N_k$ and $u_k := (1-6\lambda^2)\mathbb{E}\left[\alpha_k^2|\mathcal{F}_k\right] r(x^k)^2/2$, in order to conclude that a.s. $\{\|x^k - x^*\|^2\}$ converges and

$$
\frac{\lambda\theta}{|L(\xi)|_2^2}\sum_{k=0}^{\infty} r(x^k)^2 \leq \sum_{k=0}^{\infty}\mathbb{E}\left[\alpha_k^2|\mathcal{F}_k\right] r(x^k)^2 < \infty, \tag{64}
$$

using Lemma 13. In particular, $\{x^k\}$ is a.s.-bounded. From (64), we get that a.s.

$$
0 = \lim_{k\to\infty} r(x^k)^2 = \lim_{k\to\infty}\left\|x^k - \Pi\left[x^k - T(x^k)\right]\right\|^2. \tag{65}
$$

The fact that $\lim_{k\to\infty}\mathbb{E}[r(x^k)^2] = 0$ is proved in a similar way, using first the fact that $\mathbb{E}\left[\alpha_k^2|\mathcal{F}_k\right] \geq \frac{\lambda\theta}{|L(\xi)|_2^2}$ and taking then the total expectation in (63).

Relation (65) and the continuity of T (Assumption 4) and Π (Lemma 1(iii)) imply that a.s. every cluster point $\bar{x}$ of $\{x^k\}$ satisfies

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

From Lemmas 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 $\lim_{k\to\infty}\mathrm{d}(x^k, X^*) = 0$ as claimed.

3.3 Convergence rate and complexity analysis

In this section we focus on the derivation of the rate of convergence and the oracle complexity in terms of the mean-square natural residual for Algorithms 1 and 2. We perform analysis for Algorithm 1 and subsequently for Algorithm 2 in separate subsections.

3.3.1 ***The stochastic hyperplane projection method***

We start by giving explicit bounds on the p-norm of the sequence generated by Algorithm 1, invoking the definitions in Lemma 10 and Proposition 1.

Proposition 4 (Uniform boundedness in L^p) *Let Assumptions 1-6 hold. Assume that Algorithm 1 generates an infinite sequence. Then, for all $x^* \in X^*$ and $\phi \in (0, \frac{\sqrt{5}-1}{2})$, given $k_0 := k_0(x^*) \in \mathbb{N}_0$ such that*

$$\sum_{k \geq k_0} \frac{1}{\sqrt{N_k}} \leq \frac{\phi}{\widehat{c}\max\{1, \widetilde{\beta}L\}\widehat{\mathsf{C}}(x^*)}, \tag{66}$$

the following estimate holds:

$$\sup_{k \geq k_0} \left|\|x^k - x^*\|\right|_p^2 \leq \frac{1 + \left|\|x^{k_0} - x^*\|\right|_p^2}{1 - \phi - \phi^2}.$$

Proof Fix $x^* \in X^*$. Set $d_k := \|x^k - x^*\|$. From Lemmas 8 and 10 and the fact that $x^k \in \mathcal{F}_k$, we can take $|\cdot|\mathcal{F}_k|_q$ in Lemma 8 and obtain

$$\left|d_{k+1}^2 | \mathcal{F}_k\right|_q \leq d_k^2 + \widehat{c}\left[\frac{\widetilde{\beta}L\widehat{\mathsf{C}}(x^*)}{\sqrt{N_k}} + \frac{\widehat{\mathsf{C}}(x^*)^2}{N_k}\right]\left(1 + d_k^2\right) \tag{67}$$

for all $k \in \mathbb{N}_0$ and $x^* \in X^*$. Using the fact that $\left||\cdot|\mathcal{F}_k|_q\right|_q = |\cdot|_q$, we take $|\cdot|_q$ in (67) and sum from k_0 to $k-1$ in order to obtain, for all $k > k_0$,

$$|d_k|_p^2 \leq |d_{k_0}|_p^2 + \widehat{c}\widetilde{\beta}L\widehat{\mathsf{C}}(x^*)\sum_{i=k_0}^{k-1}\frac{1 + |d_i|_p^2}{\sqrt{N_i}} + \widehat{c}\widehat{\mathsf{C}}(x^*)^2\sum_{i=k_0}^{k-1}\frac{1 + |d_i|_p^2}{N_i}. \tag{68}$$

By Assumption 5(i), we can choose $k_0 \in \mathbb{N}_0$ and $\gamma > 0$ as in (66). In particular, $\sum_{i \geq k_0} N_i < \gamma^2$. 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 (66), (68) and the definition of τ_a, we have

$$\begin{aligned} a^2 \leq |d_{\tau_a}|_p^2 &\leq |d_{k_0}|_p^2 + \widehat{c}\widetilde{\beta}L\widehat{\mathsf{C}}(x^*)\sum_{i=k_0}^{\tau_a - 1}\frac{1 + a^2}{\sqrt{N_i}} + \widehat{c}\widehat{\mathsf{C}}(x^*)^2\sum_{i=k_0}^{\tau_a - 1}\frac{1 + a^2}{N_i} \\ &\leq |d_{k_0}|_p^2 + \widehat{c}\widetilde{\beta}L\widehat{\mathsf{C}}(x^*)\frac{\phi(1 + a^2)}{\widehat{c}\widetilde{\beta}L\widehat{\mathsf{C}}(x^*)} + \widehat{c}\widehat{\mathsf{C}}(x^*)^2\frac{\phi^2(1 + a^2)}{\widehat{c}^2\widehat{\mathsf{C}}(x^*)^2} \\ &\leq |d_{k_0}|_p^2 + (\phi + \phi^2)(1 + a^2), \end{aligned} \tag{69}$$

using the fact that $\widehat{c} \geq 1$. Relation (69) and the fact that $0 < \phi + \phi^2 < 1$ imply

$$a^2 \leq \frac{|d_{k_0}|_p^2 + 1}{1 - \phi - \phi^2}. \tag{70}$$

Since (70) holds for an arbitrary $a > |d_{k_0}|_p$ and $\phi + \phi^2 \in (0, 1)$, it follows that $\sup_{k \geq k_0} |d_k|_p^2 \leq \left[1 - \phi - \phi^2\right]^{-1}\left[1 + |d_{k_0}|_p^2\right]$. 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 (68) obtaining

$$|d_k|_p^2 \leq |d_{k_0}|_p^2 + \widehat{c}\widetilde{\beta}L\widehat{\mathsf{C}}(x^*)\gamma(1 + \hat{a}^2) + \widehat{c}\widehat{\mathsf{C}}(x^*)^2\gamma^2(1 + \hat{a}^2), \tag{71}$$

where $\gamma := \frac{\phi}{\widehat{c}\max\{1, \widetilde{\beta}L\}\widehat{\mathsf{C}}(x^*)}$ is the constant in the right hand side of (66). Note that (71) holds trivially for $k := k_0$. Thus, after taking the supremum over $k \geq k_0$ in (71), we proceed as done after the inequalities (69)-(70) but with $\hat{a}$ substituting for $\hat{a}$, proving the required claim.

We now proceed to the convergence rate result for Algorithm 1. We need first the following lemma:

Lemma 14 (Lower bound on stepsize) *Consider Assumptions 2 and 4 with $\delta = 1$. Define $\widetilde{L}_k := \frac{1}{N_k}\sum_{j=1}^{N_k} L(\xi_j^k)$. If Algorithm 1 does not stop at iteration $k+1$, then*

$$\frac{1-\lambda}{\theta\beta_k} \leq \alpha_k \widetilde{L}_k. \tag{72}$$

Proof By the definition of the linear search rule (12), we have that, for $\widehat{z}^k = z^k(\theta\alpha_k)$,

$$\lambda\|x^k - \Pi(g^k)\|^2 > \beta_k\langle\widehat{F}(\xi^k,\widehat{z}^k), x^k - \Pi(g^k)\rangle$$

$$= \beta_k\langle\widehat{F}(\xi^k,\widehat{z}^k) - \widehat{F}(\xi^k,x^k), x^k - \Pi(g^k)\rangle + \langle\beta_k\widehat{F}(\xi^k,x^k), x^k - \Pi(g^k)\rangle$$

$$= \beta_k\langle\widehat{F}(\xi^k,\widehat{z}^k) - \widehat{F}(\xi^k,x^k), x^k - \Pi(g^k)\rangle + \langle x^k - g^k, x^k - \Pi(g^k)\rangle$$

$$\geq \beta_k\langle\widehat{F}(\xi^k,\widehat{z}^k) - \widehat{F}(\xi^k,x^k), x^k - \Pi(g^k)\rangle + \|x^k - \Pi(g^k)\|^2,$$

using Lemma 1(v) in the last inequality. From (73) and the fact that $\lambda \in (0,1)$, we get

$$(1-\lambda)\|x^k - \Pi(g^k)\|^2 \leq \beta_k\langle\widehat{F}(\xi^k,x^k) - \widehat{F}(\xi^k,\widehat{z}^k), x^k - \Pi(g^k)\rangle$$

$$\leq \beta_k\|\widehat{F}(\xi^k,x^k) - \widehat{F}(\xi^k,\widehat{z}^k)\|\|x^k - \Pi(g^k)\|$$

$$\leq \beta_k\widetilde{L}_k\theta\alpha_k\|x^k - \Pi(g^k)\|^2, \tag{73}$$

using the Lipschitz-continuity of $F(\xi_j^k,\cdot)$ for every $j \in [N_k]$ and the fact that $x^k - \widehat{z}^k = \theta\alpha_k[x^k - \Pi(g^k)]$ in the last inequality. Since $x^k \neq \Pi(g^k)$, (73) proves the claim in (72).

Define $\widehat{\mathsf{a}}_0^k := \sum_{i=0}^k \frac{1}{\sqrt{N_i}}$, $\mathsf{a}_0^k := \sum_{i=0}^k \frac{1}{N_i}$ and $\mathsf{b}_0^k := \sum_{i=0}^k \frac{1}{N_i^2}$. In the remainder of the subsection we will invoke the definitions given in Lemma 10 and Proposition 1.

Theorem 8 (Convergence rate) *Consider Assumptions 1-6. Suppose that $F(\xi,\cdot)$ is Lipschitz-continuous, i.e., that Assumption 4 holds with $\delta = 1$. Moreover, assume that Algorithm 1 generates an infinite sequence. Take $M > 0$ such that $\sup_{k\in\mathbb{N}_0}\left|\|\widehat{F}(\xi^k,z^k)\|\right|_4 \leq M$. Given $x^* \in X^*$, take $\phi \in (0,\frac{\sqrt{5}-1}{2})$ and $k_0 \in \mathbb{N}$ such that:*

$$\sum_{k\geq k_0}\frac{1}{\sqrt{N_k}} \leq \frac{\phi}{\widehat{c}\max\{1,\widetilde{\beta}L\}\widehat{\mathsf{C}}(x^*)}. \tag{74}$$

Define

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

$$\widetilde{\mathsf{J}}(x^*) := 2\frac{\lambda^2(1-\lambda)^2}{\theta^2\widetilde{\beta}^2|L(\xi)|_4^2M^2}C_2^2n^2\sigma(x^*)^4.$$

Then for all $\epsilon > 0$ there exists $K_\epsilon \in \mathbb{N}$ such that

$$\mathbb{E}\left[r_{\widehat{\beta}}(x^{K_\epsilon})^2\right] \leq \epsilon \leq \frac{\widehat{\mathsf{Q}}_\infty(x^*)}{\sqrt{K_\epsilon}},$$

where, for all $k \in \mathbb{N}_0 \cup \{\infty\}$,

$$\widehat{\mathsf{Q}}_k(x^*)^2 := \frac{4\theta^2\widetilde{\beta}^4|L(\xi)|_4^2M^2}{\lambda^2(1-\lambda^2)}.$$

$$\left\{\|x^0 - x^*\|^2 + \left[1+\widehat{\mathsf{J}}(x^*)\right]\left[\widehat{c}\widetilde{\beta}L\widehat{\mathsf{C}}(x^*)\widehat{\mathsf{a}}_0^k + \widehat{c}\widehat{\mathsf{C}}(x^*)^2\mathsf{a}_0^k\right] + \widetilde{\mathsf{J}}(x^*)\left[1+\widehat{\mathsf{J}}(x^*)^2\right]\mathsf{b}_0^k\right\}.$$

Proof Take total expectation in Proposition 1(i) and obtain, for all $k \in \mathbb{N}$ and $x^* \in X^*$,

$$\begin{aligned}\mathbb{E}\big[\|x^{k+1}-x^*\|^2\big] &\leq \mathbb{E}\left[\|x^k-x^*\|^2\right] - \mathbb{E}\big[\|y^k-x^k\|^2\big] \\ &+\widehat{c}\left[\frac{\tilde{\beta}L\widehat{\mathsf{C}}(x^*)}{\sqrt{N}_k}+\frac{\widehat{\mathsf{C}}(x^*)^2}{N_k}\right]\left(1+\mathbb{E}\left[\|x^k-x^*\|^2\right]\right), \qquad (75)\end{aligned}$$

using the hereditary property $\mathbb{E}[\mathbb{E}[\cdot|\mathcal{F}_k]] = \mathbb{E}[\cdot]$.

Denote temporarily $H_k := \widehat{F}(\xi^k,\cdot)$ and $\varrho_k := \mathbb{E}\left[r_{\widetilde{\beta}}(H_k,x^k)^2\right]$. From the linear search rule (12), the definition of γ_k and the fact that $x^k - z^k = \alpha_k(x^k - \Pi(g^k))$, we get

$$\begin{aligned}\|y^k-x^k\|^2 = \gamma_k^2\|\widehat{F}(\xi^k,z^k)\|^2 &= \frac{\langle\widehat{F}(\xi^k,z^k),x^k-z^k\rangle^2}{\|\widehat{F}(\xi^k,z^k)\|^4}\|\widehat{F}(\xi^k,z^k)\|^2 \\ \geq \frac{\lambda^2\alpha_k^2\|x^k-\Pi(g^k)\|^4}{\beta_k^2\|\widehat{F}(\xi^k,z^k)\|^2} &\geq \frac{\lambda^2(1-\lambda)^2\|x^k-\Pi(g^k)\|^4}{\theta^2\tilde{L}_k^2\|\widehat{F}(\xi^k,z^k)\|^2\beta_k^4} \\ &\geq \frac{\lambda^2(1-\lambda)^2 r_{\widetilde{\beta}}(H_k,x^k)^4}{\theta^2\tilde{L}_k^2\|\widehat{F}(\xi^k,z^k)\|^2\widetilde{\beta}^4} \qquad (76)\end{aligned}$$

using the bound of Lemma 14 in the second inequality and the facts that $x^k - \Pi(g^k) = r_{\beta_k}(H_k,x^k)$ and $0 < \beta_k \leq \widetilde{\beta}$, together with Lemma 2, in the last inequality. We take $\mathbb{E}[\sqrt{\cdot}]$ in (76), obtaining

$$\begin{aligned}\varrho_k = \mathbb{E}\left[r_{\widetilde{\beta}}(H_k,x^k)^2\right] &\leq \frac{\theta\widetilde{\beta}^2}{\lambda(1-\lambda)}\mathbb{E}\left[\tilde{L}_k\|\widehat{F}(\xi^k,z^k)\|\|y^k-x^k\|\right] \\ &\leq \frac{\theta\widetilde{\beta}^2}{\lambda(1-\lambda)}|\tilde{L}_k|_4\cdot\left|\|\widehat{F}(\xi^k,z^k)\|\right|_4\cdot\left|\|y^k-x^k\|\right|_2 \\ &\leq \frac{\theta\widetilde{\beta}^2}{\lambda(1-\lambda)}|L(\xi)|_4\cdot\left|\|\widehat{F}(\xi^k,z^k)\|\right|_4\cdot\left|\|y^k-x^k\|\right|_2, \qquad (77)\end{aligned}$$

using Hölder's inequality in the second inequality, and the fact that ξ^k is an i.i.d. sample of ξ in the last inequality. We now observe that from Assumptions 4 and 6 and Proposition 4, $|L(\xi)|_4 < \infty$ and $\sup_{k\in\mathbb{N}_0}\left|\|\widehat{F}(\xi^k,z^k)\|\right|_4 \leq M$ for some $M > 0$. Hence, we define

$$\varpi := \frac{\lambda^2(1-\lambda)^2}{\theta^2\tilde{\beta}^4|L(\xi)|_4^2M^2}. \qquad (78)$$

We observe that

$$\begin{aligned}r_{\widehat{\beta}}(x^k)^2 \leq r_{\beta_k}(x^k)^2 &\leq 2\|x^k-\Pi(g^k)\|^2 + 2\left\|\Pi\left[x^k-\beta_kT(x^k)\right]-\Pi(g^k)\right\|^2 \\ &\leq 2r_{\widetilde{\beta}}(H_k,x^k)^2 + 2\tilde{\beta}^2\|\bar{\epsilon}_1^k\|^2, \qquad (79)\end{aligned}$$

using the fact that $\beta \mapsto r_\beta(H,x)$ is a non-decreasing function for any $x \in \mathbb{R}^n$ and any operator H, the fact that $(a+b)^2 \leq 2a^2 + 2b^2$ and Lemma 1(iii). We take $\mathbb{E}[\cdot]$ in (79) and obtain

$$\begin{aligned}\widehat{\varrho}_k := \mathbb{E}\left[r_{\widehat{\beta}}(x^k)^2\right] &\leq 2\varrho_k + 2\tilde{\beta}^2\mathbb{E}\left[\|\bar{\epsilon}_1^k\|^2\right] \\ &\leq 2\varrho_k + 2\tilde{\beta}^2C_2n\sigma(x^*)^2\frac{1+\mathbb{E}\left[\|x^k-x^*\|^2\right]}{N_k}, \qquad (80)\end{aligned}$$

using Lemma 5, the fact that $x^K \in \mathcal{F}_K$ and the independence between ξ^K and $\mathcal{F}_K$ in the second inequality.

In view of (75)-(78), taking squares in (80) and using the convexity of $t \mapsto t^2$, we get for all $k \in \mathbb{N}_0$ and $x^* \in X^*$,

$$\begin{aligned}\mathbb{E}\big[\|x^{k+1}-x^*\|^2\big] &\leq \mathbb{E}\left[\|x^k-x^*\|^2\right] - \frac{\varpi}{4}\widehat{\varrho}_k^2 \\ +\widehat{c}&\left[\frac{\tilde{\beta}L\widehat{\mathsf{C}}(x^*)}{\sqrt{N}_k}+\frac{\widehat{\mathsf{C}}(x^*)^2}{N_k}\right]\left(1+\mathbb{E}\left[\|x^k-x^*\|^2\right]\right)\end{aligned}$$

$$+\frac{2\varpi\tilde{\beta}^2C_2^2n^2\sigma(x^*)^4}{N_k^2}\left(1+\mathbb{E}\left[\|x^k-x^*\|^2\right]^2\right). \tag{81}$$

We sum (81) with i from 0 to k, obtaining:

$$\frac{\varpi}{4}\sum_{i=0}^{k}\widehat{\varrho}_i^2\leq\|x^0-x^*\|^2+\widehat{c}\tilde{\beta}L\widehat{\mathsf{C}}(x^*)\sum_{i=0}^{k}\frac{1+\mathbb{E}\left[\|x^i-x^*\|^2\right]}{\sqrt{N}_i}$$

$$+\widehat{c}\widehat{\mathsf{C}}(x^*)^2\sum_{i=0}^{k}\frac{1+\mathbb{E}\left[\|x^i-x^*\|^2\right]}{N_i}+2\varpi\tilde{\beta}^2C_2^2n^2\sigma(x^*)^4\sum_{i=0}^{k}\frac{1+\mathbb{E}\left[\|x^i-x^*\|^2\right]^2}{N_i^2}\leq$$

$$\leq\|x^0-x^*\|^2+\left(1+\sup_{0\leq i\leq k}\mathbb{E}\left[\|x^i-x^*\|^2\right]\right)\left(\widehat{c}\tilde{\beta}L\widehat{\mathsf{C}}(x^*)\sum_{i=0}^{k}\frac{1}{\sqrt{N}_i}+\widehat{c}\widehat{\mathsf{C}}(x^*)^2\sum_{i=0}^{k}\frac{1}{N_i}\right)$$

$$+2\varpi\tilde{\beta}^2C_2^2n^2\sigma(x^*)^4\left(1+\sup_{0\leq i\leq k}\mathbb{E}\left[\|x^i-x^*\|^2\right]^2\right)\sum_{i=0}^{k}\frac{1}{N_i^2}$$

$$\leq\|x^0-x^*\|^2+\left[1+\widehat{\mathsf{J}}(x^*)\right]\left[\widehat{c}\tilde{\beta}L\widehat{\mathsf{C}}(x^*)\widehat{\mathsf{a}}_0^k+\widehat{c}\widehat{\mathsf{C}}(x^*)^2\mathsf{a}_0^k\right]$$

$$+\widetilde{\mathsf{J}}(x^*)\left[1+\widehat{\mathsf{J}}(x^*)^2\right]\mathsf{b}_0^k=\frac{\varpi}{4}\widehat{\mathsf{Q}}_k(x^*)^2, \tag{82}$$

using in the last inequality (66) and Proposition 4 for $p=2$, which imply

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

and, hence, $\sup_{k\geq 0}\mathbb{E}[\|x^k-x^*\|^2]\leq\widehat{\mathsf{J}}(x^*)$, since $1-\phi-\phi^2\in(0,1)$.

Given $\epsilon>0$, define

$$K=K_\epsilon:=\inf\left\{k\in\mathbb{N}_0:\widehat{\varrho}_k\leq\epsilon\right\}.$$

From the definition of K we have, for every $k<K$,

$$\frac{\varpi}{4}\epsilon^2(k+1)\leq\frac{\varpi}{4}\sum_{i=0}^{k}\widehat{\varrho}_i^2. \tag{83}$$

We claim that K is finite. Indeed, if $K=\infty$, then (82)-(83) hold for all $k\in\mathbb{N}$. Hence, we arrive at a contradiction by letting $k\to\infty$ and using the fact that $\mathsf{b}_0^\infty\leq\mathsf{a}_0^\infty\leq\widehat{\mathsf{a}}_0^\infty<\infty$, which holds by Assumption 5(i). Since K is finite, we have that $\widehat{\varrho}_K\leq\epsilon$ by definition. Setting $k:=K-1$ in (82)-(83), we get $K\epsilon^2\leq\widehat{\mathsf{Q}}_{K-1}(x^*)^2\leq\widehat{\mathsf{Q}}_\infty(x^*)^2$, using the definition of $\widehat{\mathsf{Q}}_k(x^*)$. Invoking this fact and the definition of $\widehat{\varrho}_K$ in (80), we establish the result.

We now give the oracle complexity of Algorithm 1 for a natural choice of the sampling rate.

Proposition 5 (Rate and oracle complexity) *Assume that the hypotheses of Theorem 8 hold. Define N_k as*

$$N_k=\left\lceil\Theta n\sigma(x^*)^2(k+\mu)^2(\ln(k+\mu))^{2+2b}\right\rceil$$

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 be the minimum natural number satisfying

$$k_0\geq\exp\left[\left(\frac{\widehat{c}\max\{1,\tilde{\beta}L\}\max\{C_p,C_{p,L}\}\tilde{\beta}}{\phi b\sqrt{\Theta}\lambda}\right)^{1/b}\right]-\mu+1. \tag{84}$$

Define

$$\mathit{\Lambda}_1:=\frac{\widehat{c}\widetilde{\beta}^2L\max\{C_p,C_{p,L}\}}{\lambda},\quad\mathit{\Lambda}_2:=\frac{\widehat{c}\widetilde{\beta}^2\max\{C_p^2,C_{p,L}^2\}}{\lambda^2},\quad\mathit{\Lambda}_3:=\frac{2\lambda^2(1-\lambda)^2C_2^2}{\theta^2\tilde{\beta}^2|L(\xi)|_4^2M^2},$$

$$\widehat{\mathcal{A}}:=\frac{\mathit{\Lambda}_1}{b(\ln(\mu-1))^b},\quad\widehat{\mathcal{B}}:=\frac{\mathit{\Lambda}_2}{(\mu-1)(1+2b)[\ln(\mu-1)]^{1+2b}},\quad\mathcal{C}:=\frac{\mathit{\Lambda}_3}{3(\mu-1)^3}.$$

Then Theorem 6 holds, and for all $\epsilon > 0$ *there exists* $K := K_\epsilon \in \mathbb{N}$ *such that* $\mathbb{E}[r_{\widehat{\beta}}(x^K)^2] \leq \epsilon$ *and*

$$\epsilon \leq \frac{2\theta\widetilde{\beta}^2|L(\xi)|_4 M \max\{1, \Theta^{-1}\}}{\lambda\sqrt{1-\lambda^2}\sqrt{K}} \times$$

$$\left\{\|x^0 - x^*\| + \left[\widehat{\mathcal{A}} + \widehat{\mathcal{B}}\right]^{-\frac{1}{2}} \left[1 + \sqrt{\widehat{\mathsf{J}}(x^*)}\right] + \sqrt{\mathcal{C}}\left[1 + \widehat{\mathsf{J}}(x^*)\right]\right\}, \tag{85}$$

$$\sum_{k=1}^{K} 2N_k \lesssim \max\{1, \Theta^{-6}\} \max\{1, \Theta n\sigma(x^*)^2\} \frac{\widehat{\mathsf{P}}(x^*)}{\epsilon^6} \widehat{\mathsf{I}}(x^*),$$

$$\widehat{\mathsf{P}}(x^*) := \left\{\ln\left[(\widehat{\mathsf{Q}}_\infty(x^*) + \epsilon)\epsilon^{-2}\right]\right\}^{2+2b} + 2\mu^{-2},$$

$$\widehat{\mathsf{I}}(x^*) \lesssim \frac{\theta^6\widetilde{\beta}^{12}|L(\xi)|_4^6 M^6}{\lambda^6(1-\lambda^2)^3} \left\{\|x^0 - x^*\|^6 + (\widehat{\mathcal{A}} + \widehat{\mathcal{B}})^3\left[1 + \widehat{\mathsf{J}}(x^*)^3\right] + \mathcal{C}^3\left[1 + \widehat{\mathsf{J}}(x^*)^6\right]\right\}. \tag{86}$$

Proof For $\phi \in (0, \frac{\sqrt{5}-1}{2})$, we look for k_0 satisfying (74). We have

$$\begin{aligned}\sum_{k\geq k_0} \frac{1}{\sqrt{N_k}} &\leq \Theta^{-1/2} n^{-1/2} \sigma(x^*)^{-1} \sum_{k\geq k_0} \frac{1}{(k+\mu)(\ln(k+\mu))^{1+b}} \\ &\leq \Theta^{-1/2} n^{-1/2} \sigma(x^*)^{-1} \int_{k_0-1}^{\infty} \frac{dt}{(t+\mu)(\ln(t+\mu))^{1+b}} \\ &= \frac{\Theta^{-1/2} n^{-1/2} \sigma(x^*)^{-1}}{b(\ln(k_0 - 1 + \mu))^b}.\end{aligned} \tag{87}$$

From (74) and (87), it is enough to choose k_0 as the minimum natural number such that the right hand side of (87) is less than $\frac{\phi}{\widehat{c}\max\{1,\widetilde{\beta}L\}\widehat{\mathsf{C}}(x^*)}$. In view of the definition of $\widehat{\mathsf{C}}(x^*)$, it suffices to choose k_0 as in (84).

We now give an estimate of $\widehat{\mathsf{Q}}_\infty(x^*)$. Using the definitions of $\widehat{\mathsf{C}}(x^*)$ and N_k, we get the bound

$$\widehat{c}\tilde{\beta}L\widehat{\mathsf{C}}(x^*)\widehat{\mathsf{a}}_0^k + \widehat{c}\widehat{\mathsf{C}}(x^*)^2\mathsf{a}_0^k \leq \int_{-1}^{\infty} \frac{\Lambda_1\Theta^{-1/2}dt}{(t+\mu)(\ln(t+\mu))^{1+b}} + \tag{88}$$

$$+ \int_{-1}^{\infty} \frac{\Lambda_2\Theta^{-1}dt}{(t+\mu)^2(\ln(t+\mu))^{2+2b}} \leq \frac{\Lambda_1\Theta^{-1/2}}{b(\ln(\mu-1))^b} + \frac{\Lambda_2\Theta^{-1}}{(\mu-1)(1+2b)[\ln(\mu-1)]^{1+2b}}.$$

From the definitions of $\widetilde{\mathsf{J}}(x^*)$ and N_k we also have

$$\widetilde{\mathsf{J}}(x^*)\mathsf{b}_0^k \leq \int_{-1}^{\infty} \frac{\Lambda_3\Theta^{-2}dt}{(t+\mu)^4(\ln(t+\mu))^{4+4b}} \leq \frac{\Lambda_3\Theta^{-2}}{3(\mu-1)^3}. \tag{89}$$

At this point, we obtain (85) from Theorem 8, (88)-(89) and the definitions of $\widehat{\mathsf{Q}}_\infty(x^*)$, $\widehat{\mathsf{J}}(x^*)$, $\widetilde{\mathsf{J}}(x^*)$, $\widehat{\mathcal{A}}$, $\widehat{\mathcal{B}}$ and $\mathcal{C}$

We now prove (86). Using the facts that $K := K_\epsilon \leq \widehat{\mathsf{Q}}_\infty(x^*)^2/\epsilon^2$, $\mu\epsilon \leq 1$ and $N_k \leq \Theta n\sigma(x^*)^2(k+\mu)^2(\ln(k+\mu))^{2+2b} + 1$, we have

$$\begin{aligned}\sum_{k=1}^{K} 2N_k &\leq \max\{\Theta n\sigma(x^*)^2, 1\} \sum_{k=1}^{K} 2\left[(k+\mu)^2(\ln(k+\mu))^{2+2b} + 1\right] \\ &\leq \max\{\Theta n\sigma(x^*)^2, 1\} 2K(K+\mu)^2\left[(\ln(K+\mu))^{2+2b} + \frac{2}{(K+\mu)^2}\right] \\ &\leq \max\{\Theta n\sigma(x^*)^2, 1\} \frac{\widehat{\mathsf{Q}}_\infty(x^*)^2}{\epsilon^2}\left(\frac{\widehat{\mathsf{Q}}_\infty(x^*)^2 + \epsilon}{\epsilon^2}\right)^2 \cdot \\ &\left\{\left[\ln\left(\left(\widehat{\mathsf{Q}}_\infty(x^*)^2 + \epsilon\right)\epsilon^{-2}\right)\right]^{2+2b} + 2\mu^{-2}\right\}.\end{aligned}$$

The definitions of $\widehat{\mathsf{Q}}_\infty(x^*)$, $\widehat{\mathsf{J}}_\infty(x^*)$, $\widetilde{\mathsf{J}}_\infty(x^*)$, $\widehat{\mathcal{A}}$, $\widehat{\mathcal{B}}$ and $\mathcal{C}$, (88)-(89) and the convexity of $t \mapsto t^6$ imply (86).

3.3.2 ***The stochastic extragradient method with linear search***

The following result gives explicit bounds on the p-norm of the sequence generated by Algorithm 2. The proofline is identical to that of Proposition 4 in [11] up to changes in the values of certain constants due to the introduction of the linear search and the use of Lemma 6 in Proposition 2, and hence we skip the proof.

Proposition 6 (Uniform boundedness in L^p) *Let Assumptions 1-6 hold. Take $p \in \{2\} \cup [4, \infty)$. Assume that Algorithm 2 generates an infinite sequence. Then, for every $x^* \in X^*$, there exist constants $\{\mathsf{D}(x^*), \mathsf{B}_p(x^*)\}$, with $\mathsf{B}_2(x^*) = 0$, for which, given $k_0 := k_0(x^*) \in \mathbb{N}$ and $\gamma := \gamma(x^*) > 0$ such that*

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

the following estimate holds:

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

with $\mathsf{c}_2(x^*) = [1 - \beta(x^*)]^{-1}$ and $\mathsf{c}_p(x^*) = 4[1 - \beta(x^*)]^{-2}$ for $p \geq 4$.

Remark 2 The following bounds, in terms of the variance and the space dimension, hold for the constants guaranteed by Proposition 6:

$$\mathsf{D}(x^*) \lesssim C_p^2 n \sigma(x^*)^2,$$

$$\mathsf{B}_p(x^*) \lesssim C_q C_p \sigma(x^*) \left[(1 + L\hat{\alpha})^2 + (1 + L\hat{\alpha}) C_p \sqrt{n} \sigma(x^*) + C_p^2 n \sigma(x^*)^2 \right].$$

Remark 3 In the statement of the proposition, for $p = 2$, it is sufficient to set $\phi \in (0, \frac{\sqrt{5}-1}{2})$ and $k_0 := k_0(x^*)$ such that $\sum_{k \geq k_0} N_k^{-1} \leq \phi \mathsf{D}(x^*)^{-1}$ in order 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 state the convergence rate estimate for Algorithm 2. The proof of Theorem 9 uses Propositions 3 and 6 for $p = 2$ with a proofline similar to that of Theorem 4 in [11]. A minor modification is the need of the lower bound:

$$\mathbb{E}\left[r_{\alpha_k}(x^k)^2 \middle| \mathcal{F}_k \right] \geq \mathbb{E}\left[\alpha_k^2 \middle| \mathcal{F}_k \right] r(x^k)^2 \geq \frac{\lambda^2 \theta^2}{|L(\xi)|_2^2} r(x^k)^2,$$

which follows from Lemmas 2, 13 and the fact that $x^k \in \mathcal{F}_k$.

Theorem 9 (Convergence rate) *Consider Assumptions 1-6. Assume that Algorithm 2 generates an infinite sequence. Given $x^* \in X^*$, take $\phi \in (0, \frac{\sqrt{5}-1}{2})$ and $k_0 \in \mathbb{N}$ such that:*

$$\sum_{k \geq k_0} \frac{1}{N_k} \leq \frac{\phi}{\mathsf{D}(x^*)}.$$

Define

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

Then for all $\epsilon > 0$ there exists $K_\epsilon \in \mathbb{N}$ such that $\mathbb{E}[r(x^{K_\epsilon})^2] \leq \epsilon \leq \frac{\mathsf{Q}_\infty(x^*)}{K_\epsilon}$, where

$$\mathsf{Q}_k(x^*) := \frac{2|L(\xi)|_2^2}{(1 - 6\lambda^2)\lambda^2\theta^2} \left\{ \|x^0 - x^*\|^2 + [1 + \mathsf{J}(x^*)] \left[\mathsf{D}(x^*)\mathsf{a}_0^k + \mathsf{D}(x^*)^2 \mathsf{b}_0^k \right] \right\}$$

for all $k \in \mathbb{N}_0 \cup \{\infty\}$.

Finally we state the oracle complexity result for a specific choice of sample rate. The proof of Proposition 7 uses Theorem 9 and follows a proofline identical to that of Proposition 6 in [11], so that we skip it.

Note that the estimate of $\mathsf{D}(x^*)$ presented in Remark 2 can be rewritten in a more precise way as $\mathsf{D}(x^*) \leq c C_p^2 n \sigma(x^*)^2$ for some constant $c > 0$ which satisfies $c \lesssim \hat{\alpha}^2 L^2$. We will use this constant c in the statement of the following proposition.

Proposition 7 (Rate and oracle complexity) *Assume that the hypotheses of Theorem 9 hold. Define N_k as*

$$N_k = \left\lceil \Theta n\sigma(x^*)^2(k+\mu)(\ln(k+\mu))^{1+b} \right\rceil$$

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 be the minimum natural number satisfying

$$k_0 \geq \exp\left[\left(\frac{2c\hat{\alpha}^2C_p^2}{\phi b\Theta}\right)^{1/b}\right] - \mu + 1.$$

Define

$$\Lambda := 2c\hat{\alpha}^2C_p^2, \ \ \mathcal{A} := \frac{\Lambda}{b(\ln(\mu-1))^b}, \ \ \mathcal{B} := \frac{\Lambda^2}{(\mu-1)(1+2b)[\ln(\mu-1)]^{1+2b}}.$$

Then Theorem 7 holds and for all $\epsilon > 0$, there exists $K := K_\epsilon \in \mathbb{N}$ such that $\mathbb{E}[r(x^K)^2] \leq \epsilon$ and

$$\epsilon \leq \frac{2|L(\xi)|_2^2 \max\{1, \Theta^{-2}\}}{(1-6\lambda^2)\lambda^2\theta^2K} \cdot \left\{\|x^0 - x^*\|^2 + (\mathcal{A}+\mathcal{B})\left[1 + \mathsf{J}(x^*)\right]\right\},$$

$$\sum_{k=1}^{K} 2N_k \leq 12\max\{1, \Theta^{-2}\}\max\{1, \Theta n\sigma(x^*)^2\}\frac{\mathsf{P}(x^*)}{\epsilon^2}\mathsf{I}(x^*),$$

$$\mathsf{P}(x^*) := \left\{\ln\left[(\mathsf{Q}_\infty(x^*)+1)\epsilon^{-1}\right]\right\}^{1+b} + \mu^{-1},$$

$$\mathsf{I}(x^*) := \left[\frac{|L(\xi)|_2^2}{(1-6\lambda^2)\lambda^2\theta^2}\right]^2 \|x^0 - x^*\|^4 + \left[\frac{|L(\xi)|_2^2}{(1-6\lambda^2)\lambda^2\theta^2}\right]^2 (\mathcal{A}+\mathcal{B})^2\left[1 + \mathsf{J}(x^*)\right]^2 + 1.$$

Remark 4 (Number of oracle calls in linear search) A notable difference between Algorithm 2 and Algorithm (5)-(6) in [11] (which uses explicitly the Lipschitz constant L), is that during the linear search (18) of iteration k, the oracle is called $j_k := \log_{\frac{1}{\theta}}\left(\frac{\alpha_k}{\hat{\alpha}}\right)$ times. This is the price to be paid when no information on L is available. Nevertheless, Lemma 13 implies the upper bound $j_k \leq 1 + \log_{\frac{1}{\theta}}\left(\frac{\lambda}{\hat{\alpha}\widetilde{L}_k}\right)$. Note that, since α_k is a random variable, this bound holds in an almost sure sense. Hence, with respect to the tolerance $\epsilon > 0$, the total number of oracle calls is a.s. not greater than

$$\sum_{k=1}^{K_\epsilon} j_k \cdot 2N_k \lesssim \left[1 + \log_{\frac{1}{\theta}}\left(\frac{\lambda}{\hat{\alpha}}\right) - \min_{k\in[K_\epsilon]}\log_{\frac{1}{\theta}}\widetilde{L}_k\right] \cdot \frac{\left[\ln(\epsilon^{-1})\right]^{1+b}}{\epsilon^2}, \tag{90}$$

in view of Proposition 7. Hence, even though the absence of the Lipschitz constant requires more oracle calls than with its knowledge, the additional multiplicative constant in estimate (90) is not significantly higher for practical purposes. We remark also that, by the strong law of large numbers, a.s. $\lim_{k\to\infty}\widetilde{L}_k = L$ and, under light-tail assumptions on $L(\xi)$, we can infer that $\widetilde{L}_k \approx L$ with high-probability for large enough k or Θ.

4 Appendix

4.1 Recursion lemmas

Proof of Lemma 8

Proof By Lemma 7(ii), we have that $\gamma_k > 0$. Thus

$$\begin{aligned}\|x^{k+1} - x\|^2 &= \|\Pi(y^k) - x\|^2 \\ &\leq \|y^k - x\|^2 - \|y^k - \Pi(y^k)\|^2 \\ &\leq \|y^k - x\|^2 \\ &= \|(x^k - x) - \gamma_k(T(z^k) + \bar{\epsilon}_2^k)\|^2 \\ &= \|x^k - x\|^2 + \gamma_k^2\|T(z^k) + \bar{\epsilon}_2^k\|^2 - 2\gamma_k\langle T(z^k) + \bar{\epsilon}_2^k, x^k - x\rangle,\end{aligned} \tag{91}$$

using Lemma 1(ii) in the first inequality. Concerning the last term in the rightmost expression of (91), we have

$$\begin{aligned}
-2\gamma_k\langle T(z^k)+\bar{\epsilon}_2^k, x^k-x\rangle &= -2\gamma_k\langle T(z^k)+\bar{\epsilon}_2^k, x^k-z^k\rangle + \\
&\quad 2\gamma_k\langle T(z^k), x-z^k\rangle + 2\gamma_k\langle\bar{\epsilon}_2^k, x-z^k\rangle \\
&= -2\gamma_k(\gamma_k\|T(z^k)+\bar{\epsilon}_2^k\|^2) \\
&\quad +2\gamma_k\langle T(z^k), x-z^k\rangle + 2\gamma_k\langle\bar{\epsilon}_2^k, x-z^k\rangle \\
&\leq -2\gamma_k^2\|T(z^k)+\bar{\epsilon}_2^k\|^2 + 2\gamma_k\langle\bar{\epsilon}_2^k, x-z^k\rangle, \qquad (92)
\end{aligned}$$

using the definition of γ_k in the second equality, and the facts that $\gamma_k > 0$ and $\langle T(z^k), x-z^k\rangle \leq 0$ (which follows from the pseudo-monotonicity of T, and the facts $x \in X^*$, $z^k \in X$) in the inequality. Combining (91)-(92) we get

$$\begin{aligned}
\|x^{k+1}-x\|^2 &\leq \|x^k-x\|^2 + \gamma_k^2\|T(z^k)+\bar{\epsilon}_2^k\|^2 - 2\gamma_k^2\|T(z^k)+\bar{\epsilon}_2^k\|^2 + 2\gamma_k\langle\bar{\epsilon}_2^k, x-z^k\rangle \\
&= \|x^k-x\|^2 - \gamma_k^2\|T(z^k)+\bar{\epsilon}_2^k\|^2 + 2\gamma_k\langle\bar{\epsilon}_2^k, x-z^k\rangle \\
&= \|x^k-x\|^2 - \|y^k-x^k\|^2 + 2\gamma_k\langle\bar{\epsilon}_2^k, x-z^k\rangle, \qquad (93)
\end{aligned}$$

using the fact that $\|y^k-x^k\| = \gamma_k\|T(z^k)+\epsilon_2^k\|$ (which follows from the definition of γ_k), in the last equality.

Proof of Lemma 12

Proof We shall modify part of the proof of Lemma 3 in [11]. Relations (A.1)-(A.3) in the proof of Lemma 3 in the Appendix of [11] imply that, for all $x^* \in X^*$ and $k \in \mathbb{N}_0$,

$$\begin{aligned}
\|x^{k+1}-x^*\|^2 \leq \|x^k-x\|^2 - \|x^k-z^k\|^2 + 2\alpha_k^2\|\widehat{F}(\eta^k,z^k)-\widehat{F}(\xi^k,x^k)\|^2 \\
+2\langle x-z^k, \alpha_k\widehat{F}(\eta^k,z^k)\rangle. \qquad (94)
\end{aligned}$$

The last term in the rightmost expression of (94) is bounded by

$$\begin{aligned}
2\langle x^*-z^k, \alpha_k\widehat{F}(\eta^k,z^k)\rangle &= 2\langle x^*-z^k, \alpha_k(T(z^k)+\epsilon_2^k)\rangle \\
&= 2\langle x^*-z^k, \alpha_k T(z^k)\rangle + 2\langle x^*-z^k, \alpha_k\epsilon_2^k\rangle \qquad (95) \\
&\leq 2\alpha_k\langle x^*-z^k, \epsilon_2^k\rangle,
\end{aligned}$$

using the fact that $\langle x^*-z^k, T(z^k)\rangle \leq 0$ (which follows from Assumption 3 and the facts that $x^* \in X^*$, $z^k \in X$), in the last inequality.

Concerning the third term in the right hand side of (94), we have

$$\begin{aligned}
\alpha_k^2\|\widehat{F}(\eta^k,z^k)-\widehat{F}(\xi^k,x^k)\|^2 &\leq 3\alpha_k^2\|\widehat{F}(\xi^k,z^k)-\widehat{F}(\xi^k,x^k)\|^2 \\
&\quad +3\alpha_k^2\|\widehat{F}(\eta^k,z^k)-T(z^k)\|^2 + 3\alpha_k^2\|\widehat{F}(\xi^k,z^k)-T(z^k)\|^2 \\
&\leq 3\lambda^2\|z^k-x^k\|^2 + 3\hat{\alpha}^2\|\epsilon_2^k\|^2 + 3\hat{\alpha}^2\|\epsilon_3^k\|^2, \qquad (96)
\end{aligned}$$

using (18)-(19) and the convexity of $t \mapsto t^2$ in the first inequality and the linear search (18) in the last inequality.

Since $z^k = \Pi[x^k - \alpha_k(T(x^k)+\epsilon_1^k)]$, we get

$$\begin{aligned}
r_{\alpha_k}(x^k)^2 &= \|x^k - \Pi[x^k-\alpha_k T(x^k)]\|^2 \\
&\leq 2\|x^k-z^k\|^2 + 2\|\Pi[x^k-\alpha_k(T(x^k)+\epsilon_1^k)] - \Pi[x^k-\alpha_k T(x^k)]\|^2 \\
&\leq 2\|x^k-z^k\|^2 + 2\hat{\alpha}^2\|\epsilon_1^k\|^2, \qquad (97)
\end{aligned}$$

using Lemma 1(iii) in the second inequality. We complete the proof invoking (94)-(97) and the fact that $0 < 1-6\lambda^2 < 1$.

4.2 **Proof of Lemma 6**

In this subsection we will establish Lemma 6, for which we need first some preliminary results, bounding the supremum of the empirical average of the oracle error on a closed ball. As usual, $\|\cdot\|$ is the Euclidean norm. $B[x,R]$ denotes the closed Euclidean ball with center $x\in\mathbb{R}^n$ and radius $R>0$. $\mathbb{Q}[x,R]$ denotes the set of points in $B[x,R]$ with rational coordinates.

Fix $R>0$, $x_*\in X$ and choose an i.i.d sample $\xi^N:=\{\xi_j\}_{j=1}^N$ of ξ. We need to bound the following empirical process:

$$\begin{aligned}
Z:=Z(\xi^N;x_*,R):&=\sup_{x\in B[x_*,R]}\frac{1}{N}\left\|\sum_{j=1}^N\epsilon(\xi_j,x)-\epsilon(\xi_j,x_*)\right\|\\
&=\sup_{u\in B[0,1]}\frac{1}{N}\left\|\sum_{j=1}^N\epsilon(\xi_j,x_*+Ru)-\epsilon(\xi_j,x_*)\right\|\\
&=\sup_{u\in \mathbb{Q}[0,1]}\frac{1}{N}\left\|\sum_{j=1}^N\epsilon(\xi_j,x_*+Ru)-\epsilon(\xi_j,x_*)\right\| \qquad (98)\\
&=\sup_{u\in \mathbb{Q}[0,1]}\frac{1}{N}\sup_{\alpha\in B[0,1]}\left\langle\sum_{j=1}^N\epsilon(\xi_j,x_*+Ru)-\epsilon(\xi_j,x_*),\alpha\right\rangle\\
&=\sup_{u\in \mathbb{Q}[0,1]}\sup_{\alpha\in \mathbb{Q}[0,1]}\frac{1}{N}\sum_{j=1}^N\left\langle\epsilon(\xi_j,x_*+Ru)-\epsilon(\xi_j,x_*),\alpha\right\rangle,
\end{aligned}$$

We mention that the third equality follows from the fact that $F(\xi,\cdot)$ is continuous, while the fourth equality requires the fact that $\|v\|=\sup_{u\in B[0,1]}\langle u,v\rangle$, for all $v\in\mathbb{R}^n$. For every $j\in[N]$ and for every $(u,\alpha)\in\mathbb{Q}[0,1]^2$ we shall use the notation $t:=(u,\alpha)$, $\mathcal{T}:=\mathbb{Q}[0,1]^2$ and

$$X_{j,t}:=\frac{1}{N}\left\langle\epsilon(\xi_j,x_*+Ru)-\epsilon(\xi_j,x_*),\alpha\right\rangle, \qquad (99)$$

$$X_t:=\sum_{j=1}^N X_{j,t}. \qquad (100)$$

With the notation above, we have $Z=\sup_{t\in\mathcal{T}}X_t=\sup_{t\in\mathcal{T}}|X_t|$.

We shall invoke next Theorem 2. As we shall see, concerning $Z:=Z(\xi^N;x_*,R)$ in Definition (98), bounding $|M|_p$ and $\widehat{\sigma}^2$ is not so difficult. In order to bound the expected value $\mathbb{E}[Z]$ we shall need the following lemma. Its proof follows the proofline of Lemma 13.1 in [3] with minor modifications. We must also recall the definition of metric entropy given in the Preliminaries.

Lemma 15 *Let $(\mathcal{T},d)$ be a countable pseudo-metric space. Take some $t_0\in\mathcal{T}$ and let $\theta:=\sup_{t\in\mathcal{T}}\mathrm{d}(t,t_0)$. Let $(X_t)_{t\in\mathcal{T}}$ be a collection of random variables such that, for some constant $v>0$,*

$$\ln\mathbb{E}[\exp\{\lambda(X_t-X_{t'})\}]\le\frac{v\,\mathrm{d}(t,t')^2\lambda^2}{2}, \qquad (101)$$

for all $t,t'\in\mathcal{T}$ and all $\lambda>0$. Then

$$\left|\sup_{t\in\mathcal{T}}|X_t|\right|_2\le\sqrt{2}|X_{t_0}|_2+$$

$$3\sqrt{2v}\left(\theta+2\sqrt{2}\int_0^{\theta/2}\sqrt[4]{H(u,\mathcal{T})}\,\mathrm{d}\,u+4\int_0^{\theta/2}\sqrt{H(u,\mathcal{T})}\,\mathrm{d}\,u\right).$$

Proof Define $\mathcal{T}_0:=\{t_0\}$ and, given an integer $i\ge1$, let $\theta_i:=\theta2^{-i}$, $\mathcal{T}_i$ be a θ_i-net for $\mathcal{T}$ with cardinality $N(\theta_i,\mathcal{T})$ and $\Pi_i:\mathcal{T}\to\mathcal{T}_i$ the metric projection associated to d, that is, for any $t\in\mathcal{T}$, $\Pi_i(t)\in\operatorname{argmin}_{t'\in\mathcal{T}_i}\mathrm{d}(t,t')$. By the definition of a net, we have, for all $t\in\mathcal{T}$ and $i\ge1$,

$$\mathrm{d}(t,\Pi_i(t))\le\theta_i. \qquad (102)$$

For any $t \in \mathcal{T}$, we have, since $\lim_{i\to\infty} \Pi_i(t) = t$ and $\Pi_0(t) = t_0$,

$$X_t = X_{\Pi_0(t)} + \sum_{j=0}^{\infty} (X_{\Pi_{i+1}(t)} - X_{\Pi_i(t)}). \tag{103}$$

From (103) we obtain

$$X_t^2 \leq 2X_{t_0}^2 + 2\sum_{i=0}^{\infty}\sum_{k=0}^{\infty} |X_{\Pi_{i+1}(t)} - X_{\Pi_i(t)}||X_{\Pi_{k+1}(t)} - X_{\Pi_k(t)}|,$$

and hence

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in\mathcal{T}} X_t^2\right] &\leq 2\mathbb{E}[X_{t_0}^2] + 2\sum_{i=0}^{\infty}\sum_{k=0}^{\infty} \mathbb{E}\left[\sup_{t\in\mathcal{T}} |X_{\Pi_{i+1}(t)} - X_{\Pi_i(t)}||X_{\Pi_{k+1}(t)} - X_{\Pi_k(t)}|\right] \\
&\leq 2\mathbb{E}[X_{t_0}^2] + 2\sum_{i=0}^{\infty}\sum_{k=0}^{\infty} \left|\sup_{t\in\mathcal{T}} |X_{\Pi_{i+1}(t)} - X_{\Pi_i(t)}|\right|_2 \cdot \left|\sup_{t\in\mathcal{T}} |X_{\Pi_{k+1}(t)} - X_{\Pi_k(t)}|\right|_2 \\
&= 2\mathbb{E}[X_{t_0}^2] + 2\left(\sum_{i=0}^{\infty} \left|\sup_{t\in\mathcal{T}} |X_{\Pi_{i+1}(t)} - X_{\Pi_i(t)}|\right|_2\right)^2, \qquad (104)
\end{aligned}$$

using Hölder's inequality in the second inequality. Observe that

$$|\{(\Pi_i(t), \Pi_{i+1}(t)) : t \in \mathcal{T}\}| \leq N(\theta_{i+1}, \mathcal{T})^2 = e^{2H(\theta_{i+1})}. \tag{105}$$

Note that the triangular inequality implies that, for all $t \in \mathcal{T}$,

$$\mathrm{d}(\Pi_i(t), \Pi_{i+1}(t)) \leq 3\theta_{i+1}. \tag{106}$$

By assumption (101) and Theorem 3 we also have that

$$\ln \mathbb{E}[\exp\{\lambda(X_t - X_{t'})^2\}] \leq v\,\mathrm{d}(t,t')^2\lambda + \frac{v^2\,\mathrm{d}(t,t')^4\lambda^2}{(1 - 2v\,\mathrm{d}(t,t')^2\lambda)} \tag{107}$$

for all $t, t' \in \mathcal{T}$ and all $0 < \lambda < \frac{1}{2v\,\mathrm{d}(t,t')^2}$. Relations (105)-(107) and Lemma 3 imply

$$\begin{aligned}
&\mathbb{E}\left[\sup_{t\in\mathcal{T}} |X_{\Pi_{i+1}(t)} - X_{\Pi_i(t)}|^2\right] \leq \\
&v\,\mathrm{d}(t,t')^2 + \sqrt{2v^2\,\mathrm{d}(t,t')^4 2H(\delta_{i+1},\mathcal{T})} + 2v\,\mathrm{d}(t,t')^2 2H(\delta_{i+1},\mathcal{T}) \\
&\leq 3^2\theta_{i+1}^2 v + 2\cdot 3^2\theta_{i+1}^2 v\sqrt{H(\theta_{i+1},\mathcal{T})} + 4\cdot 3^2\theta_{i+1}^2 vH(\theta_{i+1},\mathcal{T}), \qquad (108)
\end{aligned}$$

so that taking the square root in (108) we get

$$\left|\sup_{t\in\mathcal{T}} |X_{\Pi_{i+1}(t)} - X_{\Pi_i(t)}|\right|_2 \leq 3\theta_{i+1}\sqrt{v} + 3\theta_{i+1}\sqrt{2v}\sqrt[4]{H(\theta_{i+1},\mathcal{T})} + 6\theta_{i+1}\sqrt{v}\sqrt{H(\theta_{i+1},\mathcal{T})}. \tag{109}$$

Taking the square root in (104), using (109) and summing over i we finally get

$$\begin{aligned}
&\left|\sup_{t\in\mathcal{T}} |X_t|\right|_2 \leq \sqrt{2}|X_{t_0}|_2 + 3\sqrt{2v}\sum_{i=1}^{\infty}\theta_i + \\
&6\sqrt{v}\sum_{i=1}^{\infty}\theta_i\sqrt[4]{H(\theta_i,\mathcal{T})} + 6\sqrt{2v}\sum_{i=1}^{\infty}\theta_i\sqrt{H(\theta_j,\mathcal{T})} \\
&\leq \sqrt{2}|X_{t_0}|_2 + 3\sqrt{2v}\theta + \\
&12\sqrt{v}\int_0^{\theta/2}\sqrt[4]{H(u,\mathcal{T})}\,\mathrm{d}\,u + 12\sqrt{2v}\int_0^{\theta/2}\sqrt{H(u,\mathcal{T})}\,\mathrm{d}\,u,
\end{aligned}$$

using the fact that $H(u,\mathcal{T})$ is a nonincreasing function of u in the last inequality.

We prove next another preliminary lemma.

Lemma 16 *Suppose that Assumptions 2 and 4 hold. Consider Definition* (98)*. Then, for any $p \geq 2$, there exists a constant $C_{\delta,p} > 0$ (depending on δ and p), such that*

$$|Z|_p \leq \frac{C_{\delta,p}\left(|L(\xi)|_p + \left|\widehat{L}_N\right|_2\right)\sqrt{n}R^\delta}{\sqrt{N}},$$

where $\widehat{L}_N := \sqrt{\frac{1}{N}\sum_{j=1}^N L(\xi_j)^2}$.

Proof We recall the definitions (98)-(100) as well as the definitions of M and $\widehat{\sigma}^2$ in Theorem 2. The statement in the lemma will follow by bounding $\mathbb{E}[Z]$, $|M|_p$ and $\widehat{\sigma}$ and applying Theorem 2. In the sequel, C_δ is a constant (depending on δ) that might change from line to line.

PART 1: Bound of $\widehat{\sigma}^2$: We have

$$\begin{aligned}\widehat{\sigma}^2 &= \sup_{(u,\alpha)\in\mathcal{T}} \mathbb{E}\left[\frac{1}{N^2}\sum_{j=1}^N \langle \epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*), \alpha\rangle^2\right] \\ &\leq \sup_{(u,\alpha)\in\mathcal{T}} \left\{\frac{1}{N}\mathbb{E}\left[\widehat{L}_N^2\right] R^{2\delta}\|u\|^{2\delta}\|\alpha\|\right\} \\ &= \frac{1}{N}\left|\widehat{L}_N\right|_2^2 R^{2\delta}, \end{aligned} \tag{110}$$

using Assumption 4 in the inequality and the fact that $(u,\alpha) \in B[0,1]$ in the last equality.

PART 2: Bound of M:

$$\begin{aligned}|M|_p^p &= \mathbb{E}\left[\left(\max_{j\in[N]}\sup_{t\in\mathcal{T}}|X_{j,t}|\right)^p\right] = \mathbb{E}\left[\max_{j\in[N]}\sup_{t\in\mathcal{T}}|X_{j,t}|^p\right] \\ &\leq \frac{1}{N^p}\sum_{j=1}^N \mathbb{E}\left[\sup_{t\in\mathcal{T}}|\langle\epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*), \alpha\rangle|^p\right] \\ &\leq \frac{1}{N^p}\sum_{j=1}^N \mathbb{E}\left[L(\xi_j)^p R^{p\delta}\sup_{t\in\mathcal{T}}\|u\|^{p\delta}\|\alpha\|^p\right] \\ &= \frac{1}{N^{p-1}}\mathbb{E}\left[L(\xi)^p\right] R^{p\delta}, \end{aligned} \tag{111}$$

using Assumption 4 in the first inequality and the fact that $(u,\alpha) \in B[0,1]$ in the last equality. We take the p-th root in (111) and note that for $p \geq 2$ we have $N^{\frac{p-1}{p}} \geq \sqrt{N}$, obtaining

$$|M|_p \leq \frac{|L(\xi)|_p R^\delta}{\sqrt{N}}. \tag{112}$$

PART 3: Bound of $\mathbb{E}[Z]$: It is sufficient to prove that for some constant $C_\delta > 0$,

$$\mathbb{E}[Z] \leq \frac{C_\delta\left|\widehat{L}_N\right|_2\sqrt{n}R^\delta}{\sqrt{N}}. \tag{113}$$

We proceed to prove (113).

From Theorem 5, given $t = (u,\alpha) \in \mathcal{T}$ and $t = (u',\alpha') \in \mathcal{T}$, there exists constant $C > 0$ and random variable $V \geq 0$ such that for all $\lambda > 0$,

$$\mathbb{P}\left\{\left|\sum_{j=1}^N (X_{j,t} - X_{j,t'})\right| \geq C\sqrt{V(1+\lambda)}\right\} \leq e^{-\lambda}, \tag{114}$$

where

$$V := \mathbb{E}\left[\sum_{j=1}^N \left((X_{j,t} - X_{j,t'}) - (Y_{j,t} - Y_{j,t'})\right)^2 \middle| \mathcal{H}_N\right],$$

and $Y_{j,t} := \frac{1}{N}\left\langle \epsilon(\eta_j, x_* + Ru) - \epsilon(\eta_j, x_*), \alpha\right\rangle$ with $\{\eta_j\}_{j=1}^N$ an i.i.d. sample of ξ independent of $\{\xi_j\}_{j=1}^N$ and $\mathcal{H}_N := \sigma(\xi_j : j \in [N])$. Now, in view of Assumption 4 and the fact that $\mathbb{E}[X_{j,t}] = 0$ for all $t \in \mathcal{T}$ and all $j \in [N]$, it is not difficult to verify that for some constant $C_\delta > 0$ (depending on δ), it holds that

$$V \leq \frac{C_\delta}{N} Y_N^2 R^{2\delta} \left(\|u - u'\|^{2\delta} + \|\alpha - \alpha'\|^2\right), \tag{115}$$

where $\widehat{L}_N := \sqrt{\frac{1}{N}\sum_{j=1}^N L(\xi_j)^2}$ and $Y_N := \sqrt{\widehat{L}_N^2 + \mathbb{E}[\widehat{L}_N^2]}$.

For any $t = (u, \alpha) \in \mathcal{T}$ and $t = (u', \alpha') \in \mathcal{T}$, we define $\mathrm{d}(t, t') := \|u - u'\|^\delta + \|\alpha - \alpha'\|$. From (114)-(115) we have that for all $t, t' \in \mathcal{T}$ and all $\lambda > 0$,

$$\mathbb{P}\left\{|X_t - X_{t'}| \geq C_\delta \frac{Y_N}{\sqrt{N}} R^\delta\, \mathrm{d}(t, t')\sqrt{1 + \lambda}\right\} \leq e^{-\lambda}. \tag{116}$$

We conclude from (116) and Theorem 4 that for some constant $C_\delta > 0$, for all $t, t \in \mathcal{T}$ and $\lambda > 0$,

$$\ln \mathbb{E}\left[\exp\left\{\frac{(X_t - X_{t'})}{Y_N}\lambda\right\}\right] \leq C_\delta^2 \frac{R^{2\delta}\, \mathrm{d}(t, t')^2}{2N}\lambda^2. \tag{117}$$

Setting $t_0 := (0, 0)$ and $\theta := \sup_{t \in \mathcal{T}} \mathrm{d}(t, 0) \leq 2$, from (117) and Lemma 15 we get that

$$\left|\sup_{t \in \mathcal{T}}\left|\frac{X_t}{Y_N}\right|\right|_2 \leq 3\sqrt{2v}\left(2 + 2\sqrt{2}\int_0^1 \sqrt[4]{H(u, \mathcal{T})}\,\mathrm{d}\,u + 4\int_0^1 \sqrt{H(u, \mathcal{T})}\,\mathrm{d}\,u\right), \tag{118}$$

with $\sqrt{v} := C_\delta \frac{R^\delta}{\sqrt{N}}$. Finally, applying Hölder's inequality, we get

$$\mathbb{E}[Z] = \mathbb{E}\left[\sup_{t \in \mathcal{T}}|X_t|\right] = \mathbb{E}\left[\sup_{t \in \mathcal{T}}\left|\frac{X_t}{Y_N}\right| \cdot Y_N\right] \leq \left|\sup_{t \in \mathcal{T}}\left|\frac{X_t}{Y_N}\right|\right|_2 \cdot |Y_N|_2. \tag{119}$$

From Lemma 4, we also have the following bounds for $\mathcal{T} = \mathbb{Q}[0, 1]^2$:

$$\int_0^1 \sqrt[4]{H(u, \mathcal{T})}\,\mathrm{d}\,u \lesssim \sqrt[4]{n} \lesssim \sqrt{n}, \tag{120}$$

$$\int_0^1 \sqrt{H(u, \mathcal{T})}\,\mathrm{d}\,u \lesssim \sqrt{n}. \tag{121}$$

Relations (118)-(121) prove (113), and then (110)-(113), together with Theorem 2, establish the result of the lemma.

We now are ready for the proof of Lemma 6.

Proof of Lemma 6.

Proof In the following $C_{L,p}$ is a constant (depending on $L(\xi)$ and p) which might change from line to line.

Set $\widetilde{z}^N := \Pi\left[x^N - \beta\left(T(x^N) + \widehat{\epsilon}(\xi^N, x^N)\right)\right]$, so that $z^N = \alpha_N \widetilde{z}^N + (1 - \alpha_N)x^N$. By Lemma 1(iv), we have that $x^* = \Pi[x^* - \beta T(x^*)]$. Taking into account this fact and Lemma 1(iii), we get

$$\begin{aligned}\|x^* - \widetilde{z}^N\| &\leq \|x^* - x^N - \beta(T(x^*) - T(x^N)) + \beta\widehat{\epsilon}(\xi^N, x^N)\| \\ &\leq \|x_* - x^N\| + \beta\|T(x^N) - T(x^*)\| + \beta\|\widehat{\epsilon}(\xi^N, x^N)\| \\ &\leq (1 + L\beta)\|x^* - x^N\| + \beta\left\|\widehat{\epsilon}(\xi^N, x^N)\right\|,\end{aligned} \tag{122}$$

using the Lipschitz continuity of T in the last inequality.

In the sequel we use the notation $\epsilon^N := \|\widehat{\epsilon}(\xi^N, x^N)\|$. Also, for given $x \in \mathbb{R}^n$ and $s > 0$, we denote $\mathsf{R}(x, s) := (1 + L\beta)\|x - x^*\| + \beta s$, while $\mathsf{B}[x^*; x, s]$ denotes the closed Euclidean ball centered at x^* with radius $\mathsf{R}(x, s)$. From (122) we have $\widetilde{z}^N \in \mathsf{B}[x^*; x^N, \epsilon^N]$ and trivially $x^N \in \mathsf{B}[x^*; x^N, \epsilon^N]$. Hence, by convexity, $z^N \in \mathsf{B}[x^*; x^N, \epsilon^N]$, since $\alpha_N \in [0, 1]$.

Set $s_N := \|x^N - x^*\|$. We use the decomposition:

$$\left|\|\widehat{\epsilon}(\xi^N, z^N)\|\,|\mathcal{F}_N\right|_p = \left|\|\widehat{\epsilon}(\xi^N, z^N)\|\mathbb{I}_{\{\epsilon^N \leq s_N\}}|\mathcal{F}_N\right|_p$$

$$+\left|\|\widehat{\epsilon}(\xi^N,z^N)\|\mathbb{I}_{\{\epsilon^N>s_N\}}|\mathcal{F}_N\right|_p =: I_1+I_2. \tag{123}$$

We first bound the term I_1. Recall that by Remark 1,

$$\omega \mapsto \sup_{x'\in\mathsf{B}[x^*;x^N(\omega),s_N(\omega)]}\|\widehat{\epsilon}(\xi^N(\omega),x')\|$$

is a measurable function. We have

$$\begin{aligned}
I_1 &= \left|\|\widehat{\epsilon}(\xi^N,z^N)\|\mathbb{I}_{\{\epsilon^N\le s_N\}}|\mathcal{F}_N\right|_p\\
&\le \left|\sup_{x'\in\mathsf{B}[x^*;x^N,s_N]}\|\widehat{\epsilon}(\xi^N,x')\|\middle|\mathcal{F}_N\right|_p\\
&= \left|\left|\sup_{x'\in\mathsf{B}[x^*;x,s]}\|\widehat{\epsilon}(\xi^N,x')\|\right|_p\right|_{(x,s)=(x^N,s_N)},
\end{aligned} \tag{124}$$

using the fact $z^N\in\mathsf{B}[x^*;x^N,\epsilon^N]$ in the first inequality, and the fact that $(x^N,s_N)\in\mathcal{F}_N$, together with the independence between ξ^N and $\mathcal{F}_N$, in the last equality. We also have, for any $x\in\mathbb{R}^n$ and $s>0$,

$$\begin{aligned}
\left|\sup_{x'\in\mathsf{B}[x^*;x,s]}\|\widehat{\epsilon}(\xi^N,x')\|\right|_p &\le \left|\sup_{x'\in\mathsf{B}[x^*;x,s]}\|\widehat{\epsilon}(\xi^N,x')-\widehat{\epsilon}(\xi^N,x^*)\|\right|_p+\left|\|\widehat{\epsilon}(\xi^N,x^*)\|\right|_p\\
&\le \sqrt{\frac{n}{N}}C_{L,p}\mathsf{R}(x,s)+\sqrt{\frac{n}{N}}C_p\sigma(x^*),
\end{aligned} \tag{125}$$

using Lemmas 5 and 16 in the second inequality. Relations (124)-(125) and the definitions of s_N and $\mathsf{R}(x^N,s_N)$ imply that

$$I_1\le\sqrt{\frac{n}{N}}C_{L,p}(1+\beta+L\beta)\|x^N-x^*\|+\sqrt{\frac{n}{N}}C_p\sigma(x^*). \tag{126}$$

We bound next the term I_2. Defining $\widetilde{L}(\xi^N):=N^{-1}\sum_{j=1}^N L(\xi_j)$, we note that

$$\begin{aligned}
\|\widehat{\epsilon}(\xi^N,z^N)\| &\le \|\widehat{\epsilon}(\xi^N,z^N)-\widehat{\epsilon}(\xi^N,x^*)\|+\|\widehat{\epsilon}(\xi^N,x^*)\|\\
&\le \left\|\frac{1}{N}\sum_{j=1}^N\left[F(\xi_j,z^N)-F(\xi_j,x^*)\right]\right\|+\|T(z^N)-T(x^*)\|+\|\widehat{\epsilon}(\xi^N,x^*)\|\\
&\le \left(\widetilde{L}(\xi^N)+L\right)\|z^N-x^*\|+\|\widehat{\epsilon}(\xi^N,x^*)\|\\
&\le \left(\widetilde{L}(\xi^N)+L\right)(1+L\beta)\|x^N-x^*\|+\beta\left(\widetilde{L}(\xi^N)+L\right)\epsilon^N+\|\widehat{\epsilon}(\xi^N,x^*)\|,
\end{aligned} \tag{127}$$

using Assumption 4 in the third inequality and the fact that $z^N\in B[x^*;x^N,\epsilon^N]$ in the last inequality. From (127) we have

$$\begin{aligned}
I_2 &= \left|\|\widehat{\epsilon}(\xi^N,z^N)\|\mathbb{I}_{\{\epsilon^N>s_N\}}|\mathcal{F}_N\right|_p\\
&\le (1+L\beta)\|x^N-x^*\|\left|\left(\widetilde{L}(\xi^N)+L\right)\mathbb{I}_{\{\epsilon^N>s_N\}}|\mathcal{F}_N\right|_p\\
&\quad+\beta\left|\left(\widetilde{L}(\xi^N)+L\right)\epsilon^N|\mathcal{F}_N\right|_p+\left|\|\widehat{\epsilon}(\xi^N,x^*)\||\mathcal{F}_N\right|_p\\
&\le (1+L\beta)\|x^N-x^*\|\left|\widetilde{L}(\xi^N)+L\right|_{2p}\cdot\left|\mathbb{I}_{\{\epsilon^N>s_N\}}|\mathcal{F}_N\right|_{2p}\\
&\quad+\beta\left|\widetilde{L}(\xi^N)+L\right|_{2p}\cdot\left|\epsilon^N|\mathcal{F}_N\right|_{2p}+\left|\|\widehat{\epsilon}(\xi^N,x^*)\|\right|_p,
\end{aligned} \tag{128}$$

using the fact that $x^N\in\mathcal{F}_N$ in the first inequality and Hölder's inequality, together with the fact that ξ^N is independent of $\mathcal{F}_N$, in the last inequality.

With respect to the last term in the rightmost expression of (128), we have, in view of Lemma 5,

$$\left|\|\widehat{\epsilon}(\xi^N,x^*)\|\right|_p\le\sqrt{\frac{n}{N}}C_p\sigma(x^*). \tag{129}$$

Concerning the third term in the rightmost expression of (128), we observe that Lemma 5, the fact that $x^N \in \mathcal{F}_N$, and the independence between ξ^N and $\mathcal{F}_N$ imply

$$\left|\epsilon^N|\mathcal{F}_N\right|_{2p} = \left|\|\widehat{\epsilon}(\xi^N, x^N)\||\mathcal{F}_N\right|_{2p} \leq \sqrt{\frac{n}{N}}C_{2p}\sigma(x^*)(1+\|x^N - x^*\|). \tag{130}$$

Finally, concerning the second term in the rightmost expression of (128), we have, in view of the fact that $(x^N, s_N) \in \mathcal{F}_N$ and the independence between ξ^N and $\mathcal{F}_N$,

$$\left|\mathbb{I}_{\{\epsilon^N > s_N\}}|\mathcal{F}_N\right|_{2p} = \sqrt[2p]{\mathbb{E}\left[\mathbb{I}_{\{\|\widehat{\epsilon}(\xi^N, x^N)\| > s_N\}}\middle|\mathcal{F}_N\right]} = \sqrt[2p]{\mathbb{P}\left(\|\widehat{\epsilon}(\xi^N, x)\| > s\right)}\Big|_{(x,s)=(x^N, s_N)}. \tag{131}$$

From Markov's inequality we obtain

$$\begin{aligned} \sqrt[2p]{\mathbb{P}\left(\|\widehat{\epsilon}(\xi^N, x)\| > s\right)} &\leq \sqrt[2p]{\frac{\mathbb{E}\left[\|\widehat{\epsilon}(\xi^N, x)\|^{2p}\right]}{s^{2p}}} \\ &= \frac{\left|\|\widehat{\epsilon}(\xi^N, x)\|\right|_{2p}}{s} \leq \sqrt{\frac{n}{N}}\frac{C_{2p}}{s}\sigma(x^*)(1+\|x - x^*\|), \end{aligned} \tag{132}$$

using Lemma 5 in the last inequality. We conclude from (131) and (132) that

$$\left|\mathbb{I}_{\{\epsilon^N > s_N\}}|\mathcal{F}_N\right|_{2p} \leq \sqrt{\frac{n}{N}}\frac{C_{2p}}{s_N}\sigma(x^*)(1+\|x^N - x^*\|). \tag{133}$$

Putting together relations (128)-(130) and (133) we get

$$\begin{aligned} I_2 &\leq (1+L\beta)\frac{\|x^N - x^*\|}{s_N}\left|\widetilde{L}(\xi^N) + L\right|_{2p}\sqrt{\frac{n}{N}}C_{2p}\sigma(x^*)(1+\|x^N - x^*\|) \\ &\quad +\beta\left|\widetilde{L}(\xi^N) + L\right|_{2p}\sqrt{\frac{n}{N}}C_{2p}\sigma(x_*)(1+\|x^N - x^*\|) + \sqrt{\frac{n}{N}}C_p\sigma(x^*) \\ &\lesssim_{L,p} \sqrt{\frac{n}{N}}\sigma(x^*)(1+\|x^N - x^*\|), \end{aligned} \tag{134}$$

using the fact that $s_N = \|x^N - x^*\|$. Relations (123), (126) and (134) prove the required result.

References

1. Auslender, A. and Teboulle, M.: Interior projection-like methods for monotone variational inequalities, Mathematical Programming, **104**, 39-68 (2005)
2. Bach, F. and Moulines, E.: Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning, Advances in Neural Information Processing Systems (NIPS), conference paper (2011)
3. Boucheron, S., Lugosi G. and Massart, P.: Concentration inequalities: A Nonasymptotic Theory of Independence, Oxford University Press, Oxford (2013)
4. Chen, X., Wets, R.J-B. and Zhang, Y.: Stochastic Variational Inequalities: Residual Minimization Smoothing/Sample Average approximations, SIAM Journal on Optimization, **22**, 649-673 (2012)
5. Chen, Y., Lan, G. and Ouyang, Y. Accelerated schemes for a class of variational inequalities, *http://arxiv.org/abs/1403.4164*, preprint.
6. Facchinei, F. and Pang, J.-S.: Finite-Dimensional Variational Inequalities and Complementarity Problems, Springer, New York (2003)
7. Ferris, M.C. and Pang, J.S.: Engineering and economic applications of complementarity problems, SIAM Review, **39**(4), 669-713 (1997)
8. Gürkan, G., Özge, A.Y. and Robinson, S.M.: Sample-path solution of stochastic variational inequalities, Mathematical Programmming, **84**, 313-333 (1999)
9. Hsu, D., Kakade, S.M. and Zhang, T.: A tail inequality for quadratic forms of subgaussian random vectors, Electronic Communications in Probability, **17**, 1-6 (2012)
10. Iusem, A.N., Jofré A. and Thompson, P.: Incremental constraint projection methods for monotone stochastic variational inequalities, submitted.
11. Iusem, A.N., Jofré A., Oliveira, R. and Thompson, P.: Extragradient method with variance reduction for pseudo-monotone stochastic variational inequalities, submitted.
12. Iusem, A.N. and Svaiter, B.F.: A variant of Korpelevich's method for variational inequalities with a new search strategy, Optimization, **42**, 309-321 (1997)
13. Jiang, H. and Xu, H.: Stochastic approximation approaches to the stochastic variational inequality problem, IEEE Transactions on Automatic Control, **53**, 1462-1475 (2008)
14. Juditsky, A., Nemirovski A. and Tauvel, C.: Solving variational inequalities with stochastic mirror-prox algorithm, Stochastic Systems, **1**, 17-58 (2011)

15. Kannan, A. and Shanbhag, U.V.: Distributed computation of equilibria in monotone Nash games via iterative regularization techniques, SIAM Journal on Optimization,**22**, 1177-1205 (2012)
16. Kannan, A. and Shanbhag, U.V.: The pseudomonotone stochastic variational inequality problem: Analytical statements and stochastic extragradient schemes, American Control Conference (ACC), Portland, USA, 2930-2935 (2014)
17. Kannan, A. and Shanbhag, U.V.: The pseudomonotone stochastic variational inequality problem: analysis and optimal stochastic approximation schemes, *http://arxiv.org/pdf/1410.1628.pdf*, pre-print.
18. Khobotov, E.N.: Modifications of the extragradient method for solving variational inequalities and certain optimization problems, USSR Computational Mathematics and Mathematical Physics, **27**, 120-127 (1987)
19. Konnov, I.V.: Equilibrium Models and Variational Inequalities, Elsevier, Amsterdam (2007)
20. Korpelevich, G.M.: The extragradient method for finding saddle points and other problems, Ekonomika i Matematcheskie Metody, **12**, 747-756 (1976)
21. Koshal, J., Nedić, A. and Shanbhag U.V.: Regularized Iterative Stochastic Approximation Methods for Stochastic Variational Inequality Problems, IEEE Transactions on Automatic Control, **58**, 594-609 (2013)
22. Kushner, H.J. and Yin, G.G.: Stochastic approximation and recursive algorithms and applications, Springer, New York (2003)
23. Nemirovski, A.: Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems, SIAM Journal on Optimization, **15**, 229-251 (2004)
24. Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A.: Robust stochastic approximation approach to stochastic programming, SIAM Journal on Optimization, **19**, 1574-1609 (2009)
25. Panchenko, D.: Symmetrization approach to concentration inequalities for empirical processes, The Annals of Probability, **1**, 2068-2081 (2003)
26. Robbins, H. and Monro, S.: A Stochastic Approximation Method, The Annals of Mathematical Statistics, **22**, 400-407 (1951)
27. Robbins, H. and Siegmund, D.O.: A convergence theorem for non negative almost supermartingales and some applications, Optimizing methods in statistics, Academic Press, New York, 233-257 (1971)
28. Rockafellar, R.T. and Wets, R.J-B.: Variational Analysis, Springer, Berlin (1998)
29. Shapiro, A., Dentcheva, D. and Ruszczynski, A.: Lectures on Stochastic Programming: Modeling and Theory, SIAM, Philadelphia (2009)
30. Solodov, M.V. and Svaiter, B.F.: A new projection method for monotone variational inequality problems, SIAM Journal on Control and Optimization, **37**, 765-776 (1999)
31. Wang, M. and Bertsekas, D.P.: Incremental Constraint Projection-Proximal Methods for Nonsmooth Convex Optimization, Lab. for Information and Decision Systems Report LIDS-P-2907, MIT (2013)
32. Wang, M. and Bertsekas, D.P.: Incremental Constraint Projection Methods for Variational Inequalities, Mathematical Programming, **150**(2), 321-363 (2015)
33. Wang, Y.J., Xiu, N.H. and Wang, C.Y.: Unified Framework of Extragradient-Type Methods for Pseudomonotone Variational Inequalities, Journal of Optimization Theory and Applications, **111**, 641-656 (2001)
34. Yousefian, F., Nedić, A. and Shanbhag, U.V.: Distributed adaptive steplength stochastic approximation schemes for cartesian stochastic variational inequality problems, *http://arxiv.org/abs/1301.1711*, pre-print.
35. Yousefian, F., Nedić, A. and Shanbhag, U.V.: Optimal robust smoothing extragradient algorithms for stochastic variational inequality problems, *http://arxiv.org/abs/1403.5591*, pre-print.
36. Yousefian, F., Nedić, A. and Shanbhag, U.V.: On Smoothing, Regularization and Averaging in Stochastic Approximation Methods for Stochastic Variational Inequalities, *http://arxiv.org/abs/1411.0209*, pre-print.