ALGORITHMS FOR STOCHASTIC VARIATIONAL INEQUALITIES: CONVERGENCE AND COMPLEXITY ANALYSIS.

Philip Thompson

December 2016

I dedicate this thesis to my parents.
To their unconditional love, incentive and patience.

If you hit a wrong note,
it's the next one that you play
that determines if it's good or bad.

MILES DAVIS

Acknowledgements

I warmly and honestly feel very fortunate to thank three persons with whom I have learned and collaborated during my PhD studies. I thank my two advisors, Alfredo and Alejandro, and Roberto, my teacher and collaborator. I thank them for their patience, friendship and to help me engage in this beautiful world where optimization, variational analysis and mathematical statistics meet.

I thank IMPA for my mathematical studies and financial support. I acknowledge IMPA and CMM for providing an excellent and inspirational research environment where this thesis was conducted. I am grateful for CNPq and CMM for providing financial support during my PhD degree. I also warmly thank the staff of IMPA and CMM for their professional work (and daily coffee, which is essential to generate theorems).

I thank my parents, my grandparents and family for their love and support.

I thank Marcus Reache for the friendship, rewarding conversations and common enthusiasm for stochastic optimization and statistics. I thank Caio, Felipe and Vanderson for the friendship and sharing years of living together. It was a pleasurable experience. I thank Braulio and Rômulo for their creativeness and the fun of playing music together. I warmly thank my aunt Brigitte Madeleine for her deep support. My best wishes and love to her. Warm thanks to Luiz Paulo, Jorge Neri, Rafael Monteiro, Gracie, Rodrigo, Cynthia and all the friends and people I met in IMPA, Rio and Chile!

Abstract:

Stochastic approximation methods are well established for optimization problems. The appeal of these methods is due largely to their ability to cope efficiently and robustly with inexact information about the underlying optimization problem. This thesis proposes stochastic approximation methods for the solution of stochastic variational inequalities, paying attention to asymptotic convergence (stability), convergence rate, oracle complexity, knowledge of problem parameters, data availability and distributed solution. In chapter 3, we propose a method that combines stochastic approximation with incremental constraint projections, meaning that, at each iteration, the random operator is sampled and a component of the intersection defining the feasible set is chosen at random. Our method allows the distributed solution of Cartesian stochastic variational inequalities with partial coordination between users of a network. Such sequential scheme is well suited for applications involving large data sets, online optimization and distributed learning. We analyse this method for the class of weak-sharp monotone operators (without regularization) and for the class of plain monotone operators with regularization. In chapter 4, we propose a stochastic extragradient method for pseudo-monotone operators with a novel iterative variance reduction procedure. We present convergence and complexity analysis relaxing previous assumptions used for stochastic approximation and accelerating the convergence rate while maintaining a near-optimal oracle complexity. Our extragradient method is also suitable for the distributed case. In chapter 5, we propose two stochastic extragradient methods with linear search with the same set of assumptions as in chapter 4, except that we do not require the knowledge of the Lipschitz constant or Lipschitz continuity.

Keywords: Stochastic approximation, randomized algorithms, stochastic variational inequalities, incremental methods, extragradient method, variance-reduction, weak-sharpness, Tykhonov regularization.

Resumo:

Métodos de aproximação estocástica já são bem estabelecidos para otimização. Uma vantagem destes métodos é a habilidade de lidar eficientemente de forma robusta com informações inexatas sobre o problema de otimização em questão. Esta tese propõe métodos de aproximação estocástica para a solução de desigualdades variacionais estocásticas, com atenção para convergência assintótica (estabilidade), taxa de convergência, complexidade do oráculo, conhecimento de parâmetros do problema, disponibilidade de dados e solução distribuída. No capítulo 3, é proposto um método que combina aproximação estocástica com projeções incrementais, significando que, a cada iteração, o operador aleatório é amostrado e uma das componentes da interseção definindo o conjunto viável é escolhida aleatoriamente. Este método pode ser usado na solução distribuída de desigualdades variacionais Cartesianas com coordenação parcial entre usuários de uma rede. Este é um esquema sequencial adequado para problemas de alta dimensão, otimização online e aprendizagem distribuída. Este método é analizado para a classe de operadores monótonos weak-sharp (sem regularização) e para operadores apenas monótonos com regularização. No capítulo 4, é proposto um método extragradiente estocástico para operadores pseudo-monótonos usando um novo procedimento de redução de variância iterativa. É apresentado resultados de convergência e complexidade, relaxando-se hipóteses usadas anteriormente em aproximação estocástica e acelerando-se a convergência. Este método também é adequado para a solução distribuída. No capítulo 5, são propostos dois métodos extragradiente estocásticos com busca linear usando-se as mesmas hipóteses do capítulo 4, exceto que não é requerido o conhecimento da constante de Lipschitz ou a continuidade Lipschitz.

Palavras-chave: Aproximação estocástica, algoritmos randomizados, desigualdades variacionais estocásticas, métodos incrementais, método extragradiente, redução de variância, weak-sharpness, regularização de Tykhonov.

Contents

1 Introduction **8**

1.1 The stochastic variational problem and the sampling methodology . 8

1.2 Accessing the feasible set and distributed solution 11

1.3 Projection methods 13

1.4 Stochastic approximation methods 17

1.5 Contributions of the thesis 19

1.5.1 Incremental methods [36] 19

1.5.2 Stochastic extragradient methods [37, 38] 21

2 Preliminaries **29**

2.1 Projection operator and notation 29

2.2 Probabilistic tools 33

2.3 Weak-sharpness 34

3 Stochastic incremental constraint projection methods **38**

3.1 An incremental projection method under weak sharpness 38

3.1.1 Statement of the algorithm 39

3.1.2 Discussion of the assumptions 40

3.1.3 Convergence analysis 43

3.1.4 Convergence rate analysis 47

3.2 An incremental projection method with Tykhonov regularization . . 60

3.2.1 Cartesian structure 60

3.2.2 Constraint structure 60

3.2.3 Statement of the algorithm 61

3.2.4 Discussion of the assumptions 62

3.2.5 Convergence analysis 64
3.3 Appendix of Chapter 3 72

4 A variance-based stochastic extragradient method 74
4.1 Discussion of the assumptions 76
4.2 Convergence analysis 83
4.3 Convergence rate and complexity analysis 90
4.3.1 Comparison of complexity estimates 103
4.4 Appendix of Chapter 4 106

5 Stochastic extragradient methods with line search 111
5.1 An empirical process theory for DS-SA line search schemes 116
5.1.1 The $\mathcal{L}^2$-norm of suprema of sub-Gaussian processes 118
5.1.2 Heavy-tailed Hölder continuous operators: self-normalization and sup-norms 122
5.1.3 The proof of Theorem 12 128
5.2 Analysis of Algorithm 5 for Lipschitz continuous operators 133
5.2.1 Derivation of an error bound 135
5.2.2 Bound on oracle error 140
5.2.3 Asymptotic convergence, convergence rate and oracle complexity 141
5.3 Analysis of Algorithm 6 for Hölder continuous operators 147
5.4 Discussion on the complexity constants of Algorithm 5 154

6 Conclusions and open questions 158

Bibliography 160

Chapter 1

Introduction

1.1 The stochastic variational problem and the sampling methodology

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

$$\langle T(x^*), x - x^* \rangle \geq 0. \tag{1.1}$$

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

$$T(x^*) = 0.$$

When $X = \mathbb{R}^n_+$, the VI problem becomes the *complementarity problem*, i.e., find $x^* \in \mathbb{R}^n$ such that

$$0 \leq x^* \perp T(x^*) \geq 0.$$

See Section 1.5 of [27]. Se also Section 1.4 of [27] and [28] for an excellent review on the applications of such formulations, which include fields such as engineering (e.g., mechanics problems, structural design problems, obstacle problems, traffic equilibrium problems, optimal control) and economics (e.g., General equilibria and game theory). The complementarity problem and systems of equations are important classes of problems where the feasible set is *unbounded*. The *saddle-point problem* is the problem

$$\min_{y \in Y} \max_{z \in Z} f(y, z),$$

for $Y \times Z \subset \mathbb{R}^m \times \mathbb{R}^n$ and $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$. This problem can be cast as an a VI under suitable conditions with

$$T(y, z) = [\nabla_y f(y, z) \quad -\nabla_z f(y, z)]$$

and $X = Y \times Z$. The saddle-point problem is an equivalent formulation of the *zero-sum Nash game*. Relevant equilibrium problems which can be formulated as a VI include: the *Nash equilibrium*, the *Wardrop traffic equilibrium* and the *General Economic Equilibrium*. See, for example, [41] and Section 1.4 of [27].

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 $\mathbf{P} := \mathbb{P}_\xi$ of ξ, that is, $\mathbf{P}(A) = \mathbb{P}(\xi \in A)$ for any measurable $A \in \mathcal{G}$. 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 thesis consists of solving VI(T, X) as defined by (1.1), where $T : \mathbb{R}^n \to \mathbb{R}^n$ is the expected value of $F(\xi, \cdot)$, i.e.,

$$T(x) = \mathbb{E}[F(\xi, x)], \quad \forall x \in \mathbb{R}^n. \tag{1.2}$$

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

Definition 1 (SVI)**.** *Under the setting of* (1.2)*, 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 [31], as a natural generalization of stochastic optimization problems

(SP). Recently, a more general definition of stochastic variational inequality was considered in [19, 69] where the feasible set is also affected by randomness, that is, $X : \Xi \rightrightarrows \mathbb{R}^n$ is a random set-valued map. This setting appears, e.g., in economical or traffic equilibrium problems where an uncertain demand is present in the constraints.

Methods for the deterministic VI(T, X) have been extensively studied (see [27]). 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 (1.2), unavailability of $\mathbb{P}_\xi$ or absence of a closed form for $F(\xi, \cdot)$. This situation 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). It is important to remark that the framework in (1.2) includes the relevant discrete case where T is a prohibitively *large* sum of operators, that is,

$$T(x) = \sum_{i=1}^{S} T_i(x), \forall x \in \mathbb{R}^n,$$

with $S \gg 1$. Computing the above operator is computationally prohibitive. Problems in this framework require methods which have the ability to make progress by examining only a small fraction of the data set rather than scanning it entirely - an operation that is too expensive for "big-data" modern applications, e.g., machine learning, stochastic equilibrium problems and empirical risk minimization.

Depending on how sampling is incorporated with the algorithm, solution methods for SVIs can be classified into two basic categories. The first category consists of the stochastic approximation (SA) methods, which perform sampling in an "interior" manner, by applying an algorithm for deterministic VIs and resorting to sampling whenever the algorithm requires values of the operator at given points. In that respect, SA-typed methods are *explicit* methods in the sense that a direct (deterministic) algorithm is used along the stochastic oracle calls. The second category corresponds to sample average approximation (SAA) methods, which sample in an "exterior" manner. These methods replace the mean operator T by the empirical average operator to obtain the SAA problem, and then use a solution to the SAA problem as an estimate of a solution to the true problem. SAA-typed

methods are *implicit* methods in the sense that the *sampled* problem is solved by means of a deterministic method of preferred choice. In this thesis we focus on the SA approach. For analysis of the SAA methodology for SP and SVI, see e.g., [31, 70] 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 [66]. 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 (1.2), almost surely find $x^*(\xi) \in \mathbb{R}^n$ such that $T(x^*(\xi)) = 0$. The SA methodology has been applied to SVI in [40], [42], [75], [50], [73], [36], [37], [20], [76], [44], [45], [77]. 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 a stochastic error $F(\xi, x) - T(x)$ for $x \in X$ in the trajectory of the method. See also [51], [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).

1.2 Accessing the feasible set and distributed solution

A frequent additional difficulty is the possibly complicated structure of the *feasible set* X. Often, the feasible set takes the form

$$X = \cap_{i \in \mathcal{I}} X_i,$$

where $\{X_i : i \in \mathcal{I}\}$ is an arbitrarily family of closed convex sets. There are different motivations for considering the design of algorithms which, at every iteration, use only a component X_i rather than the whole feasible set X. First, in the case of projection methods, when the orthogonal projection onto each X_i, namely $\Pi_i : \mathbb{R}^n \to X_i$, is much easier to compute than projection onto X, namely $\Pi : \mathbb{R}^n \to X$, a natural idea consists of replacing, at iteration k, Π by one of the Π_i's, say Π_{i_k}, or even by an approximation of Π_i. This occurs, for instance, when X is a polyhedron

and the X_i's are halfspaces. This procedure is the basis of the so called sequencial or parallel *row action methods* for solving systems of equations (see [17]) and methods for the *feasibility problem*, useful in many applications, including image restoration and tomography (see, e.g., [5] and [16]). Second, in some cases X is not known a priori, but is rather revealed through the random realizations of its components X_i in time through a learning process. Such problems currently arise in fair rate allocation problems in wireless networks where the channel state is unknown but the channel states X_i are observed in time (see e.g. [57] and [34]). Third, in some cases X is known but the number of constraints is prohibitively very large (e.g., in machine learning and signal processing).

In *Cartesian variational inequalities*, a network of m agents is associated to a coupled variational inequality with constraint set

$$X = X^1 \times \cdots \times X^m$$

and operator

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

where the i-th agent is associated to a constraint set $X^i \subset \mathbb{R}^{n_i}$ and a map $F_i : \Xi \times \mathbb{R}^n \to \mathbb{R}^{n_i}$ such that

$$n = \sum_{i=1}^{m} n_i.$$

Relevant problems which are included in the above setting are stochastic *Nash equilibrium* (SNE) problems and stochastic *multi-agent optimization* problems. In the SNE, for $i = 1, \ldots, m$, $X^i \subset \mathbb{R}^{n_i}$ is closed and convex, and the problem consists of finding, almost surely, a point $x^* = (x_1^*, \ldots, x_m^*)$ such that for all $i \in \{1, \ldots, m\}$, x_i^* solves the optimization problem

$$\min_{x_i \in X^i} \mathbb{E}[f_i(\xi, x_1, \ldots, x_i, \ldots, x_m)].$$

The equilibrium conditions of SNE can be formulated as a SVI with

$$F(\xi, x) := (\nabla_{x_1} f_1(\xi, x), \ldots, \nabla_{x_m} f_m(\xi, x))$$

and $X = X^1 \times \cdots \times X^m$. The stochastic multi-agent optimization problem is the problem

$$\min_{x \in X} \sum_{i=1}^{m} \mathbb{E}[f_i(\xi, x)],$$

with the additional constraint that the i-th user has only access to its objective f_i and constraints X^i when deciding its variable $x_i \in X^i$. This problem is reformulated as an SVI with $X := X^1 \times \cdots \times X^m$ and

$$F(\xi, \cdot) := (\nabla_{x_1} f(\xi, \cdot), \ldots, \nabla_{x_m} f(\xi, \cdot)),$$

where $f(\xi, \cdot) := \sum_{i=1}^m f_i(\xi, \cdot)$. In these mentioned problems, the i-th agent has only access to constraint set X^i and operator F_i (which depends on other agents' decisions) so that a *distributed solution* of the SVI is required in *large networks*. An important class of distributed methods for Cartesian VI's are designed so that agents update their stepsizes *independently* under some partial coordination, since communication along a large network is costly and requiring the constraint that the agents use exactly the same stepsize or additional parameters can be a non-robust requirement (see [43]). As an example, the distributed variant of the classical projection method studied in [75] takes the form: for all $i = 1, \ldots, m$,

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

where Π_i is the Euclidean projection onto X^i (see also [43]).

1.3 Projection methods

In the deterministic setting (1.1), the classical projection method for $\mathrm{VI}(T, X)$, akin to the projected gradient method for convex optimization, is

$$x^{k+1} = \Pi[x^k - \alpha_k T(x^k)], \tag{1.3}$$

where Π is the projection operator onto X and $\{\alpha_k\}$ is an exogenous sequence of positive stepsizes. Convergence of this method is guaranteed assuming T is strongly monotone, Lipschitz continuous and the stepsizes satisfy $\alpha_k \in (0, 2\sigma/L^2)$ and $\inf_k \alpha_k > 0$, where $\sigma > 0$ is the modulus of strong monotonicity and L is the Lipschitz constant, see e.g. [27].

The strong monotonicity assumption is too demanding in some applications, and convergence of (1.3) is not guaranteed when the operator is just monotone. In order to deal with this situation, the following extragradient algorithm was

proposed by Korpelevich [49]:

$$\begin{aligned} z^k &= \Pi[x^k - \alpha_k T(x^k)], \\ x^{k+1} &= \Pi[x^k - \alpha_k T(z^k)], \end{aligned} \tag{1.4}$$

in which an additional auxiliary projection step is introduced. Convergence of the method is guaranteed when the stepsizes satisfy $\alpha_k \equiv \alpha \in (0, 1/L)$. In [59], the extragradient method was generalized and convergence rates were established assuming compactness of the feasible set.

The next step relevant in applications is to relax the *knowledge* of the Lipschitz constant or even the Lipschitz continuity in the extragradient method (1.4). In [47], Khobotov proposed the following linear stepsize search for the extragradient method: given iterate x^k, the stepsize α_k is chosen as the maximum $\alpha \in \{\theta^j \hat{\alpha} : j \in \mathbb{N}_0\}$ such that

$$\alpha \left\| T(z^k(\alpha)) - T(x^k) \right\| \leq \lambda \|z^k(\alpha) - x^k\|, \tag{1.5}$$

where for all $\alpha > 0$, $z^k(\alpha) := \Pi\left[x^k - \alpha T(x^k)\right]$. In the above line search, $\hat{\alpha} > 0$, $\theta \in (0, 1)$ and $\lambda > 0$ are exogenous parameters.

The above extragradient method with line search does not use the Lipschitz constant but requires Lipschitz continuity. It also requires as many projection computations as the number of iterations in the line search. In [39], Iusem and Svaiter proposed the hyperplane projection method in which a different line search is introduced based on the geometric interpretation of separating the current iterate and the solution set by a hyperplane. An advantage is that only continuity and two projections per iteration are required. It takes the form: given iterate x^k, take α_k as the maximum $\alpha \in \{\theta^j \hat{\alpha} : j \in \mathbb{N}_0\}$ such that

$$\left\langle T(\bar{z}^k(\alpha)), x^k - \Pi(g^k) \right\rangle \geq \frac{\lambda}{\beta_k} \|x^k - \Pi(g^k)\|^2, \tag{1.6}$$

where $g^k := x^k - \beta_k T(x^k)$ 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 $x^{k+1} := \Pi\left[x^k - \gamma_k T(z^k)\right]$, where

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

Again, $\hat{\alpha} \in (0, 1]$, $\theta \in (0, 1)$, $\{\beta_k\}$ and $\lambda > 0$ are parameters to be defined a priori. It is not difficult to see that $x^{k+1} = \Pi\left[\Pi_{H_k}(x^k)\right]$ where H_k is the hyperplane

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

See also [71, 74, 1] for improvements.

Observe that the projection method (1.3) and the extragradient method (1.4) are *explicit*, i.e., the formula for obtaining x^{k+1} is an explicit one, up to the computation of the orthogonal projection Π. An *implicit* approach for the solution of monotone variational inequalities is through a Tykhonov or proximal regularization scheme (see [27], Chapter 12). In these methods, a sequence of regularized variational inequality problems are approximately solved at each iteration.

As commented before, a typical case occurs when the feasible set takes the form $X = \cap_{i=1}^{m} X_i$, where all the X_i's are closed and convex. Row action methods and alternate (or cyclic) projection algorithms for convex feasibility problems exploit the computation of projections onto the components iteratively (see [3]). In such case, the order in which the sets X_i are used along the iterations, i.e. the so called *control sequence* $\{\omega_k\} \subset \{1, \ldots, m\}$, must be specified. Several options have been considered in the literature (such as cyclic control, almost cyclical control, most violated constraint control and random control). A negative consequence of the use of approximate projections is the need to use *small stepsizes*, i.e., satisfying $\sum_k \alpha^2 < \infty$ and $\sum_k \alpha_k = \infty$, which significantly reduces the rate of convergence of the method, despite improving the access to the constraints or computational complexity (in terms of projection calculations). We thus have a trade-off between easier projection computation and slower convergence. Data availability is costly in large-scale applications and often a solution with poorer quality but with easy computation is the best available option.

The use of approximate projections requires some condition on the feasible set, so that the projections onto the sets X_i's are reasonable approximations of the projection onto X. For this, some form of error bound, linear regularity or Slater-type conditions on the sets X_i must be assumed (e.g., Assumption 7 in Chapter 3 and the comments following it). See [4, 23]. Explicit methods for monotone variational inequalities using approximate projections were studied e.g. in [30] and [18], imposing rather demanding coercivity assumptions on T, in [6] assuming paramonotonicity of T, and then in [8] assuming just monotonicity of T. Another method of this type, using an Armijo search as in [39] for determining the stepsizes, and approximate projections with the most violated constraint control, can be found in [7].

Related to row-action and alternate projective methods are the so called *incremental* methods, introduced by Kibardin in [46] (see also [52, 9, 57] and references therein). These methods are used for the minimization of a *large* sum of convex functions, e.g. in machine learning applications. In such a context, instead of using the gradient of the sum, the gradient of one of the terms is selected iteratively under different control rules. In [63, 57], *incremental constraint* methods with random control rules were proposed for minimizing a convex function over an intersection of a *large* number of convex sets. If the feasible set takes the form

$$X = X_0 \cap (\cap_{i\in\mathcal{I}} X_i)\,, \tag{1.7}$$

where $\{X_0\} \cup \{X_i : i \in \mathcal{I}\}$ is a collection of closed and convex subsets of $\mathbb{R}^n$ and for every $i \in \mathcal{I}$,

$$X_i = \{x \in \mathbb{R}^n : g_i(x) \leq 0\}, \tag{1.8}$$

for some convex function g_i with positive part $g_i^+(x) := \max\{g_i(x), 0\}$, then the method in [57] is given by

$$y^k = \Pi_{X_0}\left[x^k - \alpha_k \nabla f(x^k)\right], \tag{1.9}$$

$$x^{k+1} = \Pi_{X_0}\left[y^k - \beta_k \frac{g^+_{\omega_k}(y^k)}{\|d^k\|^2} d^k\right], \tag{1.10}$$

where $d^k \in \partial g^+_{\omega_k}(y^k) - \{0\}$ if $g^+_{\omega_k}(y^k) > 0$, and $d^k = d$ for any $d \in \mathbb{R}^n - \{0\}$ if $g^+_{\omega_k}(y^k) = 0$. In method (1.9)-(1.10), $\{\omega_k\}$ is a random control sequence taking values in $\mathcal{I}$ and satisfying certain conditions and $f : \mathbb{R}^n \to \mathbb{R}$ is a convex smooth function (the non-smooth case is also analysed). In (1.9), the method takes a step towards the gradient direction, while in (1.10), the method takes a step towards feasibility. Together with row-action and alternate projection methods, incremental constraint projection methods can be viewed as the dual version of (standard) incremental methods. More recently, stochastic approximation was incorporated in incremental constraint projections methods for stochastic convex minimization problems in [72].

1.4 Stochastic approximation methods

The first SA method for SVI was analyzed by Jiang and Xu in [40]. Their method is:

$$x^{k+1} = \Pi[x^k - \alpha_k F(\xi^k, x^k)], \tag{1.11}$$

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

$$\mathbb{E}\left[\|F(\xi, x) - T(x)\|^2\right] \leq \sigma^2. \tag{1.12}$$

After the above mentioned work, recent research on SA methods for SVI have been developed in [42, 75, 50, 73, 36, 20, 76, 44, 45, 77, 37]. 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 analyse the monotone case, SA methodologies based on the extragradient method of Korpelevich [49] and the mirror-prox algorithm of Nemirovski [59], and iterative Tykhonov and proximal regularization procedures (see [43]), were used in the above mentioned works. Other objectives were the use of incremental constraint projections in the case of difficulties when accessing the feasible set in [73], the convergence analysis in the absence of the Lipschitz constant in [77, 75, 76], and the distributed solution of Cartesian variational inequalities in [75, 50].

We finalize this section commenting on recent methods that our proposals in Chapter 3 improve upon.

In [73], method (1.11) is improved by incorporating an incremental projection scheme, instead of exact ones. They take $X = \cap_{i \in \mathcal{I}} X_i$, where $\mathcal{I}$ is a finite index set, and use a random control sequence, where *both* the random map F and the control sequence $\{\omega_k\}$ are jointly sampled, giving rise to the following algorithm:

$$\begin{aligned} y^k &= x^k - \alpha_k F(\xi^k, x^k) \\ x^{k+1} &= y^k - \beta_k (y^k - \Pi_{\omega_k}(y^k)). \end{aligned} \tag{1.13}$$

When $\beta_k \equiv 1$, the method is the version of method (1.11) with incremental constraint projections. For convergence, the operator is assumed to be *strongly monotone* and Lipschitz-continuous. In this setting, method (1.13) improves on method (1.9)-(1.10) for the particular case of $X_0 = \mathbb{R}^n$, $\mathcal{I}$ finite and $\{X_i : i \in \mathcal{I}\}$ with easy projections.

In [50], regularized *iterative* Tychonov and proximal point methods for monotone stochastic variational inequalities were introduced. In such methods, instead of solving a sequence of regularized variational inequality problems, the regularization parameter is updated in each iteration and a *single projection step* associated with the regularized problem is taken. This is desirable since (differently from the deterministic case), termination criteria are generally hard to meet in the stochastic setting. The algorithm proposed allows for a Cartesian structure on the variational inequality, so as to encompass, for example, equilibrium conditions of monotone stochastic Nash games with a limited coordination between the player's stepsize and regularization sequences. Namely, the feasible set $X \subset \mathbb{R}^n$ has the the form $X = X^1 \times \cdots \times X^m$, where each Cartesian component $X^j \subset \mathbb{R}^{n_j}$ is a closed and convex set, and the operator has components $F = (F_1, \ldots, F_m)$ with $F_j : \Xi \times \mathbb{R}^n \to \mathbb{R}^{n_j}$ for $j = 1, \ldots, m$ and $\sum_{j=1}^m n_j = n$. The algorithm in [50] is thus described as follows: given the k-th iterate $x^k \in X$ with components $x_j^k \in X^j$, for $j = 1, \ldots, m$, the next iterate is given by the projection

$$x_j^{k+1} = \Pi_{X^j}[x_j^k - \alpha_{k,j}(F_j(\xi^k, x^k) + \epsilon_{k,j} x_j^k)], \tag{1.14}$$

for $i = 1, \ldots, m$, where $\{\alpha_{k,1}, \ldots, \alpha_{k,m}\}$ are the stepsize sequences and $\{\epsilon_{k,1}, \ldots, \epsilon_{k,m}\}$ are the regularization parameter sequences. This method is shown to converge under monotonicity and Lipschitz-continuity of T and a partial coordination between the stepsize and regularization parameter sequences (see Assumption 11 in Section 3.2). The iterative proximal point follows a similar pattern but differently from the Tykhonov method, this method requires strict monotonicity, which in particular implies uniqueness of solutions.

1.5 Contributions of the thesis

The contributions of this thesis are summarized in the following and corresponds to the papers [36, 37, 38].

1.5.1 Incremental methods [36]

As commented before, accessing data in modern large-scale problems is a challenge. In such cases, the use of stochastic approximation for computing the operator and/or easily computable approximate projections instead of exact ones is a preferred option (or even the only one). Additionally, in many cases the constraint set X is known but it contains a very large number of constraint components or X is not known a priori, but is rather learned along time through random samples of its constraint components. An important feature of incremental constraint projection methods is that they process sample operators and sample constraints *sequentially*. This incremental structure is well suited for a variety of applications involving large data sets, online optimization and distributed learning. For big-data problems, an incremental method can update simultaneously as passing through the data set. For problems that require online learning, incremental projection methods of the type (1.9)-(1.10) or (1.13) are practically the only option to use without the knowledge of all the constraints.

In collaboration with Iusem, A. and Jofré, A., we propose, in Chapter 3, methods which incorporate the incremental constraint projection method with the stochastic approximation to compute the operator. One of our main goals is to weaken the strong monotonicity assumed in [73] by plain *monotonicity*. Another important goal is to incorporate the *distributed* solution of stochastic Cartesian variational inequalities in the framework of incremental constraint projection methods. The incorporation of incremental constraint projection methods to distributed solution of network equilibria seems to be new.

We thus propose an incremental constraint projection method for monotone SVIs. Precisely, assuming the structures (1.7)-(1.8), in the centralized case ($m =$

1), the method takes the form

$$y^k = \Pi_{X_0}\left[x^k - \alpha_k\left(F(v^k, x^k) + \epsilon_k x^k\right)\right], \tag{1.15}$$

$$x^{k+1} = \Pi_{X_0}\left[y^k - \beta_k \frac{g^+_{\omega_k}(y^k)}{\|d^k\|^2} d^k\right], \tag{1.16}$$

where $\{\omega_k\}$ is the random control, and $d^k \in \partial g^+_{\omega_k}(y^k) - \{0\}$ if $g_{\omega_k}(y^k) > 0$; $d^k = d \in \mathbb{R}^n - \{0\}$ if $g_{\omega_k}(y^k) \leq 0$. In Section 3.1 of Chapter 3, this method is analyzed with no regularization, i.e., $\epsilon^k \equiv 0$ and the monotone operator satisfies the *weak sharpness* property (see Section 2.3 of Chapter 2) while in Section 3.2 of Chapter 3, we consider the same method with positive regularization parameters without assuming weak sharpness. Just for simplicity, the distributed case ($m \gg 1$) is analysed only for the second variant (but the case $m \gg 1$ could be generalized in an obvious way to the case when the SVI has the weak sharpness property).

We mention the following contributions of method (1.15)-(1.16):

(i) Excepting for method (1.13) for strongly monotone stochastic variational inequalities, all the above mentioned works on stochastic approximation for SVI use *exact* projections. We generalize (1.13) by analyzing the case of infinite number of constraints, $X_0 \neq \mathbb{R}^n$, and incremental constraints for a larger class of closed convex sets. For instance, as in method (1.9)-(1.10), our method allows the components X_i of X to have difficult projection operators, as long as they take the form $X_i = \{x \in \mathbb{R}^n : g_i(x) \leq 0\}$ where g_i is a convex function with computable subgradients. We also extend the incremental projection framework to the class of weak-sharp monotone operators (without regularization) and to plainly monotone operators (with required regularization), thus extending [57, 9, 72] to the framework of (stochastic) variational inequalities. Differently from [57], we cope with unbounded operators. Under weak sharpness, we prove that the sequence is bounded in L^2 and give explicit estimates on the convergence rate $O(1/\sqrt{k})$ up to logarithm terms. Sharper estimates are possible for bounded operators.

(ii) Method (1.15)-(1.16) is a variation of (1.14) with incremental projections. We also incorporate the *distributed* solution, which appears to be new in the setting of incremental projection methods. Due to the use of approximate

projections instead of exact ones, an additional coordination requirement is imposed, which is satisfied by usual choices of stepsizes and regularization parameters (see Section 3.2.4 of Chapter 3, Assumption 11 and comments following it and [50], Lemma 4).

(iii) It seems that the use of *weak sharpness* as a suitable property for incremental projections is new. In fact, differently from the strongly monotone case in [73], where the convergence rate deteriorates with the use of incremental projections, we prove that, under weak sharpness, the rate $O(1/\sqrt{k})$ in terms of $\mathbb{E}[\mathrm{d}(x^k, X^*)]$ is the same both with exact and incremental projections (see [45, 76]). Surprisingly, these results still hold true for the cases of unbounded operators or unbounded feasible sets, an improvement over [45, 76] where boundedness and exact projections are required alongside the weak sharpness property. We also give the exact number of iterations required for an auxiliary stochastic optimization problem over X with linear objective for solving the original SVI (which recovers a related property in the deterministic setting, see [53]).

It should be noted that under weak sharpness, our method has *robust stepsizes* in the sense of Nemirovski et al. [60], without knowledge of the Lipschitz constant or the weak-sharp modulus. In that respect, we improve upon [76], where under weak sharpness, robust stepsizes are given for exact projections, compact feasible set, strict-monotonicity, knowledge of the weak sharpness modulus and requiring a smoothing procedure and an extragradient scheme.

1.5.2 Stochastic extragradient methods [37, 38]

In Chapter 4, in collaboration with Iusem, A., Jofré, A. and Oliveira, R., we propose the following extragradient method: given x^k, define

$$z^k = \Pi\left[x^k - \frac{\alpha_k}{N_k}\sum_{j=1}^{N_k} F(\xi_j^k, x^k)\right], \tag{1.17}$$

$$x^{k+1} = \Pi\left[x^k - \frac{\alpha_k}{N_k}\sum_{j=1}^{N_k} F(\eta_j^k, z^k)\right], \tag{1.18}$$

where $\{N_k\} \subset \mathbb{N}$ is a non-decreasing sequence and $\{\xi_j^k, \eta_j^k : k \in \mathbb{N}, j = 1, \ldots, N_k\}$ are independent identically distributed (i.i.d.) samples of ξ. We call $\{N_k\}$ the *sample rate* sequence. In the sequel we need some notation. For any $\alpha > 0$ we consider the natural residual function r_α, defined, for any $x \in \mathbb{R}^n$, by $r_\alpha(x) := \|x - \Pi(x - \alpha T(x))\|$. It is well known that the set of zeroes of r_α coincides with X^* (see [27]). 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 $Q > 0$ (depending on the initial iterate and the parameters of the problem and the method). 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, i.e., $\sum_{k=1}^K 2N_k$. Next we synthetize the contributions of the algorithm presented in Chapter 3.

i) **Asymptotic-convergence**: Assuming *pseudo-monotonicity* of F, and using an extragradient scheme, without regularization, we prove that, almost surely, the generated sequence is bounded, its distance to the solution set converges to zero and its natural residual value converges to zero a.s. and in L^2. See [45] for recent examples where the more general setting of pseudo-monotonicity is relevant (stochastic fractional programming, stochastic optional pricing and stochastic economic equilibria). The sequence generated by our method also possesses a new stability feature: for $p = 2$ or any $p \geq 4$, if the random operator has finite p-moment then the sequence is bounded in L^p, and we are able to provide explicit upper bounds in terms of the problem parameters. Previous work required a bounded monotone operator, specific forms of (pseudo)-monotonicity (monotonicity with acute angle, pseudo-monotonicity-plus, strict pseudo-monotonicity, symmetric pseudo-monotonicity or strong pseudo-monotonicity as in [44, 45]), or regularization procedures. The disadvantage of regularization procedures in the absence of strong monotonicity is the need to introduce additional coordination between the stepsize sequence and the regularization parameters. Also, the regularization induces a suboptimal performance in terms of rate and complexity (see [77]).

ii) **Accelerated rate with oracle complexity efficiency**: To the best of our

knowledge, our work is the first SA method for SVI with *stepsizes bounded away from zero*. Such feature allows our method to achieve an accelerated convergence rate $O(1/K)$ in terms of the mean-squared natural residual under plain *pseudo-monotonicity* (with no regularization requirements). As a consequence, our method achieves a convergence rate of $O(1/K)$ in terms of the mean D-gap function [1] (see Subsection 4.3.1 of Chapter 4 and [27], Proposition 10.3.7). In previous works, methods with diminishing stepsizes satisfying $\sum_k \alpha_k = \infty$, $\sum_k \alpha_k^2 < \infty$ were used, achieving a $O(1/K)$ rate in terms of the mean-squared distance to X^*, with more demanding monotonicity assumptions (namely, bounded strongly pseudo-monotone operators and bounded monotone weak-sharp VI) and a rate $O(1/\sqrt{K})$ in terms of mean gap functions [2] for bounded monotone operators. Importantly, our method preserves the optimal oracle complexity $O(\epsilon^{-2})$ up to first order logarithmic term. By accelerating the rate, we reduce the number of projection computations, preserving the optimal oracle complexity. We provide explicit upper bounds for the rate and complexity in terms of the problem parameters. As a corollary of our result, we provide new classes of SVIs for which a convergence rate of $O(1/K)$ holds in terms of the mean-squared distance to the solution set (see Section 4.3.1). In the context of large dimension data ($n \gg 1$), our algorithm complexity is proportional to n up to a scaling factor in the sample rate (see Proposition 8).

iii) **Unbounded setting**: The results in items (i)-(ii) are valid for an *unbounded feasible set* and *unbounded operator*. Important examples of such a setting include complementarity problems and systems of equations. Asymptotic convergence for an unbounded feasible set is analyzed in [75, 73, 45, 77] and Section 3.1 of Chapter 3 with more demanding monotonicity hypotheses, and in [50] and Section 3.2 of Chapter 3 for the monotone case, but with an additional regularization procedure. To the best of our knowledge, convergence rates in the case of an unbounded feasible set were treated only in [73, 20]. In [73], a convergence rate is given only for strongly monotone operators. In

[1] Given $a > 0$, the *regularized gap function* is defined as $g_a(x) := \sup_{y \in X}\{\langle T(x), x-y\rangle - \frac{a}{2}\|x-y\|^2\}$, for $x \in \mathbb{R}^n$. Given $b > a > 0$, the *D-gap function* is $g_{a,b}(x) := g_a(x) - g_b(x)$, for $x \in \mathbb{R}^n$.

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

[20], assuming *uniform variance* over X (in the sense of (1.12)), a convergence rate of $O(1/\sqrt{K})$ in terms of the mean value of a relaxed gap function [3] recently introduced by Monteiro and Svaiter [55]-[56] is achieved. However, we provide in Example 1 (Chapter 4, Section 4.1), a simple case showing that *asymptotically* the method in [20] is not stable in the unbounded setting: a.s. the generated sequence has an unbounded subsequence (even though the gap function value converges in mean to zero). Our convergence analysis in items (i)-(ii) does not depend upon boundedness assumptions, and we prove the accelerated rate $O(1/K)$ in terms of the mean (quadratic) natural residual and the mean D-gap function, which are new results. The natural residual and the D-gap function are better behaved than the (standard) gap function: the former are finite valued and Lipschitz continuous over $\mathbb{R}^n$ while the later is finite valued and continuous only for a compact X.

(iv) **Non-uniform variance**: To the best of our knowledge, all previous works require that the variance of the oracle error be *uniform* over X (in the sense of (1.12)), excepting in [73] for the strongly monotone case, and in Chapter 3 for the case of a weak-sharp monotone operator, and also for the monotone case with an iterative Tykhonov regularization (with no convergence rate results). Such uniform variance assumption holds for bounded operators, but not for unbounded ones, on a unbounded feasible set. Typical situations where this assumption fails to hold include affine complementarity problems and systems of equations. In such cases, the variance of the oracle error tends (quadratically) to ∞ in the horizon (see Example 2 of Chapter 4, Section 4.1). The performance of our method, in terms of the oracle complexity, depends on the point $x^* \in X^*$ with *minimal* trade-off between variance and distance to initial iterates "ignoring" points with high variance (see comments after Theorem 10 and Section 4.3.1). This result also improves on the case where (1.12) *does* holds but $\sigma(x^*)^2 \ll \sigma^2$ or, on the case X is compact but $\|x^0 - x^*\| \ll \text{diam}(X)$. In conclusion, the performance of method (1.17)-(1.18) depends on solution points x^* with minimal variance, compared to the conservative upper bound σ^2, and minimal distance to initial iterates. In

[3] Such gap-function is 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 case of uniform variance over X^* or X, we obtain sharper estimates of rate and complexity in item (ii).

(v) **Distributed solution of multi-agent system**: The analysis in items (i)-(iv) also holds true for the distributed solution of stochastic Cartesian variational inequalities, in the spirit of [75, 50, 43]. In our framework (see Algorithm (4.3)-(4.4)), agents may update stepsizes bounded away from zero independently over the range $(0, 1/2L)$. An advantage of the extragradient approach in the distributed case is that we do not require iterative regularization procedures as in [50, 43] and the method of Section 3.2 of Chapter 3, for coping with the plain monotone case. This implies that no coordination is required between users' stepsizes and regularization parameters and an optimal convergent rate is achievable. As discussed later on, our algorithm requires the choice of a sampling rate for dealing with the setting of items (i)-(iv). Hence, in the distributed solution case, agents should have the choice of sharing their oracle calls or not, and we allow both options. In the later case of fully distributed sampling, the oracle complexity has higher order dependence in terms of the network dimension m, which may be demanding in the context of large networks ($m \gg 1$). For this case, if an estimate of m is available (up to a scaling factor in the sample rate) and a decreasing sequence of (deterministic) parameters $\{b_i\}_{i=1}^m$ is shared (in any order) among agents, then our algorithm has oracle complexity of order $m(a^{-1}\epsilon^{-1})^{2+a}$ for arbitrary $a > 0$ (see Proposition 10), that is, linear in m. Further dimension reduction possibilities will be the subject of future work.

For achieving the results of items (i)-(v), we employ an iterative *variance reduction* procedure. This means that, instead of calling the oracle once per iteration (as in previous SA methods for SVI studied so far), our method calls the oracle N_k times at iteration k and uses the associated empirical average of the values of the random operator F at the current iterates x^k and z^k (see (1.17)-(1.18)). Since the presence of the stochastic error destroys the strict Fejér property (satisfied by the generated sequence in the deterministic setting), the mentioned variance reduction procedure is the mechanism that allows our extragradient method to converge in an unbounded setting with stepsizes bounded away from zero, and to achieve an

accelerated rate in terms of the natural residual. To obtain these results, we use martingale moment inequalities and a supermartingale convergence theorem (see Section 2.2). Our sampling procedure also possesses a *robust* property: a scaling factor on the sampling rate maintains the progress of the algorithm with proportional scaling in the convergence rate and oracle complexity (see Propositions 8, 9 and 10. See also [60] for robust methods). In Examples 1 and 2 of Section 4.1 of Chapter 4, we show typical situations where such variance reduction procedure is necessary.

To the best of our knowledge the variance reduction procedure mentioned above is new for SA solution of SVI. During the preparation of this thesis we became aware of references [24, 15, 29, 32, 22], where variable sample-size methods are studied for stochastic optimization. We treat the general case of pseudo-monotone variational inequalities with weaker assumptions. Also, our analysis and assumptions differ from these works relying on martingale and optimal stopping techniques. In [24, 15, 29] the SA approach is studied for convex stochastic optimization problems. In [15, 29], the focus is on gradient descent methods applied to unconstrained strongly convex optimization problems. In [15], second order information is assumed and an adaptive sample size selection is used. In [29] uniform boundedness assumptions are required. In [24], a variant of the dual averaging method of Nesterov [61] is applied for solving non-smooth stochastic convex optimization, assuming a compact feasible set and uniform variance. A constant oracle call per iteration $N_k \equiv N > 1$ is used, obtaining a convergence rate of $O(1/\sqrt{KN})$, while we typically use $N_k = O(k(\ln k)^{1+b})$ with $b > 0$ obtaining a rate of $O(1/K)$. In [32, 22], the SAA approach for stochastic optimization is studied. This is an implicit method, unlike the SA methodology. Also, uniform boundedness assumptions are required. In [22] the focus is on unconstrained optimization, with second order information, using Bayesian analysis for an adaptive choice of N_k.

To better appreciate our variance reduction procedure in our SA extragradient method, we make some remarks regarding SAA methods. In such methods, the random variable is sampled once in an exterior manner and an *uniform* strong law of large numbers guarantees the convergence of the method to the true solution. As a consequence, X has to be compact and the performance of the method depends on the maximum variance of the oracle over X and on the diameter of X. By

exploring monotonicity along stochastic approximations, our variance reduction makes use of progressive empirical averages (with increasing accuracy) at points of the trajectory of the method. Hence, intuitively, the law of large numbers is invoked *locally* along neighbourhoods of the trajectory of the method. As a consequence, the feasible set can be unbounded and the performance of the method depends on the distance of the initial iterate to the solution set and of the variance at *points of the trajectory* and at the *solution set* only. Such type of results appear to be new. [4]

In Chapter 5, in collaboration with A. Iusem, A. Jofré and R. Oliveira, we extend the analysis of Chapter 4 by proposing stochastic extragradient methods without requiring knowledge of the Lipschitz constant or weakening the Lipschitz continuity assumption. Motivated by the results of Chapter 4, we propose two methods which are SA variants of the extragradient method with line search (1.5) and the line search (1.6) of the hyperplane projection method respectively. Again, we incorporate the variance reduction mechanism of method (1.17)-(1.18) and are able to recover the results of items (i)-(iv) presented above for method (1.17)-(1.18). To the best of our knowledge, these are the first extragradient methods with line search for SVI. The introduction of line searches for the stepsize have the objective to deal with inexistent, unknown or too large Lipschitz constant while improving over the alternative of summable stepsizes (which require a too small "stepsize" with a detrimental effect on the convergence rate). It is widely recognized that line searches substantially enhance the numerical performance of the 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 stochastic extragradient methods with line search we propose more implementable.

It should be noticed that methods which avoid the use of the Lipschitz constant or Lipschitz continuity, were proposed in [76, 77], but by means of a very different procedure. Instead of line 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

[4] We remark that, as pointed out in [60], SA methods can be competitive and even outperform SAA methods in some classes of convex problems.

cope with unbounded sets, non-uniform variance for pseudo-monotone operators and achieve the rate $O(1/K)$.

We comment on the results of methods with line search of Chapter 5. In the deterministic case, the hyperplane projection method (1.6) requires only continuity. Our stochastic variant 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). Our stochastic variant of the Khobotov's line search (1.5) requires Lipschitz continuity of the random operator, (a few more) projections per iteration [5], sample rate $N_k \sim k$ (up to logarithmic terms) and oracle complexity $O(\epsilon^{-2})$ (up to logarithmic terms). Hence, the choice between these two line 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 the second variant, i.e., Algorithm 5 of Chapter 5 (if Lipschitz continuity is available).

[5] The number of iterations in the line search is of logarithmic order on the Lipschitz constant as in the deterministic case.

Chapter 2

Preliminaries

2.1 Projection operator and notation

We shall define specific constants in every chapter with *no relation* to constants outside that chapter. 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 $\mathrm{d}(x, C) := \inf\{\|x - y\| : y \in C\}$ and $\mathcal{D}(C) := \sup\{\|x - y : x, y \in C\|\}$. Given a positive-definite symmetric matrix $A \in \mathbb{R}^{n\times n}$ and $x \in \mathbb{R}^n$, we denote by $\langle x, x\rangle_A := \langle x, Ax\rangle$ and $\|x\|_A := \sqrt{\langle x, Ax\rangle}$ the correspondent inner product and Euclidean norm. For a closed and convex set $C \subset \mathbb{R}^n$, we use the notation $\Pi_{C,A}(x) := \mathrm{argmin}_{y\in C} \|y - x\|_A^2$ for $x \in \mathbb{R}^n$. We use the simplified notation $\Pi_C := \Pi_{C,I}$ when $I \in \mathbb{R}^{n\times n}$ is the identity matrix. Given $H : \mathbb{R}^n \to \mathbb{R}^n$, $\mathrm{S}(H, C)$ denotes the solution set of $\mathrm{VI}(H, C)$. For a matrix $B \in \mathbb{R}^{n\times n}$, we use the notation $\|B\| := \sup_{x\neq 0} \|Bx\|/\|x\|$. The following properties of the projection operator are well known.

Lemma 1. *Take a closed and convex set $C \subset \mathbb{R}^n$ and a positive-definite symmetric matrix $A \in \mathbb{R}^{n\times n}$.*

i) Given $x \in \mathbb{R}^n$, $\Pi_{C,A}(x)$ is the unique point of C satisfying the property: $\langle x - \Pi_{C,A}(x), y - \Pi_{C,A}(x)\rangle_A \leq 0$, for all $y \in C$.

Moreover, let $v \in \mathbb{R}^d$ and $x \in C$ with $z := \Pi_C[x - v]$. Then, for all $u \in C$, $2\langle v, z - u\rangle \leq \|x - u\|^2 - \|z - u\|^2 - \|z - x\|^2$.

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 $H : \mathbb{R}^n \to \mathbb{R}^n$, $\mathrm{S}(H, C) = \{x \in \mathbb{R}^n : x = \Pi_{C,A}[x - A^{-1}H(x)]\}$.

v) For all $x \in C, y \in \mathbb{R}^n$, $\langle x - y, x - \Pi_C(y)\rangle \geq \|x - \Pi_C(y)\|^2$.

The following lemma will be used in the analysis of the methods of Chapter 3. It was used in [57, 63] but in a slightly different form, suitable for convex optimization problems.

Lemma 2. *Consider a closed and convex* $X_0 \subset \mathbb{R}^n$, *and let* $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *be a convex function with* $\mathrm{dom}(g) \subset X_0$. *Suppose that there exists* $C_g > 0$ *such that* $\|z\| \leq C_g$ *for all* $x \in X_0$ *and all* $z \in \partial g^+(x)$. *Take* $x_1 \in X_0$, $u \in \mathbb{R}^n$, $\alpha > 0$, $\beta \in (0, 2)$ *and* $d \in \mathbb{R}^n - \{0\}$ *such that* $d \in \partial g^+(y) - \{0\}$ *if* $g^+(y) > 0$. *Define* $y, x_2 \in X_0$ *as*

$$\begin{aligned} y &= \Pi_{X_0}[x_1 - \alpha u], \\ x_2 &= \Pi_{X_0}\left[y - \beta \frac{g^+(y)}{\|d\|^2} d\right], \end{aligned}$$

Then for any $x_0 \in X_0$ *such that* $g^+(x_0) = 0$ *and any* $\tau > 0$, *it holds that*

$$\|x_2 - x_0\|^2 \leq \|x_1 - x_0\|^2 - 2\alpha\langle x_1 - x_0, u\rangle + [1 + \tau\beta(2 - \beta)]\,\alpha^2\|u\|^2 - \frac{\beta(2-\beta)}{C_g^2}\left(1 - \frac{1}{\tau}\right)\left(g^+(x_1)\right)^2.$$

Proof. We shall first give an upper bound of $\|x_2 - x_0\|$ in terms of $\|y - x_0\|$. Precisely, we shall prove:

$$\|x_2 - x_0\|^2 \leq \|y - x_0\|^2 - \beta(2 - \beta)\frac{\left(g^+(y)\right)^2}{C_g^2}. \tag{2.1}$$

Suppose first $g^+(y) > 0$ with $d \in \partial g^+(y) - \{0\}$. In this case we have

$$\begin{aligned} \|x_2 - x_0\|^2 &= \left\|\Pi_{X_0}\left[y - \beta\frac{g^+(y)}{\|d\|^2}d\right] - \Pi_{X_0}[x_0]\right\|^2 \\ &\leq \left\|y - \beta\frac{g^+(y)}{\|d\|^2}d - x_0\right\|^2 \\ &= \|y - x_0\|^2 - 2\beta\frac{g^+(y)}{\|d\|^2}\langle y - x_0, d\rangle + \beta^2\frac{\left(g^+(y)\right)^2}{\|d\|^2}, \end{aligned} \tag{2.2}$$

where in first equality we used $x_0 \in X_0$ and in the inequality we used Lemma 1(iii). Since g is convex on X_0, so is g^+. Since $x_0, y \in X_0$, $g^+(x_0) = 0$ and $d \in \partial g^+(y)$, the definition of subgradient implies $0 = g^+(x_0) \geq g^+(y) + \langle x_0 - y, d\rangle$, or equivalently:

$$-2\beta \frac{g^+(y)}{\|d\|^2} \langle y - x_0, d\rangle \leq -2\beta \frac{(g^+(y))^2}{\|d\|^2}. \tag{2.3}$$

Relations (2.2)-(2.3) and $-\beta(2-\beta)/\|d\|^2 \leq -\beta(2-\beta)/C_g^2$ imply relation (2.1) which was claimed. Suppose now $g^+(y) = 0$ and $d \neq 0$. In this case, $x_2 = \Pi_{X_0}[y] = y$ and hence $\|x_2 - x_0\|^2 = \|y - x_0\|^2$, so that (2.1) holds trivially with equality.

We now will relate $g^+(y)$ with $g^+(x_1)$. We have

$$\begin{aligned} \left(g^+(y)\right)^2 &= \left[\left(g^+(y) - g^+(x_1)\right) + g^+(x_1)\right]^2 \\ &= \left(g^+(y) - g^+(x_1)\right)^2 - 2\left(g^+(y) - g^+(x_1)\right) g^+(x_1) + \left(g^+(x_1)\right)^2 \\ &\geq -2\left|g^+(y) - g^+(x_1)\right| g^+(x_1) + \left(g^+(x_1)\right)^2. \end{aligned} \tag{2.4}$$

Using that $x_1, y \in X_0$ and the definition of C_g we have

$$\left|g^+(y) - g^+(x_1)\right| \leq C_g \|y - x_1\| = C_g \left\|\Pi_{X_0}\left[x_1 - \alpha u\right] - \Pi_{X_0}[x_1]\right\| \leq C_g \alpha \|u\|,$$

where in last inequality we used Lemma 1(iii). Multiplying the previous relation by $2g^+(x_1)$ gives

$$\begin{aligned} 2\left|g^+(y) - g^+(x_1)\right| g^+(x_1) &\leq 2C_g \alpha \|u\| g^+(x_1) \\ &\leq \tau C_g^2 \alpha^2 \|u\|^2 + \frac{1}{\tau}\left(g^+(x_1)\right)^2, \end{aligned} \tag{2.5}$$

for arbitrary $\tau > 0$. In last inequality above we used relation $2ab \leq \tau a^2 + \frac{1}{\tau} b^2$ for any $\tau > 0$. Relations (2.4) and (2.5) imply that for any $\tau > 0$,

$$-\left(g^+(y)\right)^2 \leq \tau C_g^2 \alpha^2 \|u\|^2 - \left(1 - \frac{1}{\tau}\right)\left(g^+(x_1)\right)^2, \tag{2.6}$$

which is the desired relation between $g^+(y)$ and $g^+(x_1)$.

From (2.1) and (2.6) we obtain for every $\tau > 0$,

$$\|x_2 - x_0\|^2 \leq \|y - x_0\|^2 + \tau\beta(2-\beta)\alpha^2\|u\|^2 - \frac{\beta(2-\beta)}{C_g^2}\left(1 - \frac{1}{\tau}\right)\left(g^+(x_1)\right)^2. \tag{2.7}$$

We also have

$$\begin{aligned} \|y - x_0\|^2 &= \|\Pi_{X_0}[x_1 - \alpha u] - \Pi_{X_0}[x_0]\|^2 \\ &\leq \|(x_1 - x_0) - \alpha u\|^2 \\ &= \|x_1 - x_0\|^2 - 2\alpha\langle x_1 - x_0, u\rangle + \alpha^2\|u\|^2, \end{aligned} \tag{2.8}$$

where we used Lemma 1(iii) in the inequality. Relations (2.7)-(2.8) prove the claim. □

Remark 1. We remark that if $\mathrm{dom}(g) = \mathbb{R}^n$ and the subgradients of g^+ are uniformly bounded over $\mathbb{R}^n$, then the result of Lemma 2 holds with $y \in \mathbb{R}^n$ given as $y = x_1 - \alpha u$, instead of $y = \Pi_{X_0}[x_1 - \alpha u]$.

We denote by $\mathbb{R}^n_{>0}$ the interior of the nonnegative orthant $\mathbb{R}^n_+$. We use the notation $[m] := \{1, \ldots, m\}$ for $m \in \mathbb{N}$ and $(\alpha_i)_{i=1}^m := (\alpha_1, \ldots, \alpha_m)$ for $\alpha_i \in \mathbb{R}$ and $i \in [m]$. For $\alpha := (\alpha_i)_{i=1}^m \in \mathbb{R}^m_{>0}$, $D(\alpha)$ denotes the block-diagonal matrix in $\mathbb{R}^{n\times n}$ defined as

$$D(\alpha) := \begin{bmatrix} \alpha_1 I_{n_1} & 0 & 0 \\ & \ddots & \\ 0 & 0 & \alpha_m I_{n_m} \end{bmatrix},$$

where $I_{n_i} \in \mathbb{R}^{n_i \times n_i}$ denotes the identity matrix for each $i \in [m]$.

We will use the following lemma, which is proved in the Appendix of Chapter 4, in the distributed solution of Cartesian SVIs. Assume that $C = \prod_{i=1}^m C^i$ and $n = \sum_{i=1}^m n_i$, where $C^i \subset \mathbb{R}^{n_i}$ is closed and convex for $i \in [m]$. We endow C with the inner product $\langle x, y\rangle = \sum_{i=1}^m \langle x_i, y_i\rangle$ for $x = (x_i)_{i=1}^m$ and $y = (y_i)_{i=1}^m$ in C. Consider the operator $H = (H_1, \ldots, H_m)$ with $H_i : \mathbb{R}^n \to \mathbb{R}^{n_i}$ for $i \in [m]$.

Lemma 3. *For any $\alpha \in \mathbb{R}^m_{>0}$, $\mathrm{S}(D(\alpha) \cdot H, C) = \mathrm{S}(H, C)$.*

In the case of the feasible set X as in (1.1), 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

$$r_\alpha(H; x) := \|x - \Pi[x - \alpha H(x)]\|. \tag{2.9}$$

In the case of the operator T as in (1.1), 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 $r := r_1$.

When $\alpha \in \mathbb{R}^m_{>0}$, we will also use the notation $r_\alpha(x) := \|x - \Pi\left[x - D(\alpha)T(x)\right]\|$ for $x \in \mathbb{R}^n$.

We shall also use the following useful lemma (see [27], Proposition 10.3.6).

Lemma 4. *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. We denote by $\mathrm{cov}[B]$ the covariance of a random vector B. Also, we write $\xi \in \mathcal{F}$ for "ξ is $\mathcal{F}$-measurable". We denote by $\sigma(\xi_1, \ldots, \xi_k)$ the σ-algebra generated by the random variables $\xi_1, \ldots, \xi_k$. Given the random variable ξ and $p \geq 1$, $|\xi|_p$ is the L^p-norm of ξ and $|\xi\,|\mathcal{F}|_p := \sqrt[p]{\mathbb{E}\left[|\xi|^p\,|\mathcal{F}\right]}$ is the L_p-norm of ξ conditional to the σ-algebra $\mathcal{F}$. $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. Given $x \in \mathbb{R}$, we denote by $x_+ := \max\{0, x\}$ its positive part and by $\lceil x \rceil$ the smallest integer greater than x. For a function $g : \mathbb{R}^n \to \mathbb{R}^n$ we denote by g^+ its positive part, defined by $g^+(x) = \max\{0, g(x)\}$ for $x \in \mathbb{R}^n$. If g is convex, we denote by ∂g its subdifferential and $\mathrm{dom}(g)$ its domain. For a matrix $B \in \mathbb{R}^{n\times n}$, B^T denotes its transpose, $\|B\|$ denotes its spectral norm, $\mathrm{tr}(B)$ denotes its trace and, if B is symmetric positive semidefinite, we denote its square root matrix by $\sqrt{B}$.

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 [67], 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 $\sum u_k < \infty$.*

We will also use the following result, whose proof can be found in Lemma 10 of [64].

Theorem 2. *Let $\{y_k\}, \{a_k\}, \{b_k\}$ be sequences of nonnegative random variables, adapted to the filtration $\{\mathcal{F}_k\}$, such that a.s. $a_k \in [0,1]$, $\sum a_k = \infty$, $\sum b_k < \infty$, $\lim_{k\to\infty} \frac{b_k}{a_k} = 0$ and for all $k \in \mathbb{N}$, $\mathbb{E}\Big[y_{k+1}\Big|\mathcal{F}_k\Big] \leq (1-a_k)y_k + b_k$. Then a.s. $\{y_k\}$ converges to zero.*

For the next result see [13, 54]

Theorem 3 (Burkholder-Davis-Gundy inequality in $\mathbb{R}^d$)**.** *Let $\|\cdot\|$ be the Euclidean norm in $\mathbb{R}^d$. Then, for all $q \geq 2$, there exists $C_q > 0$ such that for any vector-valued martingale $\{y_j\}_{j=0}^N$ adapted to the filtration $\{\mathcal{G}_j\}_{j=1}^N$ with $y_0 = 0$, it holds that*

$$\left|\sup_{j\leq N} \|y_j\|\right|_q \leq C_q \left|\sqrt{\sum_{j=1}^N \|y_j - y_{j-1}\|^2}\right|_q \leq C_q \sqrt{\sum_{j=1}^N \left|\|y_j - y_{j-1}\|\right|_q^2}.$$

Some more probabilistic tools will be needed and presented in Chapter 5.

2.3 Weak-sharpness

We briefly discuss the *weak sharpness* property of variational inequalities which is used in Section 3.1 of Chapter 3 as an useful property for incremental constraint projection methods.

For $X \subset \mathbb{R}^n$ and $x \in X$, $\mathbb{N}_X(x)$ denotes the normal cone of X at x, given by

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

The tangent cone of X at $x \in X$ is defined as

$$\mathbb{T}_X(x) = \{d \in \mathbb{R}^n : \exists t_k > 0, \exists d^k \in \mathbb{R}^n, \forall k \in \mathbb{N}, x + t_k d^k \in X, d^k \to d\}. \tag{2.10}$$

For a closed and convex set X, the tangent cone at a point $x \in X$ has the following alternative representation (see Rockafellar and Wets [68], Proposition 6.9 and Corollary 6.30):

$$\mathbb{T}_X(x) = \mathrm{cl}\{\alpha(y - x) : \alpha > 0, y \in X\} = [\mathbb{N}_X(x)]^\circ, \tag{2.11}$$

where for a given set $Y \subset \mathbb{R}^n$, the polar set Y° is defined as $Y^\circ = \{v \in \mathbb{R}^n : \langle v, y\rangle \leq 0, \forall y \in Y\}$.

In [14], the notion of *weak sharp* minima for the problem $\min_{x\in X} f(x)$ with solution set X^* was introduced: there exists $\rho > 0$ such that

$$f(x) - f^* \geq \rho\, \mathrm{d}(x, X^*), \tag{2.12}$$

for all $x \in X$, where f^* is the minimum value of f at X. Relation (2.12) means that $f - f^*$ gives an error bound on the solution set X^*. In [22], it is proved that if f is a closed, proper, and differentiable convex function and if the sets X and X^* are nonempty, closed, and convex, then (2.12) is equivalent to the geometric condition: for all $x^* \in X^*$,

$$-\nabla f(x^*) \in \operatorname{int}\left(\bigcap_{x\in X^*} [\mathbb{T}_X(x) \cap \mathbb{N}_{X^*}(x)]^\circ\right). \tag{2.13}$$

Every linear program is weak-sharp. Also, the minimization problem associated to non-degenerate linear complementarity problems is weak-sharp [14]. Piecewise affine functions possess "corners" which are potentially weak-sharp minima.

In optimization problems, the objective function can be used for determining regularity of solutions. In variational inequalities one can use for that purpose the above geometric definition or exploit the use of gap functions associated to the VI. The dual gap function $G : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is defined as $G(x) := \sup_{y\in X}\langle T(y), x-y\rangle$. In the sequel, we denote by $B(0,1)$ the unit ball in $\mathbb{R}^n$ and by X^* the solution set of $\mathrm{VI}(T, X)$. In order to define a meaningful notion of weak sharpness for VIs, the following possible assumptions were considered in [53]:

(i) There exists $\rho > 0$, such that for all $x^* \in X^*$,

$$-T(x^*) + \rho B(0,1) \in \bigcap_{x\in X^*} [\mathbb{T}_X(x) \cap \mathbb{N}_{X^*}(x)]^\circ. \tag{2.14}$$

(ii) There exists $\rho > 0$, such that for all $x^* \in X^*$,

$$\langle T(x^*), z\rangle \geq \rho\|z\|, \forall z \in \mathbb{T}_X(x^*) \cap \mathbb{N}_{X^*}(x^*). \tag{2.15}$$

(iii) For all $x^* \in X^*$,

$$-T(x^*) \in \operatorname{int}\left(\bigcap_{x\in X^*} [\mathbb{T}_X(x) \cap \mathbb{N}_{X^*}(x)]^\circ\right). \tag{2.16}$$

(iv) There exist $\rho > 0$ such that for all $x \in X$,

$$G(x) \geq \rho \, \mathrm{d}(x, X^*). \tag{2.17}$$

Item (iii) is the definition of a weak sharp VI(T, X) given in [53]. In Theorem 4.1 of [53], it is proved that (i)-(ii) are equivalent, and that (i)-(iv) are equivalent when X is compact and T is paramonotone (also known as monotone$^+$) i.e., T is motonone and $\langle T(x) - T(y), x - y \rangle = 0 \Rightarrow T(x) = T(y)$, for all $x, y \in \mathbb{R}^n$ (see [35] for other properties of paramonotone operators).

Relation (2.17) means that the gap function G provides an error bound on the solution set X^*. Paramonotonicity implies that T is constant on the solution set X^*. Important classes of paramonotone operators are, for example, co-coercive, symmetric monotone and strictly monotone composite operators (see [27], Chapter 2).

Recently, the following assumption was introduced in [76]: there exists $\rho > 0$ such that for all $x^* \in X^*$ and all $x \in X$,

$$\langle T(x^*), x - x^* \rangle \geq \rho \, \mathrm{d}(x, X^*). \tag{2.18}$$

Clearly, (2.18) implies (2.17). Proposition 1, proved in the Appendix of Chapter 3, states that (2.18) implies (2.15) and the converse statement holds when T is constant on X^*. Thus, when T is constant on X^*, (2.14), (2.15) and (2.18) are equivalent, and when T is paramonotone and X is compact, relations (2.14)-(2.18) are all equivalent. Hence, the following proposition, which appears to be new, gives a precise relation between property (2.18) with the previous notions of weak sharpness (2.14)-(2.17) presented in [53]. Interestingly, property (2.18) is well suited for the incremental constraint projection-type methods considered here.

Proposition 1. *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous monotone operator and $X \subset \mathbb{R}^n$ a closed and convex set. The following holds:*

i) Condition (2.18) *implies* (2.15).

ii) If T is constant on X^*, then (2.15) *implies* (2.18).

Finally, we will need the following result, established in Theorem 4.2. of [53]:

Theorem 4. *If T is continuous and there exists $z \in \mathbb{R}^n$ such that*

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

then

$$\operatorname*{argmin}_{x \in X} \langle z, x \rangle \subset X^*.$$

As a consequence of Theorem 4, under weak sharpness and uniform continuity of T, any algorithm which generates a sequence $\{x^k\}$ such that $\mathrm{d}(x^k, X^*) \to 0$ has the property that after a *finite* number of iterations M, any solution of the auxiliary program $\min_{x \in X} \langle T(x^M), x \rangle$, with a *linear* objective, is a solution of the original variational inequality (see Theorem 5.1 in [53]). When X is a polyhedron, this result can be interpreted as a finite convergence property of algorithms for VI with the weak sharpness property, since a linear program is finitely solvable. Other algorithmic implications of weak sharpness are developed in [53]. However, in practice, M may be very large. In Corollary 4 of Chapter 3 we provide an *exact value* of M in terms of the "condition number" L/ρ^2 for our first projection method under the weak sharpness assumption, assuming the operator is L-Lipschitz continuous.

Chapter 3

Stochastic incremental constraint projection methods

3.1 An incremental projection method under weak sharpness

In this chapter we assume that the feasible set has the form

$$X = X_0 \cap \left(\cap_{i\in\mathcal{I}} X_i\right), \tag{3.1}$$

where $\{X_0\} \cup \{X_i : i \in \mathcal{I}\}$ is a collection of closed and convex subsets of $\mathbb{R}^n$. We assume that the evaluation of the projection onto X_0 is computationally easy and that for all $i \in \mathcal{I}$, X_i is representable as

$$X_i = \{x \in \mathbb{R}^n : g_i(x) \leq 0\}, \tag{3.2}$$

for some convex function g_i with $\mathrm{dom}(g_i) \subset X_0$. Also we assume that, for every $i \in \mathcal{I}$, subgradients of $g_i^+(x)$ at points $x \in X_0 - X_i$ are easily computable and that $\{\partial g_i^+ : i \in \mathcal{I}\}$ is uniformly bounded over X_0, that is, there exists $C_g > 0$ such that

$$\|d\| \leq C_g, \tag{3.3}$$

for all $x \in X_0$, all $i \in \mathcal{I}$, and all $d \in \partial g_i^+(x)$.

We make the important observation that the collection $\{X_i : i \in \mathcal{I}\}$ of constraints can be *infinite* as long as (3.3) and certain regularity assumptions are satisfied (see Assumption 7 and comments following it).

3.1.1 Statement of the algorithm

In this chapter, v will denote the random variable acting in the random operator F. [1] The following incremental algorithm advances in such a way that the "operator step" and the "feasibility step" are updated in separate stages. In the first stage, given the current iterate x^k, the method advances in the direction of a sample $-F(v^k, x^k)$ of the random operator, producing an auxiliary iterate y^k. In this step, the hard constraint set X_0 is considered while the soft constraints $\{X_i : i \in \mathcal{I}\}$ are "ignored". In the second stage, a soft constraint X_{ω_k} is randomly chosen for $\omega_k \in \mathcal{I}$, and the method advances in the direction opposite to a subgradient of $g^+_{\omega_k}$ at the point y^k, producing the next iterate x^{k+1}. Thus, the method exploits simultaneously the stochastic approximation of the random operator (in the first stage) and a randomization of the incremental selection of constraint projections (in the second stage).

Algorithm 1 (Incremental constraint projection method for SVI)**.**

1. **Initialization:** *Choose the initial iterate $x^0 \in \mathbb{R}^n$, the stepsizes $\{\alpha_k\}$ and $\{\beta_k\}$, the random controls $\{\omega_k\}$ and the sample $\{v^k\}$ of v.*

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

$$y^k = \Pi_{X_0}[x^k - \alpha_k F(v^k, x^k)], \tag{3.4}$$

$$x^{k+1} = \Pi_{X_0}\left[y^k - \beta_k \frac{g^+_{\omega_k}(y^k)}{\|d^k\|^2} d^k\right], \tag{3.5}$$

where $d^k \in \partial g^+_{\omega_k}(y^k) - \{0\}$ if $g^+_{\omega_k}(y^k) > 0$; $d^k = d \in \mathbb{R}^n - \{0\}$ if $g^+_{\omega_k}(y^k) = 0$.

Before analyzing the algorithm, we present some special cases which illustrate that the mentioned framework is very general. If, for $i \in \mathcal{I}$, the Euclidean projection onto X_i is easy, then we can always construct a function satisfying (3.2)-(3.3) with "easy" subgradients. Indeed, defining the function $g_i(x) := \mathrm{d}(x, X_i)$, for $x \in \mathbb{R}^n$, then g_i satisfies (3.2), is convex, nonnegative and finite valued over $\mathbb{R}^n$ for which $\|d\| \leq 1$ for all $x \in \mathbb{R}^n$, $d \in \partial g_i(x)$. Also, for any $x \notin X_i$,

$$\frac{x - \Pi_{X_i}(x)}{g_i(x)} = \frac{x - \Pi_{X_i}(x)}{\|x - \Pi_{X_i}(x)\|} \in \partial g_i(x),$$

[1] In Chapters 1, 4 and 5, we use the notation ξ for the random variable.

provides a subgradient which is easy to evaluate. In that case, using the above directions as subgradients d^k of $g^+_{\omega_k}$ at y^k, method (3.4)-(3.5) can be rewritten as

$$\begin{aligned} y^k &= x^k - \alpha_k F(v^k, x^k), \\ x^{k+1} &= \Pi_{X_0}\left[y^k - \beta_k\left(y^k - \Pi_{X_{\omega_k}}(y^k)\right)\right]. \end{aligned}$$

In first equality above, the projection onto X_0 is not required since $\mathrm{dom}(g_i) = \mathbb{R}^n$ and $\{\partial g_i^+ : i \in \mathcal{I}\}$ is uniformly bounded in $\mathbb{R}^n$(see Remark 1). If, additionally, $X_0 = \mathbb{R}^n$ and $\beta_k \equiv 1$ then the method takes the form $x^{k+1} = \Pi_{X_{\omega_k}}\left[x^k - \alpha_k F(v^k, x^k)\right]$.

3.1.2 Discussion of the assumptions

In the sequel we consider the natural filtration

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

Next we present the assumptions necessary for our convergence analysis.

Assumption 1 (Consistency)**.** *The solution set X^* of* VI(T, X) *is nonempty.*

Assumption 2 (Monotonicity)**.** *The mean operator T in* (1.2) *satisfies: for all $y, x \in \mathbb{R}^n$,*

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

Assumption 3 (Lipschitz-continuity or boundedness)**.** *We suppose $T : \mathbb{R}^n \to \mathbb{R}^n$ is continuous and, at least, one of the following assumptions hold:*

i) There exists $L > 0$, such that a.s. for all $y, x \in \mathbb{R}^n$, $k \in \mathbb{N}$,

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

ii) There exists $C_F > 0$ such that a.s.

$$\sup_{x\in X_0}\sup_{k\in\mathbb{N}} \mathbb{E}\left[\|F(v^k, x)\|^2\Big|\mathcal{F}_k\right] \leq 2C_F^2.$$

Assumption 3(i) is satisfied if there is a random variable $L(v)$ with finite second moment such that for all $x, y \in \mathbb{R}^n$,

$$\|F(v, y) - F(v, x)\| \leq L(v)\|y - x\|,$$

and $\{v^k\}$ is an i.i.d. sample of v. Assumption 3(ii) is satisfied if there exists $C_F > 0$ such that a.s.

$$\sup_{x \in X_0} \mathbb{E}\left[\|F(v,x)\|^2\right] \leq 2C_F^2,$$

and $\{v^k\}$ is an i.i.d. sample of v.

Assumption 4 (Unbiased sampling)**.** *The sequence $\{v^k\}$ has the same distribution as v and a.s. for all $x \in \mathbb{R}^n$ and all $k \in \mathbb{N}$, $\mathbb{E}\Big[F(v^k,x)\Big|\mathcal{F}_k\Big] = T(x)$.*

Assumption 5 (Finite variance)**.** *There exists $\bar{x} \in X$ and $\sigma(\bar{x}) > 0$ such that a.s. for all $k \in \mathbb{N}$, $\mathbb{E}\Big[\|F(v^k,\bar{x}) - T(\bar{x})\|^2\Big|\mathcal{F}_k\Big] \leq \sigma(\bar{x})^2$.*

Observe that since $\{v^k\}$ is a sample drawn from v, $\sigma(\bar{x})^2$ is an upper bound on the variance of $F(v,\bar{x})$. Minkowski's inequality and Assumptions 3(i)-5 imply that for all $x \in X_0$ and $k \in \mathbb{N}$,

$$\mathbb{E}\Big[\|F(v^k,x) - T(x)\|^2\Big|\mathcal{F}_k\Big] \leq \left[2L\|x - \bar{x}\| + \sigma(\bar{x})\right]^2 < \infty, \tag{3.6}$$

while Assumptions 3(ii)-5 imply that for all $x \in X_0$ and $k \in \mathbb{N}$,

$$\mathbb{E}\Big[\|F(v^k,x) - T(x)\|^2\Big|\mathcal{F}_k\Big] \leq 8C_F^2. \tag{3.7}$$

In the following we denote by $\sigma : \mathbb{R}^n \to [0,\infty)$, the function defined by, for every $x \in \mathbb{R}^n$,

$$\sigma(x)^2 := \sup_{k \in \mathbb{N}} \mathbb{E}\Big[\|F(v^k,x) - T(x)\|^2\Big|\mathcal{F}_k\Big]. \tag{3.8}$$

The variance of $F(v,x)$ is bounded above by $\sigma(x)^2$. If $\{v^k\}$ is an i.i.d. sample of v, then $\sigma(x)^2$ is exactly the variance of $F(v,x)$. Observe that $\sigma : \mathbb{R}^n \to [0,\infty)$ is measurable and locally bounded, since (3.6) or (3.7) hold. It should be noted that Assumption 5 is merely a finite variance condition, which is standard in the stochastic setting. Excepting for [73], the much stronger *uniform* condition (1.12) was asked in the previously literature of SA methods for SVI. We do not require (1.12) when the operator is Lipschitz continuous (Assumption 3(i)).

Assumption 6 (weak sharpness)**.** *There exists $\rho > 0$, such that for all $x^* \in X^*$ and all $x \in X$,*

$$\langle T(x^*), x - x^* \rangle \geq \rho\, \mathrm{d}(x, X^*). \tag{3.9}$$

We now state the assumptions concerning the incremental projections.

Assumption 7 (Constraint sampling and regularity)**.** *There exists $c > 0$ such that for all $x \in X_0$ and all $k \geq 1$,*

$$\mathrm{d}(x, X)^2 \leq c\mathbb{E}\left[\left(g^+_{\omega_k}(x)\right)^2 \middle| \mathcal{F}_k\right].$$

As commented in [57], this assumption is quite general. For instance, it is satisfied when the index set $\mathcal{I}$ is arbitrary and X has an interior point under nondemanding properties of $\{\omega_k, v^k\}$ (e.g. $\{\omega_k\}$ is i.i.d. and independent of $\{v^k\}$). The constant $c > 0$ depends on the distribution of $\{\omega_k\}$ and $\{v^k\}$ and of regularity properties of the set X. As an example, if $\mathcal{I}$ is finite, $X_0 := \mathbb{R}^n$, $g_i := \mathrm{d}(\cdot, X_i)$ such that for some $\eta > 0$ and all $x \in \mathbb{R}^n$,

$$\mathrm{d}(x, X)^2 \leq \eta \max_{i \in \mathcal{I}} \mathrm{d}(x, X_i)^2, \tag{3.10}$$

and for some $\delta \in (0, 1]$ and all $i \in \mathcal{I}$ and $k \in \mathbb{N}$,

$$\mathbb{P}\left(\omega_k = i | \mathcal{F}_k\right) \geq \frac{\delta}{|\mathcal{I}|}, \tag{3.11}$$

then Lemma 4 in [73] shows that Assumption 7 holds with $c := \eta|\mathcal{I}|/\delta$, where $|\mathcal{I}|$ is the cardinality of $\mathcal{I}$. Condition (3.11) is satisfied, for instance, when $\{\omega_k\}$ is i.i.d. and uniform over $\mathcal{I}$ and $\{\omega_k, v^k\}$ is independent. Condition (3.10), studied in [4, 23], is satisfied when X is an polyhedron. Assumption 7 is satisfied if $\mathrm{d}(x, X)^2 \leq \eta \max_{i \in \mathcal{I}}(g_i^+(x))^2$ for some $\eta > 0$ and all $x \in X_0$, which holds under nondemanding regularity conditions on the set X, e.g., a Slater condition.

Assumption 8 (Small stepsizes)**.** *For all $k \in \mathbb{N}$, $\alpha_k > 0$, $\beta_k \in (0, 2)$, and*

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

We remark here that the use of small stepsizes is forced by two factors: the use of approximate projections instead of exact ones, and the stochastic approximation. Indeed, even with *exact projections*, the method (3.4)-(3.5) still requires small stepsizes.

3.1.3 Convergence analysis

We first state a lemma which is immediate from the Lipschitz continuity 3(i) and convexity of $t \mapsto t^2$.

Lemma 5. *Suppose that Assumptions 3(i)-5 hold and define the function* $B : \mathbb{R}^n \to [0, \infty)$ *as* $B(x) := \sigma(x) + \|T(x)\|$, *for any* $x \in \mathbb{R}^n$. *Then, almost surely, for all* $x, y \in \mathbb{R}^n$, $k \in \mathbb{N}$,

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

We now prove an iterative relation to be used in the convergence analysis. We mention that (3.12) is sufficient for the convergence analysis and includes the case of unbounded X and T. If the operator is bounded or X_0 is compact, then (3.13) allows an improvement of the convergence rate.

Lemma 6 (Recursive relation)**.** *Suppose that Assumptions 1-8 hold. For all* $x^* \in X^*$, $k \in \mathbb{N}$ *and* $\tau > 1$ *define* $\mathsf{A}_{k,\tau} := \beta_k(2 - \beta_k)(\tau - 1)/(cC_g^2\tau)$, $\mathsf{B}_{k,\tau} := \beta_k(2 - \beta_k)\tau$ *and* $\mathsf{C}(x^*) := \rho + B(x^*)$.

If Assumption 3(i) holds, then for all $x^* \in X^*$, $\tau > 1$ *and* $k \in \mathbb{N}$,

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

$$+ \left[\frac{\mathsf{C}(x^*)^2}{\mathsf{A}_{k,\tau}} + 2\left(1 + \mathsf{B}_{k,\tau}\right) B(x^*)^2\right] \alpha_k^2. \tag{3.12}$$

If Assumption 3(ii) holds, then for all $x^* \in X^*$, $\tau > 1$ *and* $k \in \mathbb{N}$,

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

$$+ \left[\frac{\left(\rho + \sqrt{2C_F}\right)^2}{\mathsf{A}_{k,\tau}} + 2\left(1 + \mathsf{B}_{k,\tau}\right) C_F^2\right] \alpha_k^2. \tag{3.13}$$

Proof. Take $x^* \in X^*$, $\tau > 1$ and $k \in \mathbb{N}$. We claim that

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - 2\alpha_k\langle x^k - x^*, F(v^k, x^k)\rangle +$$

$$\left[1 + \tau\beta_k(2 - \beta_k)\right] \alpha_k^2\|F(v^k, x^k)\|^2 - \frac{\beta_k(2 - \beta_k)}{C_g^2}\left(1 - \frac{1}{\tau}\right)\left(g_{\omega_k}^+(x^k)\right)^2. \tag{3.14}$$

Indeed, by the definition of the method (3.4)-(3.5), we can invoke Lemma 2 with $g := g_{\omega_k}$, $x_1 := x^k$, $x_2 := x^{k+1}$, $y := y^k$, $x_0 := x^*$, $\alpha := \alpha_k$, $u := F(v^k, x^k)$, $\beta := \beta_k$ and $d := d^k$, obtaining (3.14).

We now take the conditional expectation with respect to $\mathcal{F}_k$ in (3.14) obtaining,

$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2 \middle| \mathcal{F}_k\right] \leq \|x^k - x^*\|^2 - 2\alpha_k \langle x^k - x^*, T(x^k)\rangle +$$

$$\left[1 + \tau\beta_k(2-\beta_k)\right]\alpha_k^2 \mathbb{E}\left[\|F(v^k, x^k)\|^2 \middle| \mathcal{F}_k\right] -$$

$$\frac{\beta_k(2-\beta_k)}{C_g^2}\left(1 - \frac{1}{\tau}\right)\mathbb{E}\left[\left(g_{\omega_k}^+(x^k)\right)^2 \middle| \mathcal{F}_k\right] \leq \|x^k - x^*\|^2 + 2\alpha_k\langle x^* - x^k, T(x^k)\rangle +$$

$$\left[1 + \tau\beta_k(2-\beta_k)\right]\alpha_k^2 \mathbb{E}\left[\|F(v^k, x^k)\|^2 \middle| \mathcal{F}_k\right] - \frac{\beta_k(2-\beta_k)}{c \cdot C_g^2}\left(1 - \frac{1}{\tau}\right)\mathrm{d}(x^k, X)^2, \tag{3.15}$$

using $x^k \in \mathcal{F}_k$ and Assumption 4 in the first inequality, and Assumption 7 in the second inequality.

Next, we will bound the second term in the right hand side of (3.15). We write

$$\langle T(x^k), x^* - x^k\rangle = \langle T(x^k) - T(x^*), x^* - x^k\rangle +$$

$$\langle T(x^*), x^* - \Pi_X(x^k)\rangle + \langle T(x^*), \Pi_X(x^k) - x^k\rangle. \tag{3.16}$$

By monotonicity of T (Assumption 2), the first term in the right hand side of (3.16) satisfies

$$\langle T(x^k) - T(x^*), x^* - x^k\rangle \leq 0. \tag{3.17}$$

Regarding the second term in the right hand side of (3.16), the weak sharpness property (Assumption 6) and the fact that $x \in X^*$ imply

$$\langle T(x^*), x^* - \Pi_X(x^k)\rangle \leq -\rho\,\mathrm{d}\left(\Pi_X(x^k), X^*\right). \tag{3.18}$$

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

$$\mathrm{d}\left(\Pi_X(x^k), X^*\right) \geq \mathrm{d}(x^k, X^*) - \mathrm{d}(x^k, X). \tag{3.19}$$

From (3.18)-(3.19), we get

$$\langle T(x^*), x^* - \Pi_X(x^k)\rangle \leq -\rho\,\mathrm{d}(x^k, X^*) + \rho\,\mathrm{d}(x^k, X). \tag{3.20}$$

Concerning the third term in the right hand side of (3.16), we have

$$\langle T(x^*), \Pi_X(x^k) - x^k\rangle \leq \|T(x^*)\|\|\Pi_X(x^k) - x^k\| \leq B(x^*)\,\mathrm{d}(x^k, X), \tag{3.21}$$

using the Cauchy-Schwarz inequality in the first inequality, and the definition of $B(x^*)$ in Lemma 5 in the second inequality. Combining (3.17), (3.20) and (3.21) with (3.16), we finally get

$$\langle T(x^k), x^* - x^k\rangle \leq -\rho\,\mathrm{d}(x^k, X^*) + (\rho + B(x^*))\,\mathrm{d}(x^k, X). \tag{3.22}$$

We use (3.22) in (3.15) and get

$$\begin{aligned}\mathbb{E}\Big[\|x^{k+1} - x^*\|^2\Big|\mathcal{F}_k\Big] \leq \|x^k - x^*\|^2 - 2\rho\alpha_k\,\mathrm{d}(x^k, X^*)\\ + [1 + \tau\beta_k(2-\beta_k)]\,\alpha_k^2\mathbb{E}\left[\|F(v^k, x^k)\|^2\Big|\mathcal{F}_k\right]\\ - \frac{\beta_k(2-\beta_k)}{c\cdot C_g^2}\left(1 - \frac{1}{\tau}\right)\mathrm{d}(x^k, X)^2 + 2(\rho + B(x^*))\alpha_k\,\mathrm{d}(x^k, X).\end{aligned} \tag{3.23}$$

Now we rearrange the last two terms in the right hand side of (3.23), using the fact that $2ab \leq \lambda^2 a^2 + \frac{b^2}{\lambda^2}$ for any $\lambda > 0$. With $a := \mathrm{d}(x^k, X)$, $b := \mathsf{C}(x^*)\alpha_k$ and $\lambda := \mathsf{A}_{k,\tau}$ we get

$$-\mathsf{A}_{k,\tau}\,\mathrm{d}(x^k, X)^2 + 2\mathsf{C}(x^*)\alpha_k\,\mathrm{d}(x^k, X) \leq \frac{\mathsf{C}(x^*)^2\alpha_k^2}{\mathsf{A}_{k,\tau}}. \tag{3.24}$$

From Lemma 5 and $x^k \in \mathcal{F}_k$, we obtain

$$\mathbb{E}\Big[\|F(v^k, x^k)\|^2\Big|\mathcal{F}_k\Big] \leq 2L^2\|x^k - x^*\|^2 + 2B(x^*)^2. \tag{3.25}$$

Putting together relations (3.23)-(3.25) and rearranging terms, we finally get (3.12), as requested.

Suppose now that Assumption 3(ii) holds. In this case, the inequalities in (3.21) can be replaced by

$$\langle T(x^*), \Pi_X(x^k) - x^k\rangle \leq \|T(x^*)\|\|\Pi_X(x^k) - x^k\| \leq \sqrt{2C_F}\,\mathrm{d}(x^k, X), \tag{3.26}$$

using, in the last inequality, Assumption 3(ii) and the fact that $\|T(x^*)\|^2 \leq \mathbb{E}\Big[\|F(v^k, x^*)\|^2\Big|\mathcal{F}_k\Big] \leq 2C_F^2$, which follows from Jensen's inequality. Hence, combining (3.17), (3.20) and (3.26) we get, instead of (3.22),

$$\langle T(x^k), x^* - x^k\rangle \leq -\rho\,\mathrm{d}(x^k, X^*) + \left(\rho + \sqrt{2C_F}\right)\mathrm{d}(x^k, X). \tag{3.27}$$

Using Assumption 3(ii) and (3.27) in (3.15) we get

$$\mathbb{E}\Big[\|x^{k+1}-x^*\|^2\Big|\mathcal{F}_k\Big] \leq \|x^k-x^*\|^2 - 2\rho\alpha_k\,\mathrm{d}(x^k,X^*) + 2C_F^2\,[1+\mathsf{B}_{k,\tau}]\,\alpha_k^2$$

$$- \mathsf{A}_{k,\tau}\,\mathrm{d}(x^k,X)^2 + 2\left(\rho+\sqrt{2C_F}\right)\alpha_k\,\mathrm{d}(x^k,X). \tag{3.28}$$

In view of Assumption 1, we define $\bar{x}^k := \Pi_{X^*}(x^k)$. Note that $\bar{x}^k \in \mathcal{F}_k$ because Π_{X^*} is continuous and $x^k \in \mathcal{F}_k$. From (3.28) we get

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

$$+ 2C_F^2\,[1+\mathsf{B}_{k,\tau}]\,\alpha_k^2 - \mathsf{A}_{k,\tau}\,\mathrm{d}(x^k,X)^2 + 2\left(\rho+\sqrt{2C_F}\right)\alpha_k\,\mathrm{d}(x^k,X), \tag{3.29}$$

using the fact that $x^k, \bar{x}^k \in \mathcal{F}_k$, $\|x^k-\bar{x}^k\| = \mathrm{d}(x^k,X^*)$ and (3.28) in the second inequality. We rearrange now the last two terms in the right hand side of (3.29) (as we did in (3.24)), and obtain (3.13).

□

Theorem 5 (Asymptotic convergence)**.** *Under Assumptions 1-8, method* (3.4)-(3.5) *generates a sequence* $\{x^k\}$ *which a.s. is bounded and* $\lim_{k\to\infty}\mathrm{d}(x^k,X^*)=0$. *In particular, a.s. all cluster points of* $\{x^k\}$ *belong to* X^*.

Proof. We suppose first that Assumption 3(i) holds. Choose some $x^* \in X^*$ (Assumption 1) and $\tau > 1$. By Assumption 8 and the definitions given in Lemma 6, we have that $\sum_k \alpha_k^2 < \infty$, $\sum_k \alpha_k^2 \mathsf{A}_{k,\tau}^{-1} < \infty$ and $0 < \mathsf{B}_{k,\tau} \leq \tau$, since $\beta_k(2-\beta_k) \in (0,1]$, for $\beta_k \in (0,2)$ for all k. Hence, we can invoke (3.12) in Theorem 1 in order to conclude that, a.s., $\{\|x^k-x^*\|\}$ converges and, in particular, $\{x^k\}$ is bounded.

In view of Assumption 1, we can define $\bar{x}^k := \Pi_{X^*}(x^k)$. We have $\bar{x}^k \in \mathcal{F}_k$ because $x^k \in \mathcal{F}_k$ and Π_{X^*} is continuous. Since (3.12) in Lemma 6 holds for any $x^* \in X^*$ and $\mathrm{d}(x^k,X^*) = \|x^k-\bar{x}^k\|$, we conclude that for all $k \in \mathbb{N}$,

$$\mathbb{E}\Big[\,\mathrm{d}(x^{k+1},X^*)^2\Big|\mathcal{F}_k\Big] \leq \mathbb{E}\Big[\|x^{k+1}-\bar{x}^k\|^2\Big|\mathcal{F}_k\Big] \leq \Big[1+2\,(1+\mathsf{B}_{k,\tau})\,L^2\alpha_k^2\Big]\,\|x^k-\bar{x}^k\|^2 -$$

$$2\rho\alpha_k\,\mathrm{d}(x^k,X^*) + \left[\frac{\mathsf{C}(\bar{x}^k)}{\mathsf{A}_{k,\tau}} + 2\,(1+\mathsf{B}_{k,\tau})\,B(\bar{x}^k)^2\right]\alpha_k^2$$

$$= \Big[1+2\,(1+\mathsf{B}_{k,\tau})\,L^2\alpha_k^2\Big]\,\mathrm{d}(x^k,X^*)^2 -$$

$$2\rho\alpha_k \, \mathrm{d}(x^k, X^*) + \left[\frac{\mathsf{C}(\bar{x}^k)}{\mathsf{A}_{k,\tau}} + 2\,(1 + \mathsf{B}_{k,\tau})\, B(\bar{x}^k)^2 \right] \alpha_k^2, \tag{3.30}$$

using relation (3.12) and the fact that $\bar{x}^k \in \mathcal{F}_k$ in the second inequality.

We observe that the function $B : X^* \to \mathbb{R}_+$ defined in Lemma 5 is locally bounded, because σ is locally bounded and T is continuous. Using this fact, the continuity of Π_{X^*}, the a.s.-boundedness of $\{x^k\}$ and the fact that $\bar{x}^k = \Pi_{X^*}(x^k)$, we conclude that $\{B(\bar{x}^k)\}$ and $\{\mathsf{C}(\bar{x}^k)\}$ are a.s.-bounded. From the a.s.-boundedness of $\{B(\bar{x}^k)\}$ and $\{\mathsf{C}(\bar{x}^k)\}$ and the conditions $\sum_k \alpha_k^2 < \infty$, $\sum_k \alpha_k^2 \mathsf{A}_{k,\tau}^{-1} < \infty$ and $0 < \mathsf{B}_{k,\tau} \leq \tau$ for all k, which hold by Assumption 8, we conclude from Theorem 1 and (3.30) that a.s. $\{\mathrm{d}^2(x^k, X^*)\}$ converges, and that

$$\sum_{k=0}^{\infty} 2\rho\alpha_k \, \mathrm{d}(x^k, X^*) < \infty. \tag{3.31}$$

By Assumption 8, we also have $\sum_k \alpha_k = \infty$, so that (3.31) implies that, almost surely, $\liminf_{k\to\infty} \mathrm{d}(x^k, X^*) = 0$. In particular, the sequence $\{\mathrm{d}(x^k, X^*)\}$ has a subsequence that converges to zero almost surely. Since $\{\mathrm{d}(x^k, X^*)\}$ a.s. converges, we conclude that the whole sequence a.s. converges to 0. The proof under Assumption 3(ii) is similar, using (3.13). □

3.1.4 Convergence rate analysis

Next, we present convergence rate results in terms of $\mathrm{d}(x^k, X^*)$ for the method (3.4)-(3.5) under the weak sharpness property (3.9). We apply these results for obtaining an estimate of the number of iterations required so that any solution of an auxiliary stochastic optimization problem with linear objective is a solution of the variational inequality (see Corollary 4).

In order to give convergence rates for the case of an unbounded feasible set X or unbounded constraint components $\{X_0\} \cup \{X_i : i \in \mathcal{I}\}$, we shall need the following proposition, which ensures that the sequence is bounded in L^2. A typical situation is the case in which X is a polyhedron, i.e. $X_0 = \mathbb{R}^n$ and the selected constraints $\{X_i\}_{i\in\mathcal{I}}$ are halfspaces, which have easily computable projections but are unbounded sets. If the uniform bound of Assumption 3(ii) holds, then sharper bounds are given in (3.34).

Proposition 2 (Boundedness in L^2)**.** *Suppose that Assumptions 1-8 hold. Under Assumption 3(i), choose $\tau > 1$, $k_0 \in \mathbb{N}$ and $0 < \gamma < \frac{1}{2(1+\tau)L^2}$ such that*

$$\sum_{k \geq k_0} \frac{\alpha_k^2}{\beta_k(2-\beta_k)} < \gamma. \tag{3.32}$$

Define $\mathsf{G}_\tau := cC_g^2\tau(\tau-1)^{-1}$ and $\mathsf{H}_\tau := 2\,(1+\tau)$. Then for all $x^* \in X^*$,

$$\sup_{k \geq k_0} \mathbb{E}\left[\|x^k - x^*\|^2\right] \leq \frac{\mathbb{E}\left[\|x^{k_0} - x^*\|^2\right] + \left[\mathsf{G}_\tau \mathsf{C}(x^*)^2 + \mathsf{H}_\tau B(x^*)^2\right]\gamma}{1 - \mathsf{H}_\tau L^2 \gamma}. \tag{3.33}$$

Define, for $\ell \leq k$, $\mathsf{a}_\ell^k := \sum_{i=\ell}^k \alpha_i^2$, $\mathsf{b}_\ell^k := \sum_{i=\ell}^k \frac{\alpha_i^2}{\beta_i(2-\beta_i)}$. If Assumption 3(ii) holds, then for all $k \in \mathbb{N}$,

$$\sup_{0 \leq i \leq k} \mathbb{E}\left[\mathrm{d}(x^k, X^*)^2\right] \leq \mathrm{d}(x^0, X^*)^2 + \mathsf{G}_\tau \left(\rho + \sqrt{2C_F}\right)^2 \cdot \mathsf{b}_0^{k-1} + \mathsf{H}_\tau C_F^2 \cdot \mathsf{a}_0^{k-1}. \tag{3.34}$$

Proof. We first prove (3.33) under Assumption 3(i). Recall definitions of $\mathsf{A}_{k,\tau}$ and $\mathsf{B}_{k,\tau}$ in Lemma 6. By Assumption 8, we can choose $k_0 \in \mathbb{N}$ and $\gamma > 0$ such that (3.32) holds. Observe that $\beta_k(2-\beta_k) \in (0,1]$, because $\beta_k \in (0,2)$, so that

$$\sum_{k \geq k_0} \alpha_k^2 \leq \sum_{k \geq k_0} \frac{\alpha_k^2}{\beta_k(2-\beta_k)} < \gamma.$$

Fix $x \in X^*$ and $\tau > 1$. Define

$$z_k := \mathbb{E}[\|x^k - x^*\|^2], \quad D_k^2 := \frac{\mathsf{C}(x^*)^2}{\mathsf{A}_{k,\tau}} + 2\,(1 + \mathsf{B}_{k,\tau})\,B(x^*)^2,$$

$$D^2 := \frac{\mathsf{C}(x^*)^2 cC_g^2\tau}{\tau - 1} + 2\,(1+\tau)\,B(x^*)^2.$$

For any $k > k_0$, we take the total expectation and sum (3.12) from k_0 to $k-1$, obtaining

$$z_k \leq z_{k_0} + \sum_{i=k_0}^{k-1} \left[2\,(1 + \mathsf{B}_{i,\tau})\,L^2\alpha_i^2 z_i + D_i^2\alpha_i^2\right]. \tag{3.35}$$

Given an arbitrary $a > z_{k_0}^{1/2}$, define

$$\Gamma_a := \inf\{k \geq k_0 : z_k > a^2\}. \tag{3.36}$$

Suppose first that $\Gamma_a < \infty$ for all $a > z_{k_0}^{1/2}$. Then by (3.32), (3.35) and (3.36) we get

$$a^2 < z_{\Gamma_a} \leq z_{k_0} + \sum_{i=k_0}^{\Gamma_a - 1} \left[2\left(1 + \mathsf{B}_{i,\tau}\right) L^2 \alpha_i^2 a^2 + D_i^2 \alpha_i^2 \right] \leq z_{k_0} + 2(1+\tau) L^2 \gamma a^2 + D^2 \gamma,$$

using $\beta_i(2-\beta_i) \in (0,1]$ in the definition of $\mathsf{B}_{i,\tau}$, and the definitions of $\mathsf{A}_{i,\tau}$, D_i^2 and D^2. Hence

$$a^2 \leq \frac{z_{k_0} + D^2\gamma}{1 - 2(1+\tau)L^2\gamma},$$

using $0 < \gamma < [2(1+\tau)L^2]^{-1}$. Since $a > z_{k_0}^{1/2}$ is arbitrary, it follows that

$$\sup_{k \geq k_0} z_k \leq \frac{z_{k_0} + D^2\gamma}{1 - 2(1+\tau)L^2\gamma}, \tag{3.37}$$

using again that $0 < \gamma < [2(1+\tau)L^2]^{-1}$. In view of (3.36)-(3.37), we have a contradiction with the assumption that $\Gamma_a < \infty$ for any $a > z_{k_0}^{1/2}$. Hence, there exists some $\bar{a} > z_{k_0}^{1/2}$ such that $\Gamma_{\bar{a}} = \infty$, so that the set in the right hand side of (3.36) is empty. In this case we have $\sup_{k \geq k_0} z_k \leq \bar{a}^2 < \infty$. If $\sup_{k \geq k_0} z_k = z_{k_0}$, then (3.33) holds trivially, since $1 - \mathsf{H}_\tau L^2 \gamma \in (0,1)$. Otherwise, $\hat{a} := (\sup_{k \geq k_0} z_k)^{1/2} > z_{k_0}^{1/2}$. From (3.32), (3.35), $\beta_i \in (0,2)$ and the definitions of $\mathsf{A}_{i,\tau}$, $\mathsf{B}_{i,\tau}$, D_i^2 and D, we have for all $k \geq k_0$,

$$z_k \leq z_{k_0} + \sum_{i=k_0}^{k-1} \left[2\left(1 + \mathsf{B}_{i,\tau}\right) L^2 \alpha_i^2 \hat{a}^2 + D_i^2 \alpha_i^2 \right] \leq z_{k_0} + 2(1+\tau) L^2 \gamma \hat{a}^2 + D^2 \gamma,$$

implying that $\hat{a}^2 = \sup_{k \geq k_0} z_k \leq z_{k_0} + 2(1+\tau)L^2\gamma\hat{a}^2 + D^2\gamma$, so that

$$\sup_{k \geq k_0} z_k = \hat{a}^2 \leq \frac{z_{k_0} + D^2\gamma}{1 - 2(1+\tau)L^2\gamma}, \tag{3.38}$$

using again $0 < \gamma < [2(1+\tau)L^2]^{-1}$. From (3.38) and the definitions of G_τ, H_τ and D, we conclude that (3.33) holds.

We now prove (3.34) under Assumption 3(ii). As before, we define

$$\widehat{D}_k^2 := \frac{\left(\rho + \sqrt{2C_F}\right)^2}{\mathsf{A}_{k,\tau}} + 2\left(1 + \mathsf{B}_{k,\tau}\right) C_F^2.$$

Taking total expectation in (3.13) and summing from 0 to $k-1$, we get

$$\mathbb{E}\left[\mathrm{d}(x^k, X^*)^2 \right] \leq \mathrm{d}(x^0, X^*)^2 + \sum_{i=0}^{k-1} \widehat{D}_i^2 \alpha_i^2$$

$$\leq \mathrm{d}(x^0, X^*)^2 + \left(\rho + \sqrt{2C_F}\right)^2 \frac{cC_g^2\tau}{\tau - 1}\mathsf{b}_0^{k-1} + 2(1+\tau)C_F^2\mathsf{a}_0^{k-1}, \tag{3.39}$$

for all $k \geq 0$, using the fact that $\beta_i \in (0,2)$ and the definitions of $\mathsf{A}_{i,\tau}$, $\mathsf{B}_{i,\tau}$, $\widehat{D}_i^2$, a_0^{k-1} and b_0^{k-1}. We conclude from (3.39) and the definitions of G_τ and H_τ that (3.34) holds. □

We give now convergence rate results. We define, for $\ell \leq k$,

$$\mathsf{S}_\ell^k := \sum_{i=\ell}^{k} \alpha_i, \quad \widehat{x}^k := \frac{\sum_{i=0}^k \alpha_i x^i}{\mathsf{S}_0^k}, \quad \widehat{x}_\ell^k := \frac{\sum_{i=\ell}^k \alpha_i x^i}{\mathsf{S}_\ell^k},$$

where $\widehat{x}^k$ is the ergodic average of the iterates and $\widehat{x}_\ell^k$ is the window-based ergodic average of the iterates. Next we will give convergence rate results for the original sequence $\{x^k\}$ and for the ergodic average sequences. We consider separately the cases of unbounded operators (Assumption 3(i)) and the case of bounded ones (Assumption 3(ii)), because in the later case sharper rates are possible.

Theorem 6 (Rate of convergence: unbounded case)**.** *Suppose that Assumptions 1-8 and Assumption 3(i) hold. Recall definitions of Proposition 2 and $\{\mathsf{S}_\ell^k\}$. Choose $\tau > 1$, $k_0 \in \mathbb{N}$ and $\phi \in (0,1)$ such that*

$$\sum_{k \geq k_0} \frac{\alpha_k^2}{\beta_k(2-\beta_k)} \leq \frac{\phi}{2(1+\tau)L^2}. \tag{3.40}$$

Define for $x^* \in X^*$,

$$\mathsf{E}_k(x^*) := (2\rho)^{-1} \cdot \left\{ \|x^0 - x^*\|^2 + \left[\mathsf{I}(x^*)L^2 + B(x^*)^2\right] \mathsf{H}_\tau \mathsf{a}_0^k + \mathsf{G}_\tau \mathsf{C}(x^*)^2 \mathsf{b}_0^k \right\}, \tag{3.41}$$

$$\mathsf{I}(x^*) := \frac{\max_{0 \leq k \leq k_0} \mathbb{E}\left[\|x^k - x^*\|^2\right] + \left[\mathsf{G}_\tau \mathsf{H}_\tau^{-1} \mathsf{C}(x^*)^2 L^{-2} + B(x^*)^2 L^{-2}\right] \phi}{1 - \phi}. \tag{3.42}$$

Then

a) For any $\epsilon > 0$, there exists $M := M_\epsilon \in \mathbb{N}$, such that $\mathbb{E}\left[\mathrm{d}(x^M, X^*)\right] < \epsilon$ and for all $x^* \in X^*$, $\mathsf{S}_0^{M-1} \leq \mathsf{E}_\infty(x^*)/\epsilon$.

b) For all $k \in \mathbb{N}$ and all $x^* \in X^*$, $\mathbb{E}\left[\mathrm{d}(\widehat{x}^k, X^*)\right] \leq \mathsf{E}_k(x^*)/\mathsf{S}_0^k$.

Proof. Fix $\tau > 1$, $k_0 \in \mathbb{N}$ and $\phi \in (0,1)$ as in (3.40). This is possible since $\sum_{i\geq k} \alpha_i^2 \beta_i^{-1}(2-\beta_i)^{-1}$ converge to 0 as $k \to \infty$ by Assumption 8. We now invoke Lemma 6. We take the total expectation in (3.12) and sum from ℓ to k, obtaining, for every $x^* \in X^*$,

$$
\begin{aligned}
&2\rho \sum_{i=\ell}^{k} \alpha_i \mathbb{E}\left[\mathrm{d}(x^i, X^*)\right] \leq \\
&\leq \mathbb{E}\left[\|x^\ell - x^*\|^2\right] + \sum_{i=\ell}^{k} 2\left(1+\mathsf{B}_{i,\tau}\right) L^2 \alpha_i^2 \mathbb{E}\left[\|x^i - x^*\|^2\right] \\
&\quad + \sum_{i=\ell}^{k} \left[\frac{\mathsf{C}(x^*)^2}{\mathsf{A}_{i,\tau}} + 2\left(1+\mathsf{B}_{i,\tau}\right) B(x^*)^2\right] \alpha_i^2 \\
&\leq \mathbb{E}\left[\|x^\ell - x^*\|^2\right] + \left(\sup_{\ell \leq i \leq k} \mathbb{E}\left[\|x^i - x^*\|^2\right]\right) \sum_{i=\ell}^{k} 2\left(1+\mathsf{B}_{i,\tau}\right) L^2 \alpha_i^2 \\
&\quad + \sum_{i=\ell}^{k} \left[\frac{\mathsf{C}(x^*)^2}{\mathsf{A}_{i,\tau}} + 2\left(1+\mathsf{B}_{i,\tau}\right) B(x^*)^2\right] \alpha_i^2,
\end{aligned}
$$

(3.43)

$$
\leq \mathbb{E}\left[\|x^\ell - x^*\|^2\right] + \left(\sup_{i\geq 0} \mathbb{E}\left[\|x^i - x^*\|^2\right]\right) \mathsf{H}_\tau L^2 \mathsf{a}_\ell^k + \mathsf{G}_\tau \mathsf{C}(x^*)^2 \mathsf{b}_\ell^k + \mathsf{H}_\tau B(x^*)^2 \mathsf{a}_\ell^k,
$$

using $\beta_i(2-\beta_i) \in (0,1]$ and the definitions of $\mathsf{A}_{i,\tau}$, $\mathsf{B}_{i,\tau}$, G_τ, H_τ, a_ℓ^k and b_ℓ^k in the last inequality.

We now invoke Proposition 2. Setting $\gamma := \frac{\phi}{2(1+\tau)L^2}$, (3.32) can be rewritten as (3.40). From (3.33) and $1 - \mathsf{H}_\tau L^2 \in (0,1)$, we get, for all $x^* \in X^*$,

(3.44)

$$
\sup_{i\geq 0} \mathbb{E}\left[\|x^i - x^*\|^2\right] \leq \frac{\max_{0\leq i\leq k_0} \mathbb{E}\left[\|x^i - x^*\|^2\right] + \left[\mathsf{G}_\tau \mathsf{C}(x^*)^2 + \mathsf{H}_\tau B(x^*)^2\right]\gamma}{1 - \mathsf{H}_\tau L^2 \gamma} = \mathsf{I}(x^*),
$$

using the definitions of $\mathsf{H}_\tau = 2(1+\tau)$, γ and $\mathsf{I}(x^*)$.

We prove now item (a). For every $\epsilon > 0$, define

$$
M = M_\epsilon := \inf\left\{k \in \mathbb{N} : \mathbb{E}\left[\mathrm{d}(x^k, X^*)\right] < \epsilon\right\}. \tag{3.45}
$$

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

$$
2\rho\epsilon \sum_{i=0}^{k} \alpha_i \leq 2\rho \sum_{i=0}^{k} \alpha_i \mathbb{E}\left[\mathrm{d}(x^i, X^*)\right]. \tag{3.46}
$$

We claim that M is finite. Indeed, if $M = \infty$, then (3.43), (3.44) and (3.46) hold for $\ell := 0$ and all $k \in \mathbb{N}$. Hence, letting $k \to \infty$ and using that $\mathsf{a}_0^\infty < \infty$ and

$\mathsf{b}_0^\infty < \infty$, which hold by Assumption 8, we obtain $\sum_k \alpha_k < \infty$, which contradicts Assumption 8. Hence, the set in the right hand side of (3.45) is nonempty, which implies $\mathbb{E}[\mathrm{d}(x^M, X^*)] < \epsilon$. Setting $\ell := 0$ and $k := M-1$ in (3.43), (3.44) and (3.46), we get, for all $x^* \in X^*$,

$$\sum_{i=0}^{M-1} \alpha_i \leq \frac{\mathsf{E}_{M-1}(x^*)}{\epsilon} \leq \frac{\mathsf{E}_\infty(x^*)}{\epsilon},$$

using the definition of $\mathsf{E}_k(x^*)$. We thus obtain item (a).

We now prove item (b). In view of the convexity of the function $x \longmapsto \mathrm{d}(x, X^*)$, and the linearity and monotonicity of the expected value, we have

$$\mathbb{E}\left[\mathrm{d}(\widehat{x}_\ell^k, X^*)\right] = \mathbb{E}\left[\mathrm{d}\left(\frac{\sum_{i=\ell}^k \alpha_i x^i}{\sum_{i=\ell}^k \alpha_i}, X^*\right)\right] \leq \frac{\sum_{i=\ell}^k \alpha_i \mathbb{E}\left[\mathrm{d}(x^i, X^*)\right]}{\sum_{i=\ell}^k \alpha_i}. \tag{3.47}$$

Set $\ell := 0$, divide (3.43) by $2\rho \sum_{i=0}^k \alpha_i = 2\rho \mathsf{S}_0^k$ and use (3.47), the definition of $\mathsf{E}_0^k(x^*)$ together with (3.44) in order to bound $\sup_{i\geq 0} \mathbb{E}[\|x^i - x^*\|^2]$, and obtain item (b) as a consequence. □

Corollary 1 (Rate of convergence with robust stepsizes: unbounded case)**.** *Assume that the hypotheses of Theorem 6 hold. Given $\theta > 0$ and $\lambda > 0$, define $\{\alpha_k\}$ as: $\alpha_0 = \alpha_1 = \theta$ and for $k \geq 2$,*

$$\alpha_k := \frac{\theta}{\sqrt{k\,(\ln k)^{1+\lambda}}}, \tag{3.48}$$

and choose $\beta_k \equiv \beta \in (0,2)$, $\tau > 1$ and $\phi \in (0,1)$. Take $k_0 \geq 2$ as the minimum natural number such that

$$k_0 \geq \exp\left[\left(\frac{2(1+\tau)L^2\theta^2}{\lambda\beta(2-\beta)\phi}\right)^{1/\lambda}\right] + 1. \tag{3.49}$$

Define for $x^* \in X^*$, $\mathsf{J}(x^*) := [\mathsf{I}(x^*)L^2 + B(x^*)^2]\,\mathsf{H}_\tau + \mathsf{G}_\tau \mathsf{C}(x^*)^2 \beta^{-1}(2-\beta)^{-1}$.

Then $\mathrm{d}(x^k, X^*)$ a.s.-converges to 0 and the following statements hold:

a) For every $\epsilon > 0$, there exists $M = M_\epsilon \geq 2$ such that $\mathbb{E}\left[\mathrm{d}(x^M, X^*)\right] < \epsilon$ with

$$\epsilon \leq \frac{\max\{\theta, \theta^{-1}\}}{2\rho} \cdot \frac{[\ln(M-1)]^{\frac{1+\lambda}{2}}}{\sqrt{M-1}} \cdot$$

$$\inf_{x^* \in X^*} \left\{ \|x^0 - x^*\|^2 + \mathsf{J}(x^*)\left[2 + \frac{1}{2(\ln 2)^{1+\lambda}} + \frac{1}{\lambda(\ln 2)^\lambda}\right]\right\},$$

b) For all $k \geq 2$,

$$\mathbb{E}\left[\mathrm{d}(\widehat{x}^k, X^*)\right] \leq \frac{\max\{\theta, \theta^{-1}\}}{2\rho} \cdot \frac{(\ln k)^{\frac{1+\lambda}{2}}}{\sqrt{k}} \cdot$$

$$\inf_{x^* \in X^*} \left\{ \|x^0 - x^*\|^2 + \mathsf{J}(x^*) \left[2 + \frac{1}{2(\ln 2)^{1+\lambda}} + \frac{1}{\lambda(\ln 2)^{\lambda}} \right] \right\}.$$

Proof. Let $k \geq 2$. We first estimate the sum of the stepsize sequence. For any $0 \leq \ell \leq k$ we have

$$\mathsf{S}_\ell^k = \sum_{i=\ell}^{k} \alpha_i \geq \frac{\theta(k-\ell+1)}{\sqrt{k(\ln k)^{1+\lambda}}}, \tag{3.50}$$

using the fact that the minimum stepsize between ℓ and $k \geq 2$ is $\theta k^{-\frac{1}{2}} (\ln k)^{\frac{1+\lambda}{2}}$. The sum of the squares of the stepsizes sequence can be estimated as

$$\begin{aligned} \mathsf{a}_0^k \leq \mathsf{a}_0^\infty &= \sum_{i=0}^{\infty} \alpha_i^2 = 2\theta^2 + \frac{\theta^2}{2(\ln 2)^{1+\lambda}} + \sum_{i=3}^{\infty} \frac{\theta^2}{i(\ln i)^{1+\lambda}} \\ &\leq 2\theta^2 + \frac{\theta^2}{2(\ln 2)^{1+\lambda}} + \theta^2 \int_2^\infty t^{-1} (\ln t)^{-(1+\lambda)} \, \mathrm{d}\, t = \theta^2 \left[2 + \frac{1}{2(\ln 2)^{1+\lambda}} + \frac{1}{\lambda(\ln 2)^{\lambda}} \right]. \end{aligned} \tag{3.51}$$

We assume without loss on generality that we have $M \geq 2$ in (3.45). Item (a) follows from (3.50) with $k := M - 1$ and $\ell := 0$, (3.51), Theorem 6(a) and the definitions of $\mathsf{J}(x^*)$, $\mathsf{E}_\infty(x^*)$, $\mathsf{a}_0^\infty = \beta(2-\beta)\mathsf{b}_0^\infty$.

Similarly, item (b) follows from (3.50)-(3.51) with $\ell := 0$, Theorem 6(b), the definitions of $\mathsf{J}(x^*)$ and $\mathsf{E}_k(x^*)$ and the facts that $\mathsf{b}_0^k \leq \mathsf{b}_0^\infty$ and $\mathsf{a}_0^\infty = \beta(2-\beta)\mathsf{b}_0^\infty$.

Finally, we estimate k_0 in (3.42). Since

$$\sum_{k \geq k_0} \alpha_k^2 < \theta^2 \int_{k_0 - 1}^\infty t^{-1} (\ln t)^{-(1+\lambda)} \, \mathrm{d}\, t = \frac{\theta^2}{\lambda \left[\ln(k_0 - 1)\right]^{\lambda}},$$

we conclude from (3.40) that it is enough to choose the minimum $k_0 \geq 2$ such that

$$\frac{\theta^2}{\lambda \left[\ln(k_0 - 1)\right]^{\lambda}} \leq \frac{\beta(2-\beta)\phi}{2(1+\tau)L^2},$$

that is to say, the minimum $k_0 \geq 2$ such that (3.49) holds. □

Remark 2. As an immediate consequence of Corollary 1, we get that $\{\mathrm{d}(\widehat{x}^k, X^*)\}$ converges to 0 in L^1 (besides the a.s. convergence). This property may fail with $\{x^k\}$ instead of $\{\widehat{x}^k\}$.

Theorem 7 (Rate of convergence: bounded case)**.** *Suppose that Assumptions 1-8 and Assumption 3(ii) hold. Recall definitions of Proposition 2 and* $\{\mathsf{S}_\ell^k\}$*. Choose* $\tau > 1$*. Define for* $\ell \leq k$ *in* $\mathbb{N}_0 \cup \{\infty\}$*,* $R > 0$*,*

$$\mathsf{E}_\ell^k[R] := (2\rho)^{-1} \left\{ R^2 + \mathsf{G}_\tau \Big(\rho + \sqrt{2C_F} \Big)^2 \mathsf{b}_\ell^k + \mathsf{H}_\tau C_F^2 \mathsf{a}_\ell^k \right\}.$$

Then, $\mathrm{d}(x^k, X^*)$ *a.s.-converges to zero and*

a) for any $\epsilon > 0$*, there exists* $M := M_\epsilon \in \mathbb{N}$*, such that* $\mathbb{E}\left[\mathrm{d}(x^M, X^*)\right] < \epsilon$ *and* $\mathsf{S}_0^{M-1} \leq \mathsf{E}_0^\infty[\mathrm{d}(x^0, X^*)]/\epsilon$*,*

b) for all $k \in \mathbb{N}$*,* $\mathbb{E}\left[\mathrm{d}(\widehat{x}^k, X^*)\right] \leq \mathsf{E}_0^k[\mathrm{d}(x^0, X^*)]/\mathsf{S}_0^k$*,*

c) If X_0 *is compact, then for all* $\ell, k \in \mathbb{N}$ *with* $\ell < k$*,* $\mathbb{E}\left[\mathrm{d}(\widehat{x}_\ell^k, X^*)\right] \leq \mathsf{E}_\ell^k[\mathrm{diam}(X_0)]/\mathsf{S}_\ell^k$*.*

Proof. Fix $\tau > 1$. We invoke Lemma 6. We take the total expectation in (3.13) and sum from ℓ to k, obtaining

$$2\rho \sum_{i=\ell}^{k} \alpha_i \mathbb{E}\left[\mathrm{d}(x^i, X^*)\right] \leq \mathbb{E}\left[\mathrm{d}(x^\ell, X^*)^2\right] + \sum_{i=\ell}^{k} \left[\frac{\left(\rho + \sqrt{2C_F}\right)^2}{\mathsf{A}_{i,\tau}} + 2\left(1 + \mathsf{B}_{i,\tau}\right) C_F^2 \right] \alpha_i^2$$

$$\leq \mathbb{E}\left[\mathrm{d}(x^\ell, X^*)^2\right] + \mathsf{G}_\tau \left(\rho + \sqrt{2C_F}\right)^2 \mathsf{b}_\ell^k + \mathsf{H}_\tau C_F^2 \mathsf{a}_\ell^k, \tag{3.52}$$

using $\beta_i(2 - \beta_i) \in (0, 1]$ and the definitions of $\mathsf{A}_{i,\tau}$, $\mathsf{B}_{i,\tau}$, G_τ, H_τ, a_ℓ^k and b_ℓ^k in last inequality. From (3.52) on, the proofs of items (a)-(c) are similar to the proof of Theorem 6. We omit the details, but make the following remarks: differently from the proofs of items (a)-(b) in Theorem 6, the proofs of items (a)-(b) of Theorem 7 do not require Proposition 2. In the proof of item (c), we use the bound $\mathbb{E}[\mathrm{d}(x^\ell, X^*)^2] \leq \mathrm{diam}(X_0)^2$ in (3.52). □

Corollary 2 (Rate of convergence with robust stepsizes: bounded case)**.** *Assume that the hypotheses of Theorem 7 hold. Given* $\theta > 0$ *and* $\lambda > 0$*, define* $\{\alpha_k\}$ *as:* $\alpha_0 = \alpha_1 = \theta$ *and for* $k \geq 2$*,*

$$\alpha_k := \frac{\theta}{\sqrt{k \left(\ln k\right)^{1+\lambda}}}, \tag{3.53}$$

and choose $\beta_k \equiv \beta \in (0, 2)$*,* $\tau > 1$*. Define* $\widehat{\mathsf{J}} := \mathsf{H}_\tau C_F^2 + \mathsf{G}_\tau \left(\rho + \sqrt{2C_F}\right)^2 \beta^{-1}(2 - \beta)^{-1}$*.*

Then $\{\mathrm{d}(x^k, X^*)\}$ *a.s.-converges to* 0 *and*

a) For every $\epsilon > 0$, there exists $M = M_\epsilon \geq 2$ such that $\mathbb{E}\left[\mathrm{d}(x^M, X^*)\right] < \epsilon$ with

$$\epsilon \leq \frac{\max\{\theta, \theta^{-1}\}}{2\rho} \cdot \frac{[\ln(M-1)]^{\frac{1+\lambda}{2}}}{\sqrt{M-1}} \cdot \left\{ \mathrm{d}(x^0, X^*)^2 + \widehat{\mathsf{J}} \left[2 + \frac{1}{2(\ln 2)^{1+\lambda}} + \frac{1}{\lambda(\ln 2)^\lambda} \right] \right\},$$

b) for all $k \geq 2$,

$$\mathbb{E}\left[\mathrm{d}(\widehat{x}^k, X^*)\right] \leq$$

$$\frac{\max\{\theta, \theta^{-1}\}}{2\rho} \cdot \frac{(\ln k)^{\frac{1+\lambda}{2}}}{\sqrt{k}} \cdot \left\{ \mathrm{d}(x^0, X^*)^2 + \widehat{\mathsf{J}} \left[2 + \frac{1}{2(\ln 2)^{1+\lambda}} + \frac{1}{\lambda(\ln 2)^\lambda} \right] \right\},$$

c) if X_0 is compact, then given $r \in (0,1)$, for all $k \geq 2r^{-1}$, it holds that

$$\mathbb{E}\left[\mathrm{d}(\widehat{x}^k_{\lceil rk \rceil}, X^*)\right] \leq$$

$$\frac{\max\{\theta, \theta^{-1}\}}{2\rho} \cdot \frac{(\ln k)^{\frac{1+\lambda}{2}}}{\sqrt{k}} \cdot \left\{ (1-r)^{-1} \operatorname{diam}(X_0)^2 + \frac{r^{-1}\widehat{\mathsf{J}}}{[\ln k - \ln(1/r)]^{1+\lambda}} \right\}.$$

Proof. We assume that we have $M \geq 2$ in (3.45). Item (a) follows from (3.50) with $k := M-1$ and $\ell := 0$, (3.51), Theorem 7(a), the definition of $\widehat{\mathsf{J}}$, $\mathsf{E}_0^\infty[\mathrm{d}(x^0, X^*)]$ and $\mathsf{a}_0^\infty = \beta(2-\beta)\mathsf{b}_0^\infty$.

Similarly, item (b) follows from (3.50)-(3.51) with $\ell := 0$, Theorem 7(b), the definition of $\widehat{\mathsf{J}}$, $\mathsf{E}_0^k[\mathrm{d}(x^0, X^*)]$ and the facts that $\mathsf{b}_0^k \leq \mathsf{b}_0^\infty$ and $\mathsf{a}_0^\infty = \beta(2-\beta)\mathsf{b}_0^\infty$.

We now prove item (c). Let $r \in (0,1)$, $k \geq 2r^{-1}$ and set $\ell := \lceil rk \rceil$. We have $\ell \geq 2$ and $rk \leq \ell \leq rk+1$. We estimate

$$\mathsf{a}_\ell^k = \sum_{i=\ell}^k \alpha_i^2 = \sum_{i=\ell}^k \frac{\theta^2}{i(\ln i)^{1+\lambda}} \leq \frac{\theta^2(k-\ell+1)}{\ell(\ln \ell)^{1+\lambda}}. \tag{3.54}$$

From (3.50) and (3.54) we have

$$\frac{\mathsf{a}_\ell^k}{\mathsf{S}_\ell^k} \leq \frac{\theta\sqrt{k(\ln k)^{1+\lambda}}}{\ell(\ln \ell)^{1+\lambda}} \leq \frac{\theta r^{-1}}{\sqrt{k}} \cdot \frac{\sqrt{(\ln k)^{1+\lambda}}}{(\ln(rk))^{1+\lambda}} = \theta r^{-1} \frac{(\ln k)^{\frac{1+\lambda}{2}}}{\sqrt{k}\,[\ln k - \ln(1/r)]^{1+\lambda}}, \tag{3.55}$$

$$\frac{1}{\mathsf{S}_\ell^k} \leq \frac{\theta^{-1}\sqrt{k(\ln k)^{1+\lambda}}}{k-\ell+1} \leq \theta^{-1}(1-r)^{-1}\frac{(\ln k)^{\frac{1+\lambda}{2}}}{\sqrt{k}}, \tag{3.56}$$

using the inequality $\ell \geq rk$ in the second inequality of (3.55) and $k-\ell+1 \geq (1-r)k$ in the second inequality of (3.56). Item (c) follows from (3.55)-(3.56), Theorem 7(c), the definitions of $\widehat{\mathsf{J}}$ and $\mathsf{E}_\ell^k[\operatorname{diam}(X_0)]$, and the fact that $\beta(2-\beta)\mathsf{b}_\ell^k = \mathsf{a}_\ell^k$. □

Remark 3. Corollary 2(c) implies that, if X_0 is compact, then $\{\widehat{x}^k_{\lceil rk\rceil}\}$ has a better performance than $\{x^k\}$ and $\widehat{x}^k$ when stepsizes as in (3.53) are used. Indeed, in Corollary 2(c), $\lambda > 0$ can be arbitrarily small, without affecting the constant in the convergence rate, and the "stochastic error" $r^{-1}\widehat{\mathsf{J}}\,[\ln k - \ln(1/r)]^{-(1+\lambda)}$ decays to zero. For unbounded operators, (3.49) in Corollary 1 suggests the use of $\lambda > 1$ and $\theta \sim L$ so that k_0 does not become too large. As an example, if $\tau = 1.5$, $\theta = L$, $\beta = 1$, $\phi = 0.5$ and $\lambda = 2$, we have $k_0 = 11$.

In Corollaries 1-2, stepsizes of $O(1)k^{-1/2}(\ln k)^{-(1+\lambda)/2}$ are small enough to guarantee asymptotic a.s.-convergence and large enough as to ensure a rate of $O(1)k^{-1/2}(\ln k)^{(1+\lambda)/2}$. If asymptotic a.s.-convergence of the whole sequence is not the main concern, we show next that one may use larger stepsizes of $O(1)k^{-1/2}$ for ensuring convergence in L^1 (hence convergence in probability and a.s.-convergence of a subsequence) with a convergence rate of $O(1)k^{-1/2}$. When a constant stepsize α is used in method (3.4)-(3.5), we can also give an error bound on the performance proportional to α. Precisely, we have $\mathbb{E}[\mathrm{d}(\widehat{x}^k, x^*)] \lesssim k^{-1} + O(\alpha)$. Such error bounds justify rigorously the practical use of constant stepsizes in incremental methods for machine learning, where only an inexact solution is required.

Corollary 3 (Convergence rates for large stepsizes: bounded case)**.** *Assume that the hypotheses of Theorem 7 hold. Recall the definition of* $\widehat{\mathsf{J}}$ *in Corollary 2. Choose* $\theta > 0$, $\beta_k \equiv \beta \in (0,2)$ *and* $\tau > 1$.

a) If we choose a constant stepsize $\alpha_k \equiv \theta\alpha$, *then for all* $k \geq 1$,

$$\mathbb{E}\left[\mathrm{d}(\widehat{x}^k, X^*)\right] \leq \frac{\max\{\theta, \theta^{-1}\}}{2\rho}\left\{\frac{\mathrm{d}(x^0, X^*)^2}{\alpha(k+1)} + \widehat{\mathsf{J}}\alpha\right\}.$$

b) If the total number of iterations $\mathsf{K} \geq 1$ *is given a priori and for all* $k \in [\mathsf{K}]$, $\alpha_k \equiv \frac{\theta\,\mathrm{d}(x^0,X^*)}{\sqrt{\widehat{\mathsf{J}}(\mathsf{K}+1)}}$, *then*

$$\mathbb{E}\left[\mathrm{d}(\widehat{x}^{\mathsf{K}}, X^*)\right] \leq \frac{\max\{\theta, \theta^{-1}\}}{\rho} \cdot \frac{\mathrm{d}(x^0, X^*)\sqrt{\widehat{\mathsf{J}}}}{\sqrt{\mathsf{K}+1}}.$$

c) If X_0 *is compact and we choose* $\alpha_0 := \theta$ *and for* $k \geq 1$, $\alpha_k := \frac{\theta}{\sqrt{k}}$, *then, given* $r \in (0,1)$, *for all* $k \geq r^{-1}$,

$$\mathbb{E}\left[\mathrm{d}(\widehat{x}^k_{\lceil rk\rceil}, X^*)\right] \leq \frac{\max\{\theta, \theta^{-1}\}}{2\rho} \cdot \frac{1}{\sqrt{k}} \cdot \left\{(1-r)^{-1}\,\mathrm{diam}(X_0)^2 + r^{-1}\widehat{\mathsf{J}}\right\}.$$

Proof. Item (a) follows from Theorem 7(b) and the definitions of $\widehat{\mathsf{J}}$, $\mathsf{E}_0^k[\mathrm{d}(x^0, X^*)]$, S_0^k, a_0^k and b_0^k. Item (b) follows by setting $k := \mathsf{K}$ and minimizing the right hand side inequality in item (a) with respect to α. We prove now item (c). Take $r \in (0,1)$, $k \geq r^{-1}$ and set $\ell := \lceil rk \rceil$. We have $\ell \geq 1$ and $rk \leq \ell \leq rk + 1$. We estimate

$$\mathsf{S}_\ell^k = \sum_{i=\ell}^{k} \alpha_i \geq \frac{\theta(k - \ell + 1)}{\sqrt{k}}, \tag{3.57}$$

using the fact that the minimum stepsize between ℓ and $k \geq 2$ is $\theta k^{-\frac{1}{2}}$. We also estimate

$$\mathsf{a}_\ell^k = \sum_{i=\ell}^{k} \alpha_i^2 = \sum_{i=\ell}^{k} \frac{\theta^2}{i} \leq \frac{\theta^2(k - \ell + 1)}{\ell}. \tag{3.58}$$

From (3.57)-(3.58) we have

$$\frac{\mathsf{a}_\ell^k}{\mathsf{S}_\ell^k} \leq \frac{\theta\sqrt{k}}{\ell} \leq \frac{\theta r^{-1}}{\sqrt{k}}, \tag{3.59}$$

$$\frac{1}{\mathsf{S}_\ell^k} \leq \frac{\theta^{-1}\sqrt{k}}{k - \ell + 1} \leq \frac{\theta^{-1}(1-r)^{-1}}{\sqrt{k}}, \tag{3.60}$$

using $\ell \geq rk$ in the second inequality of (3.59) and $k - \ell + 1 \geq (1-r)k$ in the second inequality of (3.60). Item (c) follows from (3.59)-(3.60), Theorem 7(c), the definitions of $\widehat{\mathsf{J}}$ and $\mathsf{E}_\ell^k[\mathrm{diam}(X_0)]$, and the fact that $\beta(2-\beta)\mathsf{b}_\ell^k = \mathsf{a}_\ell^k$. □

We make a remark concerning the *robustness* of the stepsize sequence in Corollaries 1, 2 and 3 in the spirit of Nemirovski et al. [60]. The stepsizes presented above are robust in the sense that the knowledge of L is not required, and does not affect the progress of the method. Also, a scaling of θ in the stepsize implies a scaling in the convergence rate which is linear in $\max\{\theta, \theta^{-1}\}$. It is important to remark that by Corollary 1 this property holds true even in the case of an unbounded operator, a result which appears to be new, as well as the use of robust stepsizes with approximate projections, which greatly decreases the computational effort. Also, these two new features still hold with a near optimal rate (up to logarithmic terms).

We close this section by showing that, in the case of stochastic approximation, the weak sharpness property implies that after a finite number of iterations an

auxiliary stochastic program with linear objective solves the original variational inequality. This recovers a similar property satisfied in the deterministic setting (see [53], Theorem 5.1). We give exact values of the minimum number of iterations in terms of the condition number L/ρ^2, the variance and the distance of x^0 to the solution set, in case T is L-Lipschitz continuous.

Corollary 4 (An auxiliary simpler optimization problem)**.** *Suppose that T is (L,δ)-Hölder continuous with $\delta \in (0,1]$ and*

1. *the assumptions of Corollary 1 hold with $\delta = 1$ (unbounded case), or*
2. *the assumptions of Corollary 2 hold (bounded case).*

Then, there exists $\mathsf{V} > 0$, such that for all $k \geq 2$ with $\frac{k}{(\ln k)^{1+\lambda}} > \left(\frac{\mathsf{V} L^\delta}{\rho^{1+\delta}}\right)^2$, we have

$$\operatorname*{argmin}_{x\in X} \langle \mathbb{E}\left[F(v,\widehat{x}^k)\right], x\rangle \subset X^*.$$

Moreover, under condition 1,

$$\mathsf{V} := \frac{\max\{\theta,\theta^{-1}\}}{2} \cdot \inf_{x^*\in X^*}\left\{\|x^0 - x^*\|^2 + \mathsf{J}(x^*)\left[2 + \frac{1}{2(\ln 2)^{1+\lambda}} + \frac{1}{\lambda(\ln 2)^\lambda}\right]\right\},$$

while, under condition 2,

$$\mathsf{V} := \frac{\max\{\theta,\theta^{-1}\}}{2} \cdot \left\{\mathrm{d}(x^0, X^*)^2 + \widehat{\mathsf{J}}\left[2 + \frac{1}{2(\ln 2)^{1+\lambda}} + \frac{1}{\lambda(\ln 2)^\lambda}\right]\right\}.$$

Proof. Call $\bar{x}^k := \Pi_{X^*}(\widehat{x}^k)$. By the choice of k, the definition of V and item (b) of Corollaries 1-2 we have

$$\mathbb{E}\left[\|\widehat{x}^k - \bar{x}^k\|\right] = \mathbb{E}\left[\mathrm{d}(\widehat{x}^k, X^*)\right] < (\rho/L)^\delta. \tag{3.61}$$

From the Hölder-continuity of T,

$$\left\|\mathbb{E}[T(\widehat{x}^k)] - \mathbb{E}[T(\bar{x}^k)]\right\| \leq$$

$$\mathbb{E}\left[\left\|T(\widehat{x}^k) - T(\bar{x}^k)\right\|\right] \leq L\mathbb{E}\left[\|\widehat{x}^k - \bar{x}^k\|^\delta\right] \leq L\mathbb{E}\left[\|\widehat{x}^k - \bar{x}^k\|\right]^{\frac{1}{\delta}} < \rho, \tag{3.62}$$

using Jensen's inequality in the first inequality, Hölder's inequality in third inequality and (3.61) in last inequality.

From Proposition 1, Assumption 6 and the equivalence between (2.14) and (2.15), we get that the Euclidean ball with center $-T(\bar{x}^k)$ and radius ρ is contained in $\bigcap_{x\in X^*}[\mathbb{T}_X(x)\cap\mathbb{N}_{X^*}(x)]^\circ$. By the convexity of the ball and Jensen's inequality, we have

$$-\mathbb{E}\left[T(\bar{x}^k)\right]+\rho B(0,1)\subset\bigcap_{x\in X^*}[\mathbb{T}_X(x)\cap\mathbb{N}_{X^*}(x)]^\circ. \tag{3.63}$$

From (3.62) and (3.63) we get that $-\mathbb{E}[T(\widehat{x}^k)]\in\operatorname{int}\left(\bigcap_{x\in X^*}[\mathbb{T}_X(x)\cap\mathbb{N}_{X^*}(x)]^\circ\right)$. Hence we conclude from Theorem 4 that

$$\operatorname*{argmin}_{x\in X}\langle\mathbb{E}[T(\widehat{x}^k)],x\rangle\subset X^*. \tag{3.64}$$

Finally, we observe that $\mathbb{E}\left[T(\widehat{x}^k)\right]=\mathbb{E}\left[\mathbb{E}\left[F(v,\widehat{x}^k)\middle|\mathcal{F}_k\right]\right]=\mathbb{E}[F(v,\widehat{x}^k)]$, using Assumption 4, $\widehat{x}^k\in\mathcal{F}_k$ and the property $\mathbb{E}[\mathbb{E}[\cdot|\mathcal{F}_k]]=\mathbb{E}[\cdot]$. The result follows from (3.64) and the fact that $\mathbb{E}\left[T(\widehat{x}^k)\right]=\mathbb{E}[F(v,\widehat{x}^k)]$. □

We remark [2] that, when X is a compact polyhedron, T is (L,δ)-Hölder continuous and $\rho^{1+\delta}/L^\delta\gg\epsilon$ for a specified tolerance $\epsilon>0$, solving the linear program in Corollary 4 may be more advantageous than maintaining method (3.4)-(3.5). Indeed, N repeated runs of method (3.4)-(3.5) with the same initial iterate until iteration k (as stated in Corollary 4), produce N i.i.d. samples $\{(\bar{v}_i,\widehat{x}_i^k)\}_{i=1}^N$ of $(v,\widehat{x}^k_{\lceil k/2\rceil})$, which may be used to estimate the coefficient $\mathbb{E}[F(v,\widehat{x}^k_{\lceil k/2\rceil})]$. The oracle complexity needed to produce an ϵ-approximated solution of a stochastic linear program (SLP) via the SAA method is $N\sim 1/\epsilon^2$. Hence, we may solve the linear program $\min_X\langle\frac{1}{N}\sum_{i=1}^N F(\bar{v}_i,\widehat{x}_i^k),x\rangle$ by efficient methods (such as the Simplex method or interior point methods), after $kN\sim[L^\delta/(\rho^{1+\delta}\epsilon)]^2$ total iterations of method (3.4)-(3.5). Moreover, solving this SLP produces a ϵ-solution of the SVI with *high-probability* (not just in *mean*, as stated in Corollaries 2-3).

[2] Supposing that X_0 is compact, the results of Corollary 4 may be refined: for all $r\in(0,1)$, there exists $\mathsf{V}>0$, such that for all $k>(\mathsf{V}L^\delta/\rho^{1+\delta})^2$, $\operatorname{argmin}_{x\in X}\langle\mathbb{E}[F(v,\widehat{x}^k_{\lceil rk\rceil})],x\rangle\subset X^*$. The proof uses Corollary 3(c).

3.2 An incremental projection method with Tykhonov regularization

3.2.1 Cartesian structure

We assume in this section that the stochastic variational inequality (1.1)-(1.2) has a Cartesian structure. We consider the decomposition $\mathbb{R}^n = \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m}$, with $n = n_1 + \ldots + n_m$ and furnish this Cartesian space with the standard inner product $\langle x, y \rangle = \sum_{j=1}^m \langle x_j, y_j \rangle$, for $x = (x_1, \ldots, x_m)$ and $y = (y_1, \ldots, y_m)$. We suppose that the feasible set $X \subset \mathbb{R}^n$ has the form $X = X^1 \times \cdots \times X^m$, where each component $X^j \subset \mathbb{R}^{n_j}$ is a closed and convex set for $j \in [m]$.

We also assume that the random operator $F : \Xi \times \mathbb{R}^n \to \mathbb{R}^n$ has the form $F = (F_1, \ldots, F_m)$, where each component is of the form $F_j : \Xi \times \mathbb{R}^n \to \mathbb{R}^{n_j}$ for $j \in [m]$. From (1.2), the mean operator has the form $T = (T_1, \ldots, T_m)$ with $T_j(x) = \mathbb{E}[F_j(v, x)]$ for $j \in [m]$. We emphasize that the orthogonal projection under a Cartesian structure is simple: for $x = (x_1, \ldots, x_m) \in \mathbb{R}^n$ and $Y = Y^1 \times \ldots \times Y^m \subset \mathbb{R}^n$ with $x_j \in \mathbb{R}^{n_j}$ and $Y^j \subset \mathbb{R}^{n_j}$, we have $\Pi_Y(x) = (\Pi_{Y^1}(x_1), \ldots, \Pi_{Y^m}(x_m))$.

3.2.2 Constraint structure

In order to exploit the use of incremental projections (as in Section 3.1) in the Cartesian framework, we assume from now on that for $j \in [m]$, each Cartesian component X^j of $X = X^1 \times \ldots \times X^m$ has the following form:

$$X^j = X_0^j \cap \left(\cap_{i \in \mathcal{I}_j} X_i^j \right), \tag{3.65}$$

where $\{X_0^j\} \cup \{X_i^j : i \in \mathcal{I}_j\}$ is a collection of closed and convex subsets of $\mathbb{R}^{n_j}$. Given $j \in [m]$, we assume that the projection operator onto X_0^j is computationally easy to evaluate, and that for every $i \in \mathcal{I}_j$, X_i^j is representable in $\mathbb{R}^{n_j}$ as

$$X_i^j = \{x \in \mathbb{R}^{n_j} : g_i(j|x) \leq 0\}, \tag{3.66}$$

for some convex function $g_i(j|\cdot) : \mathbb{R}^{n_j} \to \mathbb{R} \cup \{\infty\}$ with domain $\operatorname{dom} g_i(j|\cdot) \subset X_0^j$. We shall denote the positive part of $g_i(j|\cdot)$ as $g_i^+(j|x) := \max\{g_i(j|x), 0\}$, for $x \in \mathbb{R}^{n_j}$. We also assume that, for every $i \in \mathcal{I}_j$, the subgradients of $g_i^+(j|\cdot)$ at

points $x \in X_0^j - X_i^j$ are easily computable and that $\{\partial g_i^+(j|\cdot) : i \in \mathcal{I}_j\}$ is uniformly bounded over X_0^j, i.e., there exists $C_g^j > 0$ such that

$$\|d\| \leq C_g^j, \tag{3.67}$$

for all $x \in X_0^j$, all $i \in \mathcal{I}_j$ and all $d \in \partial g_i^+(j|x)$.

3.2.3 Statement of the algorithm

The idea of our Tykhonov method consists of combining the stochastic approximation proposed by the explicit iterative Tykhonov method in [50], for coping with the *monotone* case, with the incremental projection method proposed in [57], exploiting *simpler constraint structures*, which reduce significantly the computational complexity. Our method improves over the results of [73], by proposing an incremental projection method for a SVI which is monotone but not strongly monotone. Our method also generalizes the work in [73] in the sense that it allows the *distributed solution of Cartesian* SVIs with approximate projections, and includes a larger class of closed and convex feasible sets (see item (i) in Subsection 1.5.1).

For problems endowed with the Cartesian structure and the constraint structure of Sections 3.2.1 and 3.2.2, our method advances in a distributed fashion for each Cartesian component $j \in [m]$, as in the incremental projection method (3.4)-(3.5) with an additional Tykhonov regularization (in order to cope with the monotone case). Precisely, fix the Cartesian component $j \in [m]$. In a first stage, given the current iterate x^k, the method advances in the direction $-F_j(v^k, x^k) - \epsilon_{k,j} x_j^k$, after taking the sample v^k of v, producing an auxiliary iterate y_j^k, where $\epsilon_{k,j} > 0$ is a regularization parameter. In the second stage, a soft constraint $X_{\omega_{k,j}}^j$ is randomly chosen with the random control $\omega_{k,j} \in \mathcal{I}_j$, and the method advances in the direction opposite to a subgradient of $g_{\omega_{k,j}}^+(j|\cdot)$ at the point y_j^k, producing the next iterate x_j^{k+1}. The iterates are collected in x^{k+1} and the method continues. Formally, the method takes the form:

Algorithm 2 (Incremental projection and Tykhonov regularization)**.**

1. **Initialization:** *Choose the initial iterate* $x^0 \in \mathbb{R}^n$*, the stepsize sequences* $\alpha^k = (\alpha_{k,1}, \ldots, \alpha_{k,m}) \in (0, \infty)^m$ *and* $\beta^k = (\beta_{k,1}, \ldots, \beta_{k,m}) \in (0, 2)^m$*, the*

regularization sequence $\epsilon^k = (\epsilon_{k,1}, \ldots, \epsilon_{k,m}) \in (0,\infty)^m$*, the random control sequence* $\omega^k = (\omega_{k,1}, \ldots, \omega_{k,m}) \in \mathcal{I}_1 \times \ldots \times \mathcal{I}_m$ *and the operator samples* $\{v^k\}$*.*

2. **Iterative step:** *Given* $x^k = (x_1^k, \ldots, x_m^k)$*, define, for each* $j \in [m]$*,*

$$y_j^k = \Pi_{X_0^j}\left[x_j^k - \alpha_{k,j}\left(F_j(v^k, x^k) + \epsilon_{k,j} x_j^k\right)\right], \tag{3.68}$$

$$x_j^{k+1} = \Pi_{X_0^j}\left[y_j^k - \beta_{k,j}\frac{g^+_{\omega_{k,j}}(j|y_j^k)}{\|d_j^k\|^2} d_j^k\right], \tag{3.69}$$

where $d_j^k \in \partial g^+_{\omega_{k,j}}(j|y_j^k) - \{0\}$ *if* $g_{\omega_{k,j}}(j|y_j^k) > 0$*, and* $d_j^k = d$ *for any* $d \in \mathbb{R}^{n_j} - \{0\}$ *if* $g_{\omega_{k,j}}(j|y_j^k) \leq 0$*.*

The first stage (3.68) of the iterative step can be written compactly as

$$y^k = \Pi_{X_0}\left[x^k - D(\alpha_k)\cdot\left(F(v^k, x^k) + D(\epsilon_k)x^k\right)\right],$$

where $X_0 := X_0^1 \times \ldots \times X_0^m$.

3.2.4 Discussion of the assumptions

We consider the natural filtration

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

Assumption 9. *We request Assumptions 1-5 and Assumption 3(i).*

We now state the assumptions concerning the approximate projections which accommodate the Cartesian structure. Basically, we require each Cartesian component X^j given by (3.65) to satisfy Assumption 7. This is formally stated in Assumption 10. Also, as in [50], the stepsize and regularization sequences require a partial coordination specified in Assumption 11.

Assumption 10 (Constraint sampling and regularity)**.** *For each* $j \in [m]$*, there exists* $c^j > 0$*, such that a.s. for all* $k \in \mathbb{N}$ *and all* $x \in X_0^j$*,*

$$\mathrm{d}(x, X^j)^2 \leq c^j \cdot \mathbb{E}\left[g^+_{\omega_{k,j}}(j|x)\middle|\mathcal{F}_k\right].$$

We observe that Assumption 10 requires a sampling coordination between the control sequences $\{\omega_{k,j}\}_{k=0}^{\infty}$ for $j \in [m]$ since the filtration $\mathcal{F}_k$ accumulates the history from the control sequence of every Cartesian component. Such assumption is satisfied, e.g., when $\{\omega_{k,j}, v^k : j \in [m]\}$ is an independent sequence and, for each $j \in [m]$, X^j satisfies nondemanding regularity assumptions (e.g., a Slater condition).

Assumption 11 (Partial coordination of stepsizes and regularization sequences)**.** *Let* $u_{k,\min} := \min_{j\in[m]} u_{k,i}$, $u_{k,\max} := \max_{j\in[m]} u_{k,j}$ *for* $u \in \{\alpha, \beta, \epsilon\}$. *Denote* $\Delta_k := \alpha_{k,\max} - \alpha_{k,\min}$, $\Gamma_k := \epsilon_{k,\max} - \epsilon_{k,\min}$ *and* $\Theta_k := \beta_{k,\min}(2 - \beta_{k,\max})$. *Then,*

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

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

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

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

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

Assumption 11 contains usual conditions on the regularization parameters of Tykhonov algorithms and on the stepsize for SA algorithms, with certain coordination across stepsizes and regularization parameters. Assumption 11 includes Assumption 2 in [50] with the following addition, due to the use of approximate projections:

$$\sum_{k=0}^{\infty} \frac{(\alpha_{k,\max} - \alpha_{k,\min})^2}{\alpha_{k,\min}\epsilon_{k,\min}} < \infty. \tag{3.70}$$

We observe that this condition is trivially satisfied when $\alpha_{k,j} = \alpha_{k,\ell}$ for all k, j, ℓ. Lemma 4 of [50] establishes that stepsizes and regularization parameters of the form $\alpha_{k,j} = (k + C_j)^{-c}$ and $\epsilon_{k,j} = (k + D_j)^{-d}$, satisfy Assumption 2 in [50], when $c, d \in (0, 1)$ are such that $c > d$ and $c + d < 1$, the C_j's belong to the interval $[\underline{C}, \overline{C}]$ and the D_i's belong to the interval $[\underline{D}, \overline{D}]$ for some $0 < \underline{C} < \overline{C}$ and $0 < \underline{D} < \overline{D}$. These stepsizes and parameters, together with $\beta_{k,j} \equiv \beta_j \in (0, 2)$

for $j \in [m]$, also satisfy our extra condition (3.70) and Assumption 11: indeed, if $C_{\max} = \max_{1\leq i\leq m} C_i$, $C_{\min} = \min_{1\leq i\leq m} C_i$ and $D_{\max} = \max_{1\leq i\leq m} D_i$, then

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

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

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

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

3.2.5 Convergence analysis

We present next our convergence result for method (3.68)-(3.69). We shall need two lemmas.

Lemma 7 (Asymptotic strong-monotonicity)**.** *Defining the operator* $H_k := D(\alpha_k)\cdot (T + D(\epsilon_k))$ *and* $\sigma_k = \alpha_{k,\min}\epsilon_{k,\min} - L(\alpha_{k,\max} - \alpha_{k,\min})$, *then for all* $y, x \in \mathbb{R}^n$ *and* $k \in \mathbb{N}$, $\langle H_k(y) - H_k(x), y - x\rangle \geq \sigma_k\|y - x\|^2$.

Proof. We consider the decomposition

(3.71)

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

Concerning the second term in the right hand side of (3.71), if D_k is the diagonal matrix with entries $(\alpha_1\epsilon_1, \ldots, \alpha_m\epsilon_m)$, then

$$\langle D(\alpha_k)D(\epsilon_k)(y - x), y - x\rangle = \langle D_k(y - x), y - x\rangle \geq \alpha_{k,\min}\epsilon_{k,\min}\|y - x\|^2. \tag{3.72}$$

The first term in the right hand side of (3.71) is equal to

$$\begin{aligned}\sum_{i=1}^m \alpha_{k,i}\langle T_i(y) - T_i(x), y_i - x_i\rangle &= \alpha_{k,\min}\sum_{i=1}^m \langle T_i(y) - T_i(x), y_i - x_i\rangle \\ &\quad + \sum_{i=1}^m (\alpha_{k,i} - \alpha_{k,\min})\langle T_i(y) - T_i(x), y_i - x_i\rangle. \end{aligned} \tag{3.73}$$

The first term in the right hand side of (3.73) is nonnegative by monotonicity of T. For the second term in the right hand side of (3.73), we have

$$\sum_{i=1}^m (\alpha_{k,i} - \alpha_{k,\min})\langle T_i(y) - T_i(x), y_i - x_i\rangle \geq -\sum_{i=1}^m (\alpha_{k,i} - \alpha_{k,\min})\|T_i(y) - T_i(x)\|\|y_i - x_i\|$$

$$\geq -(\alpha_{k,\max} - \alpha_{k,\min}) \sum_{i=1}^{m} \|T_i(y) - T_i(x)\| \|y_i - x_i\|$$
$$\geq -(\alpha_{k,\max} - \alpha_{k,\min}) \|T(y) - T(x)\| \|y - x\|$$
$$\geq -(\alpha_{k,\max} - \alpha_{k,\min}) L \|y - x\|^2, \tag{3.74}$$

using Cauchy-Schwartz inequality in the first inequality, Hölder-inequality in the third one and Lipschitz continuity of T in the last one. The result follows from (3.71)-(3.74). □

We will use the following result proved in Lemma 3 of [50]:

Lemma 8 (Properties of the Tykhonov sequence)**.** *Assume that $X \subset \mathbb{R}^n$ is convex and closed, that the operator $T : \mathbb{R}^n \to \mathbb{R}^n$ is continuous and monotone over X and that Assumption 12 hold. Assume also that the sequences $\{\epsilon_{k,i}\}_{k=1}^{\infty}$ for $i = 1, \ldots, m$ decrease to 0 and satisfy $\limsup_{k\to\infty} \frac{\epsilon_{k,\max}}{\epsilon_{k,\min}} < \infty$, with $\epsilon_{k,\max} = \max_i \epsilon_{k,i}$ and $\epsilon_{k,\min} = \min_i \epsilon_{k,i}$. Denote by t^k the solution of $\mathrm{VI}(T + D(\epsilon_k), X)$. Then*

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

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

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

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

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

Theorem 8 (Asymptotic convergence)**.** *If Assumptions 9-11 hold, then the method (3.68)-(3.69) generates a sequence $\{x^k\}$ such that:*

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

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

Proof. Let $\{t^k\}$ denote the Tykhonov sequence of Lemma 8. We claim that for all $\tau > 1$, $j \in [m]$, $k \in \mathbb{N}$,

$$\|x_j^{k+1} - t_j^k\|^2 \leq \|x_j^k - t_j^k\|^2 - 2\alpha_{k,j}\langle x_j^k - t_j^k, F_j(v^k, x^k) + \epsilon_{k,j}x_j^k\rangle +$$

(3.75)
$$[1 + \tau\beta_{k,j}(2 - \beta_{k,j})]\,\alpha_{k,j}^2 \left\|F_j(v^k, x^k) + \epsilon_{k,j}x_j^k\right\|^2 - \frac{\beta_{k,j}(2-\beta_{k,j})}{(C_g^j)^2}\left(1 - \frac{1}{\tau}\right)\left(g^+_{\omega_{k,j}}(j|x_j^k)\right)^2.$$

Indeed, in view of (3.68)-(3.69) and $t_j^k \in X^j \subset X_0^j \cap X^j_{\omega_{k,j}}$, we can invoke Lemma 2 with $g := g_{\omega_{k,j}}(j|\cdot)$, $x_1 := x_j^k$, $x_2 := x_j^{k+1}$, $x_0 := t_j^k$ $\alpha := \alpha_{k,j}$, $u := F_j(v^k, x^k) + \epsilon_{k,j}x_j^k$, $y := y_j^k$, $\beta := \beta_{k,j}$ and $d := d_j^k$ obtaining (3.75).

We define $z_j^k := x_j^k - \alpha_{k,j}(F_j(v^k, x^k) + \epsilon_{k,j}x_j^k)$ for $j \in [m]$. We sum the inequalities in (3.75) with j between 1 and m, getting

$$\|x^{k+1} - t^k\|^2 \leq \|x^k - t^k\|^2 + 2\sum_{j=1}^m \langle t_j^k - x_j^k, x_j^k - z_j^k\rangle +$$

(3.76)
$$[1 + \tau\beta_{k,\max}(2 - \beta_{k,\min})]\left\|z^k - x^k\right\|^2 - \frac{\beta_{k,\min}(2-\beta_{k,\max})}{C_g^2}\left(1 - \frac{1}{\tau}\right)\sum_{j=1}^m \left(g^+_{\omega_{k,j}}(j|x_j^k)\right)^2,$$

where $\beta_{k,\min} := \min_{j\in[m]} \beta_{k,j}$, $\beta_{k,\max} := \max_{j\in[m]} \beta_{k,j}$ and $C_g := \min_{j\in[m]} C_g^j$.

Concerning the second term in the right hand side of (3.76), Assumption 4 and the fact that $x_j^k \in \mathcal{F}_k$ imply that

$$\mathbb{E}\Big[\langle t_j^k - x_j^k, x_j^k - z_j^k\rangle\Big|\mathcal{F}_k\Big] = \alpha_{k,j}\langle t_j^k - x_j^k, \mathbb{E}\Big[F_j(v^k, x^k)\Big|\mathcal{F}_k\Big] + \epsilon_{k,j}x_j^k\rangle =$$

$$\alpha_{k,j}\langle t_j^k - x_j^k, T_j(x^k) + \epsilon_{k,j}x_j^k\rangle. \tag{3.77}$$

We now analyse the third term in the right hand side of (3.76). The triangular inequality and the inequality $(\sum_{i=1}^4 a_i)^2 \leq 4\sum_{i=1}^4 a_i^2$ imply that

$$\mathbb{E}\Big[\|z_j^k - x_j^k\|^2\Big|\mathcal{F}_k\Big] = \alpha_{k,j}^2\mathbb{E}\Big[\|F_j(v^k, x^k) + \epsilon_{k,j}x_j^k\|^2\Big|\mathcal{F}_k\Big]$$

$$= \alpha_{k,j}^2\mathbb{E}\Big[\|F_j(v^k, x^k) - F_j(v^k, t^k) + \epsilon_{k,j}(x_j^k - t_j^k) + F_j(v^k, t^k) + \epsilon_{k,j}t_j^k\|^2\Big|\mathcal{F}_k\Big]$$

$$\leq 4\alpha_{k,j}^2\mathbb{E}\Big[\|F_j(v^k, x^k) - F_j(v^k, t^k)\|^2\Big|\mathcal{F}_k\Big] + 4\alpha_{k,j}^2\epsilon_{k,j}^2\|x_j^k - t_j^k\|^2$$

$$+4\alpha_{k,j}^2\mathbb{E}\Big[\|F_j(v^k, t^k)\|^2\Big|\mathcal{F}_k\Big] + 4\alpha_{k,j}^2\epsilon_{k,j}^2\|t_j^k\|^2$$

$$\leq 4\alpha_{k,\max}^2 \mathbb{E}\Big[\|F_j(v^k, x^k) - F_j(v^k, t^k)\|^2 \Big| \mathcal{F}_k\Big] + 4\alpha_{k,\max}^2 \epsilon_{k,\max}^2 \|x_j^k - t_j^k\|^2$$

$$+ 4\alpha_{k,\max}^2 \mathbb{E}\Big[\|F_j(v^k, t^k)\|^2 \Big| \mathcal{F}_k\Big] + 4\alpha_{k,\max}^2 \epsilon_{k,\max}^2 \|t_j^k\|^2. \tag{3.78}$$

Summing the inequalities in (3.78) with j between 1 and m, we get from Assumptions 3(i), 5, Lemma 5 and $x^k \in \mathcal{F}_k$,

$$\mathbb{E}\Big[\|z^k - x^k\|^2 \Big| \mathcal{F}_k\Big] = \sum_{j=1}^m \mathbb{E}\Big[\|z_j^k - x_j^k\|^2 \Big| \mathcal{F}_k\Big]$$

$$\leq 4\alpha_{k,\max}^2 \mathbb{E}\Big[\|F(v^k, x^k) - F(v^k, t^k)\|^2 \Big| \mathcal{F}_k\Big] + 4\alpha_{k,\max}^2 \epsilon_{k,\max}^2 \|x^k - t^k\|^2$$

$$+4\alpha_{k,\max}^2 \mathbb{E}\Big[\|F(v^k, t^k)\|^2 \Big| \mathcal{F}_k\Big] + 4\alpha_{k,\max}^2 \epsilon_{k,\max}^2 \|t^k\|^2$$

$$\leq 4L^2\alpha_{k,\max}^2 \|x^k - t^k\|^2 + 4\alpha_{k,\max}^2 \epsilon_{k,\max}^2 \|x^k - t^k\|^2 + 4\alpha_{k,\max}^2 B^2(t^k) + 4\alpha_{k,\max}^2 \epsilon_{k,\max}^2 \|t^k\|^2$$

$$\leq 4(L^2 + \epsilon_{k,\max}^2)\alpha_{k,\max}^2 \|x^k - t^k\|^2 + 4\alpha_{k,\max}^2 (B_t^2 + \epsilon_{k,\max}^2 M_t^2), \tag{3.79}$$

where the last inequality follows from the fact that B_t and M_t are positive constants (depending on the Tykhonov sequence) satisfying $\max_{k\in\mathbb{N}} \|B(t^k)\| \leq B_t$ and $\max_{k\in\mathbb{N}} \|t^k\| \leq M_t$, because $\{t^k\}$ is a bounded sequence by Lemma 8, and B is a nonnegative locally bounded function by Assumption 5 and Lemma 5.

We now analyse the last term in the right hand side of (3.76). Denoting $C := \max_{j\in[m]} c^j$, we get from Assumption 10 and $x_i^k \in \mathcal{F}_k$,

$$\sum_{j=1}^m \mathbb{E}\left[\left(g^+_{\omega_{k,j}}(j|x_j^k)\right)^2 \Big| \mathcal{F}_k\right] \geq \sum_{j=1}^m \frac{1}{c^j} \|\Pi_{X^j}(x_j^k) - x_j^k\|^2$$

$$\geq \frac{1}{C} \sum_{j=1}^m \|\Pi_{X^j}(x_j^k) - x_j^k\|^2 = \frac{1}{C}\, \mathrm{d}(x^k, X)^2. \tag{3.80}$$

Now we use again the fact that $x^k \in \mathcal{F}_k$, take the conditional expectation in (3.76) and combine the result with (3.77)-(3.80), in order to obtain

$$\mathbb{E}\Big[\|x^{k+1} - t^k\|^2 \Big| \mathcal{F}_k\Big] \leq \Big[1 + H_{k,\tau}(L^2 + \epsilon_{k,\max}^2)\alpha_{k,\max}^2\Big] \|x^k - t^k\|^2 +$$

$$2\sum_{j=1}^m \alpha_{k,j} \langle t_j^k - x_j^k, T_j(x^k) + \epsilon_{k,j} x_j^k\rangle$$

$$+ H_{k,\tau}(B_t^2 + M_t^2 \epsilon_{k,\max}^2)\alpha_{k,\max}^2 - A_{k,\tau}\, \mathrm{d}(x^k, X)^2, \tag{3.81}$$

where $H_{k,\tau}, A_{k,\tau}$ and G_τ are defined as

$$H_{k,\tau} := 4\left[1 + \tau\beta_{k,\max}(2 - \beta_{k,\min})\right], \quad A_{k,\tau} := \frac{\beta_{k,\min}(2 - \beta_{k,\max})}{G_\tau}, \quad G_\tau := \frac{CC_g^2\tau}{(\tau - 1)}.$$

The sum in the second term of the right hand side of (3.81) is equal to

$$\langle D(\alpha_k) \cdot (T + D(\epsilon_k))(x^k), t^k - x^k\rangle = \langle D(\alpha_k) \cdot (T + D(\epsilon_k))(x^k)$$

$$-D(\alpha_k) \cdot (T + D(\epsilon_k))(t^k), t^k - x^k\rangle$$

$$+\langle D(\alpha_k)\cdot(T+D(\epsilon_k))(t^k), t^k-\Pi(x^k)\rangle+\langle D(\alpha_k)\cdot(T+D(\epsilon_k))(t^k), \Pi(x^k)-x^k\rangle. \tag{3.82}$$

Calling $\Delta_k := \alpha_{k,\max} - \alpha_{k,\min}$, it follows from Lemma 7 that the first term in the right hand side of (3.82) satisfies

$$\langle D(\alpha_k) \cdot (T + D(\epsilon_k))(x^k) - D(\alpha_k) \cdot (T + D(\epsilon_k))(t^k), t^k - x^k\rangle \leq$$

$$- (\alpha_{k,\min}\epsilon_{k,\min} - L\Delta_k)\|x^k - t^k\|^2. \tag{3.83}$$

The second term in the right hand side of (3.82) is equal to

$$\sum_{j=1}^m \alpha_{k,j}\langle T_j(t^k) + \epsilon_{k,j}t_j^k, t_j^k - \Pi_{X^j}(x_j^k)\rangle = \alpha_{k,\min}\sum_{j=1}^m \langle T_j(t^k) + \epsilon_{k,j}t_j^k, t_j^k - \Pi_{X^j}(x_j^k)\rangle$$

$$+ \sum_{j=1}^m (\alpha_{k,j} - \alpha_{k,\min})\langle T_j(t^k) + \epsilon_{k,j}t_j^k, t_j^k - \Pi_{X^j}(x_j^k)\rangle. \tag{3.84}$$

The first term in the right hand side of (3.84) satisfies

$$\sum_{j=1}^m \langle T_j(t^k) + \epsilon_{k,j}t_j^k, t_j^k - \Pi_{X^j}(x_j^k)\rangle = \langle (T + D(\epsilon_k))(t^k), t^k - \Pi(x^k)\rangle \leq 0, \tag{3.85}$$

since t^k solves $\mathrm{VI}(T + D(\epsilon_k), X)$. Regarding the second term in the right hand side of (3.84), we use the fact that $\Pi_{X^j}(t_j^k) = t_j^k$, so that for each $\mu \in (0, 1)$ we have

$$\sum_{j=1}^m (\alpha_{k,j} - \alpha_{k,\min})\langle T_j(t^k) + \epsilon_{k,j}t_j^k, t_j^k - \Pi_{X^j}(x_j^k)\rangle \leq$$

$$\sum_{j=1}^m (\alpha_{k,j} - \alpha_{k,\min})\|T_j(t^k) + \epsilon_{k,j}t_j^k\|\|\Pi_{X^j}(t_j^k) - \Pi_{X^j}(x_j^k)\|$$

$$\leq \Delta_k \sum_{j=1}^{m} (\|T_j(t^k)\| + \epsilon_{k,j}\|t_j^k\|)\|t_j^k - x_j^k\| \leq$$

$$\Delta_k(B_t + \epsilon_{k,\max}M_t)\|t^k - x^k\| = 2\frac{(B_t + \epsilon_{k,\max}M_t)\Delta_k}{2\sqrt{\mu\alpha_{k,\min}\epsilon_{k,\min}}} \cdot \sqrt{\mu\alpha_{k,\min}\epsilon_{k,\min}}\|t^k - x^k\|$$

$$\leq \frac{(B_t + \epsilon_{k,\max}M_t)^2\Delta_k^2}{4\mu\alpha_{k,\min}\epsilon_{k,\min}} + \mu\alpha_{k,\min}\epsilon_{k,\min}\|t^k - x^k\|^2, \tag{3.86}$$

using Cauchy-Schwartz inequality in the first inequality, Lemma 1(iii) for Π_{X^i} in the second one, the fact that $\|T(t^k)\| \leq B(t^k) \leq B_t$ and $\|t^k\| \leq M_t$ for all $k \in \mathbb{N}$ in the third one, and the relation $2ab = -(a-b)^2 + a^2 + b^2$ in the fourth one. Putting together (3.84)-(3.86), we finally get that the second term in the right hand side of (3.82) is bounded by

(3.87)

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

For the third term in the right hand side of (3.82), we have

$$\begin{aligned} \langle D(\alpha_k)\cdot(T + D(\epsilon_k))(t^k), \Pi(x^k) - x^k\rangle &\leq \|D(\alpha_k)\|\|T(t^k) + \epsilon_k t^k\|\|\Pi(x^k) - x^k\| \\ &\leq \alpha_{k,\max}(B_t + \epsilon_{k,\max}M_t)\,\mathrm{d}(x^k, X). \end{aligned} \tag{3.88}$$

Combining (3.83), (3.87) and (3.88) with (3.82), we obtain

$$2\langle D(\alpha_k)\cdot(T + D(\epsilon_k))(x^k), t^k - x^k\rangle \leq \Big[-2(1-\mu)\alpha_{k,\min}\epsilon_{k,\min} + 2L\Delta_k\Big]\|x^k - t^k\|^2$$

$$+ \frac{(B_t + \epsilon_{k,\max}M_t)^2\Delta_k^2}{2\mu\alpha_{k,\min}\epsilon_{k,\min}} + 2\alpha_{k,\max}(B_t + \epsilon_{k,\max}M_t)\,\mathrm{d}(x^k, X). \tag{3.89}$$

Now we use (3.89) in (3.81), getting

$$\mathbb{E}\Big[\|x^{k+1} - t^k\|^2\Big|\mathcal{F}_k\Big] \leq q_k\|x^k - t^k\|^2 + H_{k,\tau}(B_t^2 + M_t^2\epsilon_{k,\max}^2)\alpha_{k,\max}^2$$

$$+ \frac{(B_t + \epsilon_{k,\max}M_t)^2\Delta_k^2}{2\mu\alpha_{k,\min}\epsilon_{k,\min}} + 2\alpha_{k,\max}(B_t + \epsilon_{k,\max}M_t)\,\mathrm{d}(x^k, X) - A_{k,\tau}\,\mathrm{d}(x^k, X)^2, \tag{3.90}$$

where

$$q_k := 1 - 2(1-\mu)\alpha_{k,\min}\epsilon_{k,\min} + H_{k,\tau}(L^2 + \epsilon_{k,\max}^2)\alpha_{k,\max}^2 + 2L\Delta_k. \tag{3.91}$$

The last term in the right hand side of (3.90) becomes

(3.92)

$$-A_{k,\tau}\,\mathrm{d}(x^k,X)^2+2(B_t+\epsilon_{k,\max}M_t)\alpha_{k,\max}\,\mathrm{d}(x^k,X)\leq\frac{(B_t+\epsilon_{k,\max}M_t)^2\alpha_{k,\max}^2}{A_{k,\tau}},$$

using the relation $2ab\leq\lambda^2a^2+\frac{b^2}{\lambda^2}$ with $\lambda^2:=A_{k,\tau}$, $a:=\mathrm{d}(x^k,X)$ and $b:=(B_t+\epsilon_{k,\max}M_t)\alpha_{k,\max}$. Using (3.92) in (3.90) we get that for all $k\in\mathbb{N}$,

$$\mathbb{E}\Big[\|x^{k+1}-t^k\|^2\Big|\mathcal{F}_k\Big]\leq q_k\|x^k-t^k\|^2$$

$$+\left[H_{k,\tau}(B_t^2+M_t^2\epsilon_{k,\max}^2)+\frac{(B_t+M_t\epsilon_{k,\max})^2}{A_{k,\tau}}\right]\alpha_{k,\max}^2+\frac{(B_t+\epsilon_{k,\max}M_t)^2\Delta_k^2}{2\mu\alpha_{k,\min}\epsilon_{k,\min}}. \tag{3.93}$$

Next we relate $\|x^k-t^k\|^2$ with $\|x^k-t^{k-1}\|^2$, using the properties of the Tykhonov sequence (Lemma 8). We have

$$\|x^k-t^k\|^2\leq(\|x^k-t^{k-1}\|+\|t^k-t^{k-1}\|)^2$$

$$=\|x^k-t^{k-1}\|^2+\|t^k-t^{k-1}\|^2+2\|x^k-t^{k-1}\|\|t^k-t^{k-1}\|$$

$$\leq\|x^k-t^{k-1}\|^2+\left(M_t\frac{\epsilon_{k-1,\max}-\epsilon_{k,\min}}{\epsilon_{k,\min}}\right)^2+2M_t\frac{\epsilon_{k-1,\max}-\epsilon_{k,\min}}{\epsilon_{k,\min}}\|x^k-t^{k-1}\|. \tag{3.94}$$

Using the relation $2ab\leq\lambda^2a^2+\frac{b^2}{\lambda^2}$, the last term in the rightmost expression in (3.94) can be estimated as

$$2M_t\frac{\epsilon_{k-1,\max}-\epsilon_{k,\min}}{\epsilon_{k,\min}}\|x^k-t^{k-1}\|=$$

$$2\sqrt{\alpha_{k,\min}\epsilon_{k,\min}}\|x^k-t^{k-1}\|\cdot M_t\frac{\epsilon_{k-1,\max}-\epsilon_{k,\min}}{\sqrt{\alpha_{k,\min}\epsilon_{k,\min}}\epsilon_{k,\min}}$$

$$\leq\alpha_{k,\min}\epsilon_{k,\min}\|x^k-t^{k-1}\|^2+M_t^2\frac{\big(\epsilon_{k-1,\max}-\epsilon_{k,\min}\big)^2}{\alpha_{k,\min}\epsilon_{k,\min}^3}. \tag{3.95}$$

Putting (3.95) in (3.94) yields

(3.96)

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

We combine (3.93) and (3.96) in order to get

$$\mathbb{E}\Big[\|x^{k+1}-t^k\|^2\Big|\mathcal{F}_k\Big] \le q_k(1+\alpha_{k,\min}\epsilon_{k,\min})\|x^k-t^{k-1}\|^2$$

$$+\left[H_{k,\tau}(B_t^2+M_t^2\epsilon_{k,\max}^2)+\frac{(B_t+M_t\epsilon_{k,\max})^2}{A_{k,\tau}}\right]\alpha_{k,\max}^2+\frac{(B_t+\epsilon_{k,\max}M_t)^2\Delta_k^2}{2\mu\alpha_{k,\min}\epsilon_{k,\min}}$$

$$+q_k\left(M_t\frac{\epsilon_{k-1,\max}-\epsilon_{k,\min}}{\epsilon_{k,\min}}\right)^2\left(1+\frac{1}{\alpha_{k,\min}\epsilon_{k,\min}}\right). \tag{3.97}$$

We now estimate the coefficient $q_k(1+\alpha_{k,\min}\epsilon_{k,\min})$ in (3.97). In view of (3.91), we have

$$q_k = 1-\alpha_{k,\min}\epsilon_{k,\min}\left(2-2\mu-\frac{H_{k,\tau}(L^2+\epsilon_{k,\max}^2)\alpha_{k,\max}^2}{\alpha_{k,\min}\epsilon_{k,\min}}-\frac{2L\Delta_k}{\alpha_{k,\min}\epsilon_{k,\min}}\right). \tag{3.98}$$

Assumption 11(ii) and $0 < H_{k,\tau} = 4[1+\beta_{k,\min}(2-\beta_{k,\max})\tau] \le 4(1+\tau)$ guarantee that

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

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

$$c_k := 2\mu+\frac{H_{k,\tau}(L^2+\epsilon_{k,\max}^2)\alpha_{k,\max}^2}{\alpha_{k,\min}\epsilon_{k,\min}}+\frac{2L\Delta_k}{\alpha_{k,\min}\epsilon_{k,\min}} < c \tag{3.99}$$

for all sufficiently large k. Next we show that $q_k\in(0,1)$ for large k. Indeed, from (3.99) and $c\in(0,1)$ we have that $1<2-c_k<2$ for large enough k, so that we obtain, from (3.98),

$$1-2\alpha_{k,\min}\epsilon_{k,\min} < q_k < 1-\alpha_{k,\min}\epsilon_{k,\min}. \tag{3.100}$$

Finally, $\lim_{k\to\infty}\alpha_{k,\min}\epsilon_{k,\min}=0$ by Assumption 11(ii), so that (3.100) implies that $q_k\in(0,1)$ for sufficiently large k. Using this fact and (3.99) we get the following estimate:

$$0 < q_k(1+\alpha_{k,\min}\epsilon_{k,\min}) \le q_k+\alpha_{k,\min}\epsilon_{k,\min} = 1-\alpha_{k,\min}\epsilon_{k,\min}(2-c_k)+\alpha_{k,\min}\epsilon_{k,\min}$$

$$= 1-\alpha_{k,\min}\epsilon_{k,\min}(1-c_k) \le 1-\alpha_{k,\min}\epsilon_{k,\min}(1-c), \tag{3.101}$$

using (3.99) in the last inequality.

Combining (3.97), (3.101) and $A_{k,\tau} = \beta_{k,\min}(2-\beta_{k,\max})G_\tau^{-1}$, we obtain

$$\mathbb{E}\Big[\|x^{k+1}-t^k\|^2\Big|\mathcal{F}_k\Big] \leq (1-a_k)\|x^k-t^{k-1}\|^2 + b_k, \tag{3.102}$$

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

$$b_k := \left[H_{k,\tau}(B_t^2 + M_t^2\epsilon_{k,\max}^2) + \frac{G_\tau(B_t+M_t\epsilon_{k,\max})^2}{\beta_{k,\min}(2-\beta_{k,\max})}\right]\alpha_{k,\max}^2$$

$$+\frac{(B_t+\epsilon_{k,\max}M_t)^2\Delta_k^2}{2\mu\alpha_{k,\min}\epsilon_{k,\min}} + q_k\Big(M_t\frac{\epsilon_{k-1,\max}-\epsilon_{k,\min}}{\epsilon_{k,\min}}\Big)^2\Big(1+\frac{1}{\alpha_{k,\min}\epsilon_{k,\min}}\Big). \tag{3.103}$$

From (3.101) and $c \in (0,1)$, we conclude that $a_k \in [0,1]$, while from Assumption 11(iii) we have that $\sum_k a_k = \infty$. From Assumption 11(iv) and (3.103) we also have that $\sum_k b_k < \infty$. Finally, denoting $\Gamma_k := \epsilon_{k-1,\max} - \epsilon_{k,\min}$ and $\Theta_k := \beta_{k,\min}(2-\beta_{k,\max})$ we obtain from (3.103):

$$0 \leq \frac{b_k}{a_k} = C_1\frac{\alpha_{k,\max}^2}{\alpha_{k,\min}\epsilon_{k,\min}} + C_2\frac{\alpha_{k,\max}^2}{\Theta_k\alpha_{k,\min}\epsilon_{k,\min}} + C_3\left(\frac{\Delta_k}{\alpha_{k,\min}\epsilon_{k,\min}}\right)^2$$

$$+C_4\frac{\Gamma_k^2}{\epsilon_{k,\min}^3\alpha_{k,\min}}\Big(1+\frac{1}{\alpha_{k,\min}\epsilon_{k,\min}}\Big)$$

for some positive constants C_1, C_2, C_3 and C_4. Therefore, we get $\lim_{k\to\infty} b_k/a_k = 0$ from Assumption 11(ii) and (v). These conditions, Theorem 2 and (3.102) imply that $\lim_{k\to\infty}\|x^k - t^{k-1}\| = 0$ almost surely. The result follows from this fact and Lemma 8. □

Remark 4. *The previous convergence analysis does not present* feasibility *rate guarantees for method 1 neither rate guarantees for method 2. For such type of statements we refer to the improved version of this chapter published in [36].*

3.3 Appendix of Chapter 3

We give the proof of Proposition 1:

Proof. Suppose that (2.18) holds and let $x^* \in X^*$. If $\mathbb{T}_X(x^*) \cap \mathbb{N}_{X^*}(x^*) = \{0\}$, then (2.15) holds trivially. Otherwise, take $d \in \mathbb{T}_X(x^*) \cap \mathbb{N}_{X^*}(x^*)$ with $d \neq 0$.

Since $d \in \mathbb{N}_{X^*}(x^*)$, the definition of $\mathbb{N}_{X^*}(x^*)$ implies that X^* is a subset of the halfspace $H_d^- := \{y : \langle d, y - x^* \rangle \leq 0\}$. In view of (2.10) and $d \in \mathbb{T}_X(x^*)$, there exist sequences $d^k \in \mathbb{R}^n$, $t_k > 0$ such that $x^* + t_k d^k \in X$, $d^k \to d$ and $t^k \to 0$. We claim that, taking a subsequence if needed,

$$x^* + t_k d^k \in X - H_d^-. \tag{3.104}$$

for all k. Indeed, otherwise we would have

$$0 \geq \langle d, x^* + t_k d^k - x^* \rangle = t_k \langle d, d^k \rangle \tag{3.105}$$

for large enough k. Dividing (3.105) by t_k and letting $k \to \infty$ we get $d = 0$ which entails a contradiction. Hence, (3.104) holds. From (2.18), $x^* \in X^*$ and $x^* + t_k d^k \in X$ we get

(3.106)
$$\langle T(x^*), x^* + t_k d^k - x^* \rangle \geq \rho \, \mathrm{d}(x^* + t_k d^k, X^*) \geq \rho \, \mathrm{d}(x^* + t_k d^k, H_d^-) = \rho t_k \frac{\langle d, d^k \rangle}{\|d\|},$$

using (3.104) and $X^* \subset H_d^-$ in second inequality. Dividing (3.106) by t_k and letting $k \to \infty$, we conclude that (2.15) holds for d.

Now suppose that (2.15) holds and that T is constant on X^*. Take $x \in X$, $x^* \in X^*$ and let $\bar{x} := \Pi_{X^*}(x)$. Since $x, \bar{x} \in X$ and X is closed and convex, we have that $x - \bar{x} \in \mathbb{T}_X(\bar{x})$, using the first equality in (2.11). Since T is monotone and X is closed and convex, X^* is closed and convex (see [27], Theorem 2.3.5). From this fact, $\bar{x} = \Pi_{X^*}(x)$ and Lemma 1(i), we obtain that $x - \bar{x} \in \mathbb{N}_{X^*}(\bar{x})$, using the definition of the polar cone. Thus, $x - \bar{x} \in \mathbb{T}_X(\bar{x}) \cap \mathbb{N}_{X^*}(\bar{x})$. We conclude from (2.15) that

$$\langle T(\bar{x}), x - \bar{x} \rangle \geq \rho \|x - \bar{x}\| = \rho \, \mathrm{d}(x, X^*). \tag{3.107}$$

Since T is constant on X^*, we have

(3.108)
$$\langle T(\bar{x}), x - \bar{x} \rangle = \langle T(x^*), x - \bar{x} \rangle = \langle T(x^*), x - x^* \rangle + \langle T(x^*), x^* - \bar{x} \rangle \leq \langle T(x^*), x - x^* \rangle,$$

using the fact that $\langle T(x^*), x^* - \bar{x} \rangle \leq 0$, which holds because $x^* \in X^*$ and $\bar{x} \in X$. The desired claim (2.18) follows from (3.108) and (3.107). □

Chapter 4

A variance-based stochastic extragradient method

Our extragradient method takes the form:

Algorithm 3 (Stochastic extragradient method with stepsize away from zero)**.**

1. **Initialization:** *Choose the initial iterate* $x^0 \in \mathbb{R}^n$*, a positive stepsize sequence* $\{\alpha_k\}$*, the sample rate* $\{N_k\}$ *and initial samples* $\{\xi_j^0\}_{j=1}^{N_0}$ *and* $\{\eta_j^0\}_{j=1}^{N_0}$ *of the random variable* ξ*.*

2. **Iterative step:** *Given iterate* x^k*, generate samples* $\{\xi_j^k\}_{j=1}^{N_k}$ *and* $\{\eta_j^k\}_{j=1}^{N_k}$ *of* ξ *and define:*

$$z^k = \Pi\left[x^k - \frac{\alpha_k}{N_k}\sum_{j=1}^{N_k} F(\xi_j^k, x^k)\right], \tag{4.1}$$

$$x^{k+1} = \Pi\left[x^k - \frac{\alpha_k}{N_k}\sum_{j=1}^{N_k} F(\eta_j^k, z^k)\right], \tag{4.2}$$

where Π is the projection operator onto X. Method (4.1)-(4.2) is designed so that at iteration k the random variable ξ is sampled $2N_k$ times and the empirical average of F at x is used as the approximation of $T(x)$ at each projection step.

In order to incorporate the distributed case mentioned in Section 1.5.2, item(v), we will also analyze the case in which the SVI has a Cartesian structure. We consider the decomposition $\mathbb{R}^n = \prod_{i=1}^m \mathbb{R}^{n_i}$, with $n = \sum_{i=1}^m n_i$, and furnish this

space with the inner product mentioned before Lemma 3. We suppose that the feasible set has the form $X = \prod_{i=1}^m X^i$, where $X^i \subset \mathbb{R}^{n_i}$ is a closed and convex set for $i \in [m]$. The random operator $F : \Xi \times \mathbb{R}^n \to \mathbb{R}^n$ has the form $F = (F_1, \ldots, F_m)$, where $F_i : \Xi \times \mathbb{R}^n \to \mathbb{R}^{n_i}$ for $i \in [m]$. Given $i \in [m]$, we denote by $\Pi_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ the orthogonal projection onto X^i. We emphasize that the orthogonal projection under a Cartesian structure has a simple form: for $x = (x_i)_{i=1}^m \in \mathbb{R}^n$, we have $\Pi_X(x) = (\Pi_{X^1}(x_1), \ldots, \Pi_{X^m}(x_m))$.

In such a setting, the method takes the form:

Algorithm 4 (Stochastic extragradient method with stepsize away from zero: distributed case)**.**

1. **Initialization:** *Choose the initial iterate $x^0 \in \mathbb{R}^n$, the stepsize sequence $\alpha_k = (\alpha_{k,i})_{i=1}^m \in (0,\infty)^m$, the sample rates $N_k = (N_{k,i})_{i=1}^m \in \mathbb{N}^m$ and, for each $i \in [m]$, generate the initial samples $\{\xi_{j,i}^0\}_{j=1}^{N_{0,i}}$ and $\{\eta_{j,i}^0\}_{j=1}^{N_{0,i}}$ of the random variable ξ.*

2. **Iterative step:** *Given $x^k = (x_i^k)_{i=1}^m$, for each $i \in [m]$, generate samples $\{\xi_{j,i}^k\}_{j=1}^{N_{k,i}}$ and $\{\eta_{j,i}^k\}_{j=1}^{N_{k,i}}$ of ξ and define:*

$$z_i^k = \Pi_i\left[x_i^k - \frac{\alpha_{k,i}}{N_{k,i}} \sum_{j=1}^{N_{k,i}} F_i(\xi_{j,i}^k, x^k)\right], \tag{4.3}$$

$$x_i^{k+1} = \Pi_i\left[x_i^k - \frac{\alpha_{k,i}}{N_{k,i}} \sum_{j=1}^{N_{k,i}} F_i(\eta_{j,i}^k, z^k)\right]. \tag{4.4}$$

Method (4.1)-(4.2) is a particular case of method (4.3)-(4.4) with $m = 1$. The only additional requirement when $m > 1$ is the sampling coordination between agents (Assumption 17). We define next the stochastic errors: for each $i \in [m]$,

$$\epsilon_{1,i}^k := \frac{1}{N_{k,i}} \sum_{j=1}^{N_{k,i}} F_i(\xi_{j,i}^k, x^k) - T_i(x^k), \tag{4.5}$$

$$\epsilon_{2,i}^k := \frac{1}{N_{k,i}} \sum_{j=1}^{N_{k,i}} F_i(\eta_{j,i}^k, z^k) - T_i(z^k), \tag{4.6}$$

in which case method (4.3)-(4.4) is expressible in a compact form as:

$$z^k = \Pi[x^k - D(\alpha_k)(T(x^k) + \epsilon_1^k)], \tag{4.7}$$

$$x^{k+1} = \Pi[x^k - D(\alpha_k)(T(z^k) + \epsilon_2^k)], \tag{4.8}$$

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

4.1 Discussion of the assumptions

For simplicity of notation, we aggregate the samples as

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

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

In method (4.7)-(4.8), the sample $\{\xi^k\}$ is used in the first projection while $\{\eta^k\}$ is used in the second projection. In the case of a Cartesian SVI, $\{\xi_i^k\}$ and $\{\eta_i^k\}$ are the samples used in the first and second projections in (4.3)-(4.4) by the i-th agent respectively.

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

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

We observe that by induction, $x^k \in \mathcal{F}_k$ and $z^k \in \widehat{\mathcal{F}}_k$ but $z^k \notin \mathcal{F}_k$. The filtration $\mathcal{F}_k$ corresponds to the information carried until iteration k, to be used on the computation of iteration $k+1$. The filtration $\widehat{\mathcal{F}}_k$ corresponds to the information carried until iteration k plus the information produced at the first projection step of iteration $k+1$, namely, $\widehat{\mathcal{F}}_k = \sigma(\mathcal{F}_k \cup \sigma(\xi^k))$. The way information evolves according to filtrations $\{\mathcal{F}_k, \widehat{\mathcal{F}}_k\}$ is natural in applications. Also, the use of two filtrations will be important since we have $\mathbb{E}[\epsilon_2^k | \mathcal{F}_k] \neq 0$ but $z^k \in \widehat{\mathcal{F}}_k$, so that, given $i \in [m]$:

$$\begin{aligned}
\mathbb{E}[\epsilon_{2,i}^k | \widehat{\mathcal{F}}_k] &= \mathbb{E}\left[\frac{1}{N_{k,i}} \sum_{j=1}^{N_{k,i}} F_i(\eta_{j,i}^k, z^k) - T_i(z^k) \middle| \widehat{\mathcal{F}}_k\right] \\
&= \frac{1}{N_{k,i}} \sum_{j=1}^{N_{k,i}} \mathbb{E}\Big[F_i(\eta_{j,i}^k, z^k) \Big| \widehat{\mathcal{F}}_k\Big] - T_i(z^k) \\
&= \frac{1}{N_{k,i}} \sum_{j=1}^{N_{k,i}} T_i(z^k) - T_i(z^k) = 0, \qquad (4.9)
\end{aligned}$$

if for every $i \in [m]$, $\{\eta_{j,i}^k : j \in [N_{k,i}]\}$ is independent of $\widehat{\mathcal{F}}_k$ and identically distributed as ξ. We exploit (4.9) for avoiding first order moments of the stochastic

errors, which drastically diminishes the complexity by an order of one, and for using martingale techniques [1]. We remark that, with some minor extra effort, the same samples can be used in both projections in method (4.3)-(4.4). Next we describe the assumptions required in our convergence analysis.

Assumption 12 (Consistency). *The solution set* $X^* := \mathrm{S}(T, X)$ *is non-empty.*

Assumption 13 (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,* [2] $\xi : \Omega \to \Xi$ *is a random variable defined on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *and* $\mathbb{E}[\|F(\xi, x)\|] < \infty$ *for all* $x \in X$.

Assumption 14 (Lipschitz continuity). *The mean operator* $T : X \to \mathbb{R}^n$ *defined by* (1.2) *is Lipschitz continuous with modulus* $L > 0$.

Assumption 15 (Pseudo-monotonicity). *The mean operator* $T : \mathbb{R}^n \to \mathbb{R}^n$ *is pseudo-monotone,*[3] *i.e.,* $\langle T(x), z - x\rangle \geq 0 \Longrightarrow \langle T(z), z - x\rangle \geq 0$ *for all* $z, x \in \mathbb{R}^n$.

Assumption 16 (Sample rate). *Given* $\{N_k\}$, *define* $N_{k,\min} := \min_{i\in[m]} N_{k,i}$ *and* $\frac{1}{\mathcal{N}_k} := \frac{1}{n}\sum_{i=1}^m \frac{n_i}{N_{k,i}}$. *Then one of the following conditions is satisfied:*

i) $\sum_{k=0}^\infty \frac{1}{\mathcal{N}_k} < \infty$,

ii) $\sum_{k=0}^\infty \frac{1}{N_{k,\min}} < \infty$.

Note that $\mathcal{N}_k$ is the harmonic average of $\{N_{k,i}\}_{i=1}^m$ with weights $\{n_i/n\}_{i=1}^m$. Hence $\mathcal{N}_k \geq N_{k,\min}$, so that (ii) implies (i). Items (i) and (ii) are different only for the Cartesian SVI (see comment below). Typically a sufficient choice is, for $i \in [m]$:

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

for any $\Theta_i > 0$, $\mu_i > 0$ with $a_i > 0$, $b_i \geq -1$ or $a_i = 0$, $b_i > 0$ (the latter is the minimum requirement). It is essential to specify choices of the above parameters that induce a practical complexity of method (4.3)-(4.4), i.e., practical upper

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

[2] 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.

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

bounds on the total oracle complexity $\sum_{k=1}^{K}\sum_{i=1}^{m} 2N_{k,i}$, where K is an estimate of the total numbers of iterations needed for achieving a given specified tolerance $\epsilon > 0$. A convergence rate in terms of K is also desirable. As commented after Theorem 10, our algorithm achieves an optimal accelerated rate $O(1/K)$ and an optimal complexity $O(\epsilon^{-2})$ up to a first order logarithmic term $\ln(\epsilon^{-1})$. [4]

We offer two options of sampling coordination among the agents:

Assumption 17 (Sampling coordination)**.** *For each $i \in [m]$ and $k \in \mathbb{N}_0$, $\{\xi_i^k\}$ and $\{\eta_i^k\}$ are i.i.d. samples of ξ such that $\{\xi_i^k\}$ and $\{\eta_i^k\}$ are independent of each other. Also, one of the two next coordination conditions is satisfied:*

i) (Centralized sampling) For all $i \in [m]$, $N_{k,i} \equiv N_k$, $\xi_i^k \equiv \xi^k$ and $\eta_i^k \equiv \eta^k$.

ii) (Distributed sampling) $\{\xi^k, \eta^k : k \in \mathbb{N}\}$ is an i.i.d. sample of ξ.

We remark that, with some extra effort, it is possible to use the same samples in each projection step of method (4.3)-(4.4), that is, $\xi_i^k \equiv \eta_i^k$ for $k \in \mathbb{N}_0$ and $i \in [m]$. We ask independence in Assumption 17 to simplify the analysis. Both conditions (i) and (ii) in Assumption 17 are the same for $m = 1$.[5] Assumption 17 implies in particular that $\{\xi^k\}$ is independent of $\mathcal{F}_k$, $\{\eta^k\}$ is independent of $\widehat{\mathcal{F}}_k$

[4] In Machine Learning, the dependence of the rate and complexity estimates on the dimension is relevant in the case of large constraint dimension ($n_i \gg 1$) or large networks ($m \gg 1$). We show our method has complexity $O(n\sigma^2)$ up to a scaling factor in the sample rate, where σ^2 is the variance, even for the case of an unbounded feasible set and a non-uniform variance. Sharper constants are available in case of uniform variance (see Proposition 9). In the case of networks, although Assumption 16(ii) is sufficient, the definition of $\mathcal{N}_k$ could be exploited in the sampling procedure for reducing the dimension dependence or the sampling effort among the agents, if information about the dimensions $\{n_i\}_{i=1}^m$ of the agents' problems is available. Further dimension reduction possibilities are the subject of future work.

[5] In the case when $m > 1$, item (i) corresponds to the case where one stochastic oracle is centralized. In this case, less samples are required but the sampling process needs total coordination. Item (ii) corresponds to the other extreme case, where the agents have completely distributed oracles so that the sampling process of each agent is conducted independently. We do not explore the intermediate possibilities between (i) and (ii). In the case of item (ii), the oracle complexity has higher order dependence in terms of the network dimension m, which may be demanding in the context of large networks ($m \gg 1$). However, if a rapidly decreasing sequence of deterministic exponents $\{b_i\}_{i=1}^m$ is coordinated among agents, then the oracle complexity is linear in m (see Proposition 10) as in the case of centralized sampling.

and both are identically distributed to ξ. In particular, for any $x \in \mathbb{R}^n$, $k \in \mathbb{N}$, $i \in [m]$, $j \in [N_{k,i}]$:

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

Assumption 18 (Stepsize bounded away from zero with partial coordination)**.** *Defining* $\alpha_{k,\min} := \min_{i\in[m]} \alpha_{k,i}$ *and* $\alpha_{k,\max} := \max_{i\in[m]} \alpha_{k,i}$*, the stepsize sequence* $\{\alpha_k\}$ *satisfy* $0 < \inf_{k\in\mathbb{N}} \alpha_{k,\min} \leq \sup_{k\in\mathbb{N}} \alpha_{k,\max} < \frac{1}{2L}$.

The following two sets of assumptions ensure that the variance of the error $F(\xi, x) - T(x)$ is controlled, so that (together with Assumption 16 on the sampling rate) boundedness is guaranteed, even in the case of an unbounded operator.

Assumption 19 (Variance control)**.** *There exists* $p \geq 2$*, such that one of the following three conditions holds:*

i) There exist $x^* \in X^*$ *and* $\sigma(x^*) > 0$ *such that for all* $x \in X$*,*

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

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

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

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

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

$$\|F(\xi, x) - F(\xi, y)\| \leq L(\xi)\|x - y\|,$$

for some measurable $L : \Xi \to \mathbb{R}_+$ with finite L_p-norm for some $p \geq 2$, then Assumptions 13-14 and 19 hold with

$$\sigma(x^*) := \max\{|||F(\xi, x^*) - T(x^*)|||_p, 2|L(\xi)|_p\}$$

for $x^* \in X^*$, in view of Minkowski's inequality. Thus, Assumption 19(i)-(ii) is merely a *finite* variance assumption even for the case of an unbounded feasible set. Assumption 19(iii) means that the variance is *uniformly bounded* over the feasible set X. It has been assumed in most of the past literature [42, 50, 75, 20, 76, 44, 45, 77] on stochastic approximation algorithms for SVI. [6] Assumptions 19(i)-(ii) are much weaker than Assumption 19(iii) and, to the best of our knowledge, seem to be new for monotone operators without regularization.

The next two examples provide instances where Assumption 19(i)-(ii) and the iterative variance reduction in method (4.3)-(4.4) are *necessary* for asymptotic convergence, in the case of an unbounded feasible set (e.g., stochastic equations and stochastic complementarity problems).

Example 1 (Equation problem for zero mean random constant operator)**.** The following example shows that, in the case of an unbounded feasible set, the variation of the mirror-prox method in [20] diverges asymptotically in terms of solutions in the sense that a.s. the generated sequence is unbounded (even though it converges in terms of the gap function). Precisely, the method in [20] say that given a prescribed number of iterations K, for $k \in [K]$ compute:

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

with a final output $\bar{z}^K = \sum_{k=1}^K p_k^K z^k$, where $\{p_k^K\}$ is a positive sequence such that $\sum_{k=1} p_k^K = 1$. For an unbounded X, assuming uniformly bounded variance (Assumption 19(iii)) and a single oracle call per iteration, it is shown that there

[6] Assumption 19(iii) is weakened in previous works only in situations in which the operator satisfies more demanding monotonicity conditions (strongly monotone operator in [73] and weak-sharp monotone operator in method of Section 3.1 of Chapter 3) or when the operator is merely monotone, but with additional Tykhonov regularization (as in method of Section 3.2 of Chapter 3, without convergence rate results).

exists $\{v^K\} \subset \mathbb{R}^n$ such that $\mathbb{E}[\tilde{G}(\bar{z}^K, v^K)] \lesssim 1/\sqrt{K}$ and $\mathbb{E}[\|v^K\|] \lesssim \sqrt{K}$ (see [20], Corollary 3.4). In these statements, for $z, v \in \mathbb{R}^n$,

$$\tilde{G}(z, v) = \sup_{y \in X} \langle T(y) - v, z - y \rangle, \tag{4.10}$$

is the relaxed dual gap function recently introduced by Monteiro and Svaiter [55, 56], based on the enlargement of monotone operators introduced in [12]. The following example shows, however, that $\limsup_{K\to\infty} \|\bar{z}^K\| = \infty$ with total probability.

We shall consider $n = 1$, but one can easily generalize the argument for any $n > 1$. Consider $X = \mathbb{R}$ and the random operator given by

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

for all $x \in \mathbb{R}$, where ξ is a random variable with zero mean, finite variance σ^2 and finite third moment (one could generalize the argument assuming finite q-moment for any $q > 2$). In this case, trivially $T \equiv 0$, $X^* = \mathbb{R}$ and Assumption 19(iii) holds. It is easy to check that the mirror-prox method in [20] gives, after K iterations, for $k \in [K]$:

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

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

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

(see [20], Corollary 3.4). Using the expression of $\{p_k^K\}$ and $\sum_{k=1}^K p_k^K = 1$, we get

$$\bar{z}^K = x^1 - \sum_{k=1}^{K} \theta_k^K \cdot \xi_k^K, \tag{4.11}$$

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

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

from which the following estimates follow:

$$s_K^2 := \sum_{k=1}^{K} \left(\theta_k^K\right)^2 \sim 1, \quad \sum_{k=1}^{K} \left(\theta_k^K\right)^3 \sim K^{-5} \quad (\text{as } K \to \infty).$$

We invoke Lyapounov's criteria ([10], Theorem 7.3) with $\delta = 1$ for the sum $\sum_{k=1}^{K} \theta_k^K \cdot \xi_k^K$ of independent random variables, obtaining

$$\lim_{K\to\infty} \frac{\mathbb{E}\left[|\xi|^3\right]}{s_K^3} \sum_{k=1}^{K} \left(\theta_k^K\right)^3 = \lim_{K\to\infty} \mathbb{E}\left[|\xi|^3\right] K^{-5} = 0.$$

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

$$\begin{aligned} \mathbb{P}\left(\limsup_{K\to\infty} \bar{z}^K > R\right) &= \mathbb{P}\left(\limsup_{K\to\infty} \sum_{k=1}^{K} \frac{\theta_k^K}{\sigma s_K} \cdot \xi_k^K > CR\right) \\ &\geq \limsup_{K\to\infty} \mathbb{P}\left(\sum_{k=1}^{K} \frac{\theta_k^K}{\sigma s_K} \cdot \xi_k^K > CR\right) > 0, \end{aligned} \tag{4.12}$$

using (4.11) and $s_K \sim 1$ in the equality of (4.12). For every $R > 0$, the event $A_R := [\limsup_{K\to\infty} \bar{z}^K > R]$ is a tail event with positive probability and, hence, has total probability from Kolmogorov's zero-one law ([26], Theorem 2.5.1). We conclude then that

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

as claimed.

Example 2 (Linear SVI with unbounded feasible set)**.** The following relevant example [7] is a typical situation of a non-uniform variance over a unbounded feasible set. Let the random operator be:

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

for all $x \in \mathbb{R}^n$, where $A(\xi)$ is an random matrix whose entries have finite mean and variance, such that $\bar{A} := \mathbb{E}[A(\xi)]$ is nonnull and positive semidefinite. In this case, $T(x) = \bar{A}x$ $(x \in \mathbb{R}^n)$ is monotone and linear. For all $x \in \mathbb{R}^n$, $\mathbb{V}[F(\xi, x)] = x^t B x$, where $B := \sum_{i=1}^{m} \operatorname{cov}[A_i(\xi)]$ is positive semidefinite and $A_1(\xi), \ldots A_m(\xi)$ are the rows of $A(\xi)$. Hence, for all $x \in \mathbb{R}^n - \{0\}$ we have

$$\frac{\mathbb{V}[F(\xi, x)]}{\|x\|^2} \geq \lambda_+(\sqrt{B}) > 0,$$

[7] It includes, for instance, quadratic programming, stochastic linear equations and complementarity problems, affine convex-concave saddle-point problems and bimatrix games.

where $\lambda_+(\sqrt{B})$ is the smallest nonnull eigenvalue of $\sqrt{B}$. This shows that Assumption 19(iii) does not hold if X is unbounded (in fact, the variance grows quadratically in the infinite horizon, that is, Assumption 19(i)-(ii) hold with equality). Hence, in the case in which X is unbounded, this example cannot be studied under the uniform variance Assumption 19(iii). Note that if X is a compact set and e.g. $X \supset \{0\}$, then $0 \in X^*$ and $\mathbb{V}[\epsilon(\xi, 0)] = 0$, so that σ^2 in Assumption 19(iii) can be a very conservative upper bound on the oracle variance over X. This situation might suggest to invoke Assumption 19(i)-(ii) even in the compact case so that in the convergence analysis of the method, only the variance *at points of the trajectory* and *the solution set* matter.

4.2 Convergence analysis

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

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

We start with two key lemmas whose proofs are given in the Appendix. First, we define recursively, for $k \in \mathbb{N}_0$, $A_0 := 0$,

$$A_{k+1} := A_k + (8 + \rho_k)\alpha_{k,\max}^2 \|\epsilon_1^k\|^2 + 8\alpha_{k,\max}^2 \|\epsilon_2^k\|^2, \tag{4.13}$$

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

$$M_{k+1}(x^*) := M_k(x^*) + 2\langle x^* - z^k, D(\alpha_k) \cdot \epsilon_2^k \rangle. \tag{4.14}$$

Lemma 9 (Recursive relation)**.** *Suppose that Assumption 12, 14 and 18 hold, and let* $\rho_k := 1 - 4L^2\alpha_{k,\max}^2 > 0$*. Then, almost surely, for all* $k \in \mathbb{N}$ *and* $x^* \in X^*$*,*

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

From definitions (4.13)-(4.14), $\{A_k\}$ is a non-decreasing process for which $A_k \in \mathcal{F}_k$ and, for any $x^* \in X^*$, $\{M_k(x^*), \widehat{\mathcal{F}}_k\}$ is a martingale (since $z^k \in \widehat{\mathcal{F}}_k$, $\mathbb{E}[\epsilon_2^k | \widehat{\mathcal{F}}_k] = 0$).

Lemma 10 (Error decay)**.** *Consider Assumptions 13-19. For each $i \in [m]$ and $N_i \in \mathbb{N}$, let $\xi_i := \{\xi_{j,i} : j \in [N_i]\}$ be an i.i.d. sample of ξ with $1/\mathcal{N} := \frac{1}{n}\sum_{i=1}^m n_i/N_i$ and $N_{\min} := \min_{i\in[m]} N_i$. For any $x \in X$ set*

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

If Assumption 19(i) hold for some $x^* \in X^*$, then for all $x \in X, v \in \mathbb{R}^n$,

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

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

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

Moreover, if Assumption 19(iii) holds, then for all $x \in X, v \in \mathbb{R}^n$,

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

The following two results will establish upper bounds on $A_{k+1} - A_k$ and $M_{k+1}(x^*) - M_k(x^*)$ in terms of $\|x^k - x^*\|^2$ for any $x^* \in X^*$. Under the Assumptions 19(i)-(ii) of non-uniform variance, we need first a bound of $\|x^* - z^k\|^2$ in terms of $\|x^k - x^*\|^2$.

Proposition 3. *Consider Assumptions 12-19. If Assumption 19(i) holds for some $x^* \in X^*$, then*

$$\left| \|z^k - x^*\| \middle| \mathcal{F}_k \right|_p \le (1 + L\alpha_{k,\max} + \mathsf{H}_k(x^*)) \|x^k - x^*\| + \mathsf{H}_k(x^*),$$

where $\mathsf{H}_k(x^*) := \mathsf{G}_k(x^*)\sqrt{\frac{\mathsf{A}}{\mathcal{N}_k}}$ and $\mathsf{G}_k(x^*) := \alpha_{k,\max} C_p \sigma(x^*)$.

Moreover, if Assumption 19(iii) holds, then

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

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

Proof. Recall that $z^k = \Pi[x^k - D(\alpha_k)\,(T(x^k) + \epsilon_1^k)]$. By Lemma 1(iv) and Lemma 3, we have $x^* = \Pi[x^* - D(\alpha_k)\,T(x^*)]$.

Consider first Assumption 19(i). By Lemma 1(iii),

$$\begin{aligned} \|x^* - z^k\| &\leq \|x^* - x^k - D(\alpha_k)(T(x^*) - T(x^k)) + D(\alpha_k)\epsilon_1^k\| \\ &\leq \|x^* - x^k\| + \|D(\alpha_k)\|\|T(x^k) - T(x^*)\| + \|D(\alpha_k)\|\,\left\|\epsilon_1^k\right\| \\ &\leq (1 + L\alpha_{k,\max})\,\|x^* - x^k\| + \alpha_{k,\max}\,\left\|\epsilon_1^k\right\|, \end{aligned} \tag{4.15}$$

using the Lipschitz continuity of T and $\|D(\alpha_k)\| = \alpha_{k,\max}$ in last inequality. We now recall the definition of ϵ_1^k in (4.5). We have

$$\left\|\left\|\epsilon_1^k\right\|\left|\mathcal{F}_k\right.\right|_p \leq \left(\frac{\mathsf{A}}{\mathcal{N}_k}\right)^{\frac{1}{2}} C_p\sigma(x^*)\,(1 + \|x^k - x^*\|), \tag{4.16}$$

using Lemma 10, $x^k \in \mathcal{F}_k$ and the independence of ξ^k with $\mathcal{F}_k$. Invoking Minkowski's inequality, we take $|\cdot|\mathcal{F}_k|_p$ in (4.15) and use (4.16) together with $x^k \in \mathcal{F}_k$ in order to finish the proof.

We now consider Assumption 19(iii). In this case, (4.15) may be replaced by

$$\|x^* - z^k\| \leq (1 + \mathcal{L}\alpha_{k,\max})\,\|x^* - x^k\| + \alpha_{k,\max}\left(\mathcal{M} + \left\|\epsilon_1^k\right\|\right), \tag{4.17}$$

with $\mathcal{L} = L$ and $\mathcal{M} = 0$ as stated in the proposition. By Lemma 10, relation (4.16) is replaced by $\left\|\left\|\epsilon_1^k\right\|\left|\mathcal{F}_k\right.\right|_p \leq \frac{C_p\sigma}{\sqrt{N_{k,\min}}}$, which together with (4.17) implies the required statement, after taking $|\cdot|\mathcal{F}_k|_p$ in (4.17). □

The following proposition gives bounds on the increments of $\{A_k\}$ and $\{M_k(x^*)\}$ in terms of $\|x^k - x^*\|^2$, using the definitions given in Proposition 3.

Proposition 4 (Bounds on increments)**.** *Consider Assumptions 12-19. If Assumption 19(i) holds for some $x^* \in X^*$, then, for all $k \in \mathbb{N}_0$,*

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

$$|M_{k+1}(x^*) - M_k(x^*)|\mathcal{F}_k|_q \leq \sqrt{\frac{\mathsf{B}}{\mathsf{A}}}\mathsf{H}_k(x^*)\,[1 + L\alpha_{k,\max} + \mathsf{H}_k(x^*)]^2\,\|x^k - x^*\|^2$$

$$+\sqrt{\frac{\mathsf{B}}{\mathsf{A}}}\mathsf{H}_k(x^*)\left[1 + L\alpha_{k,\max} + (3 + 2L\alpha_{k,\max})\mathsf{H}_k(x^*) + 2H_k(x^*)^2\right]\|x^k - x^*\|$$

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

Moreover, if Assumption 19(iii) holds, then, for all $k \in \mathbb{N}_0$,

$$\begin{aligned}
|A_{k+1}-A_k|\mathcal{F}_k|_q &\leq (16+\rho_k)\,\alpha_{k,\max}^2\frac{C_p^2\sigma^2}{N_{k,\min}},\\
|M_{k+1}(x^*)-M_k(x^*)|\mathcal{F}_k|_q &\leq (1+\mathcal{L}\alpha_{k,\max})\,\alpha_{k,\max}\frac{C_p\sigma}{\sqrt{N_{k,\min}}}\|x^k-x^*\|\\
&+ \left(\mathcal{M}+\frac{C_p\sigma}{\sqrt{N_{k,\min}}}\right)\alpha_{k,\max}^2\frac{C_p\sigma}{\sqrt{N_{k,\min}}}.
\end{aligned}$$

Proof. Assume first that Assumption 19(i) holds. We start with the bound on $A_{k+1}-A_k$. Definition (4.5), Lemma 10, $x^k \in \mathcal{F}_k$, the independence of $\{\xi^k\}$ and $\mathcal{F}_k$ and relation $(a+b)^2 \leq 2a^2+2b^2$ imply

$$\left\|\left\|\epsilon_1^k\right\|^2\left|\mathcal{F}_k\right|_q = \left\|\left\|\epsilon_1^k\right\|\left|\mathcal{F}_k\right|_p^2 \leq 2\frac{\mathsf{A}}{\mathcal{N}_k}C_p^2\sigma(x^*)^2\,(1+\|x^k-x^*\|^2). \tag{4.18}$$

We proceed similarly for a bound of ϵ_2^k defined in (4.6), but with the use of the filtration $\widehat{\mathcal{F}}_k$. Lemma 10, $z^k \in \widehat{\mathcal{F}}_k$ and the independence of $\{\eta^k\}$ and $\widehat{\mathcal{F}}_k$ imply

$$\left\|\left\|\epsilon_2^k\right\|\left|\widehat{\mathcal{F}}_k\right|_p \leq \left(\frac{\mathsf{A}}{\mathcal{N}_k}\right)^{\frac{1}{2}}C_p\sigma(x^*)\,(1+\|z^k-x^*\|). \tag{4.19}$$

We condition (4.19) with $\left|\left|\cdot\left|\widehat{\mathcal{F}}_k\right|_p\right|\mathcal{F}_k\right|_p = \left|\cdot\left|\mathcal{F}_k\right|_p\right|$, and then take squares, getting

$$\left\|\left\|\epsilon_2^k\right\|^2\left|\mathcal{F}_k\right|_q = \left\|\left\|\epsilon_2^k\right\|\left|\mathcal{F}_k\right|_p^2 \leq 2\frac{\mathsf{A}}{\mathcal{N}_k}C_p^2\sigma(x^*)^2\left(1+\left|\|z^k-x^*\|\right|\mathcal{F}_k\right|_p^2\right). \tag{4.20}$$

Finally we use (4.18), (4.20), (4.13), Proposition 3 and the relation $(a+b)^2 \leq 2a^2+2b^2$, obtaining the required bounds on $A_{k+1}-A_k$.

Now we deal with $M_{k+1}(x^*)-M_k(x^*)$. Definition (4.6), Lemma 10, $z^k \in \widehat{\mathcal{F}}_k$ and the independence of $\{\eta^k\}$ and $\widehat{\mathcal{F}}_k$ imply

$$\left|\langle x^*-z^k, D(\alpha_k)\epsilon_2^k\rangle\right|\widehat{\mathcal{F}}_k\Big|_p \leq \|z^k-x^*\|\|D(\alpha_k)\|\sqrt{\frac{\mathsf{B}}{\mathcal{N}_k}}C_p\sigma(x^*)(1+\|z^k-x^*\|). \tag{4.21}$$

In (4.21), we first use $\left|\cdot|\widehat{\mathcal{F}}_k\right|_q \leq \left|\cdot|\widehat{\mathcal{F}}_k\right|_p$ and then take $\left|\cdot\left|\mathcal{F}_k\right|_q$, obtaining

$$\left|\langle x^*-z^k, D(\alpha_k)\epsilon_2^k\rangle\right|\mathcal{F}_k\Big|_q \leq$$

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

$$\alpha_{k,\max}\sqrt{\frac{\mathsf{B}}{\mathcal{N}_k}}C_p\sigma(x^*)\left(\left|\|x^*-z^k\|\right|\mathcal{F}_k\Big|_p+\left|\|x^*-z^k\|\right|\mathcal{F}_k\Big|_p^2\right), \tag{4.22}$$

using the facts that $\left|\left|\cdot\left|\widehat{\mathcal{F}}_k\right|_q\right|\mathcal{F}_k\right|_q=\left|\cdot\left|\mathcal{F}_k\right|_q$ and $\alpha_{k,\max}=\|D(\alpha_k)\|$ in the first inequality and the fact that $|\cdot|\mathcal{F}_k|_q\leq|\cdot|\mathcal{F}_k|_p$ in the second inequality. Definitions (4.14), (4.22), Proposition 3 and the relation $(a+b)^2\leq 2a^2+2b^2$ entail the required bound on $M_{k+1}(x^*)-M_k(x^*)$.

Suppose now that Assumption 19(iii) hold. First we prove the bound on $\{A_{k+1}-A_k\}$. The proof is similar to the previous case, but (4.18) and (4.20) are replaced respectively by

$$\left|\left\|\epsilon_1^k\right\|^2\right|\mathcal{F}_k\Big|_q\leq\frac{C_p^2\sigma^2}{N_{k,\min}},\quad\left|\left\|\epsilon_2^k\right\|^2\right|\mathcal{F}_k\Big|_q\leq\frac{C_p^2\sigma^2}{N_{k,\min}}, \tag{4.23}$$

using Lemma 10. From Definitions (4.13) and (4.23) we obtain the required bound on $A_{k+1}-A_k$. We deal now with $\{M_{k+1}(x^*)-M_k(x^*)\}$. The proof is similar to the previous case, but instead of (4.21) now we have

$$\left|\langle x^*-z^k,D(\alpha_k)\epsilon_2^k\rangle\right|\widehat{\mathcal{F}}_k\Big|_p\leq\|z^k-x^*\|\|D(\alpha_k)\|\frac{C_p\sigma}{\sqrt{N_{k,\min}}}, \tag{4.24}$$

using Lemma 10. In (4.24), we use $\left|\cdot|\widehat{\mathcal{F}}_k\right|_q\leq\left|\cdot|\widehat{\mathcal{F}}_k\right|_p$ and then take $\left|\cdot\left|\mathcal{F}_k\right|_q$, to get

$$\left|\langle x^*-z^k,D(\alpha_k)\epsilon_2^k\rangle\right|\mathcal{F}_k\Big|_q\leq\alpha_{k,\max}\frac{C_p\sigma}{\sqrt{N_{k,\min}}}\left|\|z^k-x^*\|\right|\mathcal{F}_k\Big|_q, \tag{4.25}$$

using the facts that $\left|\left|\cdot\left|\widehat{\mathcal{F}}_k\right|_q\right|\mathcal{F}_k\right|_q=\left|\cdot\left|\mathcal{F}_k\right|_q$ and $\alpha_{k,\max}=\|D(\alpha_k)\|$. Definition (4.14), Proposition 3 with $|\cdot|\mathcal{F}_k|_q\leq|\cdot|\mathcal{F}_k|_p$ and (4.25) imply the required bound on $M_{k+1}(x^*)-M_k(x^*)$. □

Now, we combine Lemma 9 and Proposition 4 in the following recursive relation.

Proposition 5 (Stochastic quasi-Fejér property)**.** *Consider Assumptions 12-19. Then, there exists $x^*\in X^*$ such that*

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

In the above relation, under Assumption 19(i)-(ii), $\mathcal{I}=1$, $\mathcal{N}_k'=\mathcal{N}_k$ and

$$\mathsf{C}_k(x^*) := \mathsf{AG}_k(x^*)^2\left(32\left(1+L\alpha_{k,\max}+\mathsf{H}_k(x^*)\right)^2+18\right),$$

while, under Assumption 19(iii), $\mathcal{I}=0$, $\mathcal{N}_k'=N_{k,\min}$ and

$$\mathsf{C}_k(x^*) := \mathsf{C}_k = (16+\rho_k)\alpha_{k,\max}^2 C_p^2\sigma^2.$$

Proof. We take the conditional expectation with respect to $\widehat{\mathcal{F}}_k$ in relation of Lemma 9 obtaining

$$\mathbb{E}[\|x^{k+1}-x^*\|^2|\widehat{\mathcal{F}}_k] \leq \|x^k-x^*\|^2 - \frac{\rho_k}{2}r_{\alpha_k}(x^k)^2 + \mathbb{E}[A_{k+1}-A_k|\widehat{\mathcal{F}}_k], \tag{4.26}$$

using the facts that $x^k, z^k \in \widehat{\mathcal{F}}_k$ and $\mathbb{E}[M_{k+1}-M_k|\widehat{\mathcal{F}}_k]=0$, because $\{M_k,\widehat{\mathcal{F}}_k\}$ is a martingale. We now take the conditional expectation with respect to $\mathcal{F}_k$ in (4.26) obtaining

$$\mathbb{E}[\|x^{k+1}-x^*\|^2|\mathcal{F}_k] \leq \|x^k-x^*\|^2 - \frac{\rho_k}{2}r_{\alpha_k}(x^k)^2 + \mathbb{E}[A_{k+1}-A_k|\mathcal{F}_k], \tag{4.27}$$

using the fact that $x^k \in \mathcal{F}_k$ and the hereditary property $\mathbb{E}[\mathbb{E}[\cdot|\widehat{\mathcal{F}}_k]|\mathcal{F}_k]=\mathbb{E}[\cdot|\mathcal{F}_k]$.

We have that

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

Hence, under Assumption 19(i)-(ii), the bound of $\{A_{k+1}-A_k\}$ given in Proposition 4 implies that

$$|A_{k+1}-A_k|\mathcal{F}_k|_q \leq \mathsf{C}_k(x^*)\frac{\mathcal{I}\|x^k-x^*\|^2+1}{\mathcal{N}_k'}, \tag{4.28}$$

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

Under Assumption 19(iii), Proposition 4 implies (4.28), with $\mathcal{I}=1$ and $\mathcal{N}_k'=N_{k,\min}$ and definition of C_k. The claimed relation follows from (4.27) and (4.28) for $q=1$. □

Remark 5 (Bounds of $A_{k+1}-A_k$)**.** Under Assumption 19(i)-(ii), the upper-bound on $\mathsf{C}(x^*) := \sup_k \mathsf{C}_k(x^*)$ depends only on p, L, $\hat{\alpha} := \sup_k \alpha_{k,\max}$ and the sampling rate $\mathcal{N}_k$ and $n\sigma(x^*)^2$. From the definition of $\mathsf{C}_k(x^*)$ in Proposition 5 under Assumption 19(i)-(ii), there exists $c>1$ such that

$$\frac{\mathsf{C}_k(x^*)}{\mathcal{N}_k} \leq c\mathsf{H}_k(x^*)^2\left(1+\mathsf{H}_k(x^*)^2\right) \leq c\hat{\alpha}^2C_p^2\frac{\mathsf{A}\sigma(x^*)^2}{\mathcal{N}_k}\left(1+\hat{\alpha}^2C_p^2\frac{\mathsf{A}\sigma(x^*)^2}{\mathcal{N}_k}\right), \tag{4.29}$$

that is, $\mathsf{C}(x^*) \lesssim n^2\sigma(x^*)^4$, since $\mathsf{A} \in \{n, 2n\}$. But since at least $\mathcal{N}_k \geq N_{k,\min} \approx \Theta k^{1+a}(\ln k)^{1+b}$, for some $\Theta > 0$, $a > 0$, $b \geq -1$ or $a = 0$, $b > 0$, the following non-asymptotic bound holds:

$$\mathsf{C}_k(x^*) \lesssim n\sigma(x^*)^2 \left(1 + \frac{n\sigma(x^*)^2}{\Theta k^{1+a}(\ln k)^{1+b}}\right), \tag{4.30}$$

which is $\approx n\sigma(x^*)^2$ for an iteration index k large enough as compared to [8] $n\sigma(x^*)^2$. Under Assumption 19(iii), the following *uniform* bound holds on X^*: $\mathsf{C}_k \lesssim \sigma^2$.

We finish this section with an asymptotic convergence result.

Theorem 9 (Asymptotic convergence for extragradient method with stepsize away from zero)**.** *Under Assumptions 12-19, a.s. the sequence* $\{x^k\}$ *generated by* (4.3)-(4.4) *is bounded,* $\lim_{k\to\infty} \mathrm{d}(x^k, X^*) = 0$, *and* $r_{\alpha_k}(x^k)$ *converges to* 0 *almost surely and in* L^2. *In particular, a.s. every cluster point of* $\{x^k\}$ *belongs to* X^*.

Proof. The result in Proposition 5 may be rewritten as

$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2 | \mathcal{F}_k\right] \leq \left(1 + \frac{\mathcal{I}\mathsf{C}(x^*)}{\mathcal{N}'_k}\right)\|x^k - x^*\|^2 - \frac{\rho_k}{2} r_{\alpha_k}(x^k)^2 + \frac{\mathsf{C}(x^*)}{\mathcal{N}'_k}, \tag{4.31}$$

for all $k \geq 0$ and for some $x^* \in X^*$, as ensured by Assumption 19. Taking into account Assumption 16, i.e., $\sum_k \mathcal{N}_k^{-1} < \infty$, (4.31) and the fact that $x^k \in \mathcal{F}_k$, we apply Theorem 1 with $y_k := \|x^k - x^*\|^2$, $a_k = \mathcal{I}\mathsf{C}(x^*)/\mathcal{N}'_k$, $b_k = \mathsf{C}(x^*)/\mathcal{N}'_k$ and $u_k := \rho_k r_{\alpha_k}(x^k)^2/2$, in order to conclude that a.s. $\{\|x^k - x^*\|^2\}$ converges. In particular, $\{x^k\}$ is bounded, and $\hat{\rho}\sum_k r_{\alpha_k}(x^k)^2 \leq \sum_k \rho_k r_{\alpha_k}(x^k)^2 < \infty$, where $\hat{\rho} := 1 - 4\hat{\alpha}^2 L > 0$ and $\hat{\alpha} := \sup_k \alpha_{k,\max}$ by Assumption 18. Hence, almost surely,

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

The fact that $\lim_{k\to\infty} \mathbb{E}[r_{\alpha_k}(x^k)^2] = 0$ is proved in a similar way, taking the total expectation in (4.31). The boundedness of the stepsize sequence, (4.32), and the continuity of T (Assumption 14), Π (Lemma 1(iii)) and $D(\cdot)$ imply that a.s. every cluster point $\bar{x}$ of $\{x^k\}$ satisfies

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

[8] In terms of i numerical constants, a sharper bound can be obtained by exploiting the first order term $H_k(x^*) \sim n^{1/2}\sigma(x^*)\mathcal{N}_k^{-1/2}$ in the definition of $\mathsf{C}_k(x^*)$. Using this, we get roughly $c \approx 32$, but we do not carry out this procedure here.

for some $\bar{\alpha} \in \mathbb{R}^m_{>0}$, in view of Assumption 18: i.e. the fact that the stepsizes are bounded away from zero; from Lemmas 1(iv) and 3 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. □

4.3 Convergence rate and complexity analysis

We now study the convergence rate and the oracle complexity of our algorithm. Besides the relation in Proposition 5 for $p = 2$, we can also obtain a recursive relation for higher order moments, assuming that $p \geq 4$. This recursion, derived as consequence of Propositions 4 and 6(i), will give an explicit upper-bound on the p-norm of the generated sequence (see Proposition 7). The explicit bound on the 2-norm of the sequence will be used for giving explicit estimates on the convergence rate and complexity under Assumption 19(i)-(ii), i.e., when X and T are unbounded, in Theorem 10. In this setting, we will also obtain sharper estimates of the constants assuming uniform variance over the *solution set* (see Propositions 6(ii), Proposition 7(ii) and Theorem 11). Important cases satisfying these assumptions include the cases in which X^* is a singleton or a compact set (which can occur even when the feasible set X is unbounded)[9]. Under the stronger Assumption 19(iii), that is, uniform variance over the feasible set, even sharper bounds on the estimates will be presented (see Propositions 6(iii) and 7(iii) and Theorem 11).

The definitions in Proposition 5 and Remark 5 will be used in the next proposition.

Proposition 6 (Improved stochastic quasi-Fejér properties)**.**

i) If Assumption 19(i) holds for $p \geq 4$ and some $x^* \in X^*$, then for all k_0, k

[9]This occurs when the solution set is a singleton in the case of a strictly or strongly pseudo-monotone operator. See Theorems 2.3.5 and 2.3.16 in [27] for general conditions ensuring compactness of the solution set of a pseudo-monotone VI. An example is the so called *strictly feasible* complementarity problem over a cone.

such that $0 \leq k_0 < k$*, it holds that*

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

ii) If Assumption 19(ii) holds for $p \geq 2$*, then* $\mathsf{C} := \sup_k \mathsf{C}_k : X^* \to \mathbb{R}_+$ *is a locally bounded and measurable function, and for all* $k \geq 0$*,*

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

iii) If Assumption 19(iii) holds then for all $k \geq 0$*,*

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

Proof. i) Define for simplicity $d_k := \|x^k - x^*\|^2$. Sum relation in Lemma 9 from k_0 to $k-1$ obtaining $0 \leq d_k \leq d_{k_0} + M_k(x^*) - M_{k_0}(x^*) + A_k - A_{k_0}$, which implies

$$0 \leq d_k \leq d_{k_0} + [M_k(x^*) - M_{k_0}(x^*) + A_k - A_{k_0}]_+, \tag{4.32}$$

using that $a \leq b \Rightarrow [a]_+ \leq [b]_+$ for any $a, b \in \mathbb{R}$. We take the q-norm in (4.32), getting

$$\begin{aligned} |d_k|_q &\leq |d_{k_0}|_q + |[M_k(x^*) - M_{k_0}(x^*) + A_k - A_{k_0}]_+|_q \\ &\leq |d_{k_0}|_q + |M_k(x^*) - M_{k_0}(x^*) + A_k - A_{k_0}|_q \\ &\leq |d_{k_0}|_q + |M_k(x^*) - M_{k_0}(x^*)|_q + \sum_{i=k_0+1}^{k} |A_i - A_{i-1}|_q, \end{aligned} \tag{4.33}$$

using Minkowski's inequality in the first and last inequalities and the fact that $|U_+|_q \leq |U|_q$ for any random variable U in the second inequality.

Since $q \geq 2$ ($p \geq 4$), the norm of the martingale term above may be estimated via the BDG inequality (3) applied to the martingale $\tilde{M}_i := M_{k_0+i}(x^*) - M_{k_0}(x^*)$. This gives:

$$|M_k(x^*) - M_{k_0}(x^*)|_q \leq C_q \sqrt{\sum_{i=k_0+1}^{k} |M_i(x^*) - M_{i-1}(x^*)|_q^2}. \tag{4.34}$$

Plugging (4.34) into (4.33) completes the proof of item (i).

ii) Under Assumption 19(ii), we define $\bar{x}^k := \Pi_{X^*}(x^k)$, recalling Assumption 12, and obtain from Proposition 5:

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

using the fact that $\bar{x}^k \in X^*$ in the first inequality, the facts that $\mathsf{C}_k(\bar{x}^k) \in \mathcal{F}_k$ (which holds because $x^k \in \mathcal{F}_k$, Π_{X^*} is continuous and C_k is measurable) and $\bar{x}^k \in X^*$ (cf. Proposition 5) in the second inequality, and the fact that $\mathrm{d}(x^k,X^*) = \|x^k-\bar{x}^k\|$ in the equality. Note that the function $\mathsf{C}: X^* \to \mathbb{R}_+$ is measurable and locally bounded by Assumption 19(ii) and definition of $\mathsf{C}_k(x^*)$.

iii) we use a proof line analogous to the one in item (ii), with Assumption 19(iii) and Proposition 5. □

The following result gives explicit bounds on the p-norm of the sequence in the unbounded setting. In order to make the presentation easier, we introduce some definitions. Recall the definitions of Lemma 10, Propositions 3 and 5 and Remark 5. Set $\mathsf{D}(x^*) := 2c\hat{\alpha}^2C_p^2n\sigma(x^*)^2$, $\tilde{\mathsf{G}}_p(x^*) := C_p\hat{\alpha}\sigma(x^*)$, $\mathsf{B}_2(x^*) := 0$ and for $p \geq 4$,

$$\mathsf{B}_p(x^*) := \sqrt{3\mathsf{B}}C_q\tilde{\mathsf{G}}_p(x^*)\left[(1+L\hat{\alpha})^2+(3+2L\hat{\alpha})\sqrt{\mathsf{A}}\tilde{\mathsf{G}}_p(x^*)+2\mathsf{A}\tilde{\mathsf{G}}_p(x^*)^2\right]. \tag{4.35}$$

Proposition 7 (Uniform boundedness in L^p)**.**

i) Let Assumptions 12-19(i) hold for some $x^* \in X^*$ and $p \in \{2\} \cup [4,\infty)$. Choose $k_0 := k_0(x^*) \in \mathbb{N}$ and $\gamma := \gamma(x^*) > 0$ such that

$$\beta(x^*) := \mathsf{B}_p(x^*)\sqrt{\gamma}+\mathsf{D}(x^*)\gamma+\mathsf{D}(x^*)^2\gamma^2 < 1, \quad \sum_{k\geq k_0}\frac{1}{\mathcal{N}_k} < \gamma. \tag{4.36}$$

Then

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

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

ii) Let Assumptions 12-19(ii) hold and suppose there exists $\sigma > 0$ such that $\sigma(x^*) \leq \sigma$ for all $x^* \in X^*$. Let $\phi \in (0, \frac{\sqrt{5}-1}{2})$. Choose $k_0 \in \mathbb{N}$ such that $\sum_{k\geq k_0} \frac{1}{\mathcal{N}_k} \leq \frac{\phi}{2c\hat{\alpha}^2 C_p^2 n\sigma^2}$. Then

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

iii) If Assumptions 12-19(iii) hold then

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

Proof. i) Denote $d^k := \|x^k - x^*\|$. We first unify the Fejér-type relations obtained so far under Assumption 19(i)-(ii) as: for all $k > k_0$,

$$\begin{aligned} |d_k|_p^2 \;&\leq\; |d_{k_0}|_p^2 + \mathsf{B}_p(x^*)\sqrt{\sum_{i=k_0}^{k-1} \frac{1 + |d_i|_p^2 + |d_i|_p^4}{\mathcal{N}_i}} + \\ &+\; \mathsf{D}(x^*) \sum_{i=k_0}^{k-1} \frac{1 + |d_i|_p^2}{\mathcal{N}_i} + \mathsf{D}(x^*)^2 \sum_{i=k_0}^{k-1} \frac{1 + |d_i|_p^2}{\mathcal{N}_i^2}. \end{aligned} \tag{4.37}$$

Indeed, for $p = 2$, we have $\mathsf{B}_2(x^*) = 0$ so that (4.37) results by summing the relation in Proposition 5 from k_0 to $k-1$ and using the estimate in (4.29), the facts that $\mathsf{A} \leq 2n$, $c > 1$, and the definition of $\mathsf{D}(x^*)$ as stated before this proposition. For $p \geq 4$, we recall the bounds of increments of $\{M_k(x^*)\}$ in Proposition 4. The common factor is bounded by $\sqrt{\mathsf{B}/\mathsf{A}} \cdot \mathsf{H}(x^*) \leq \sqrt{\mathsf{B}}\tilde{\mathsf{G}}_p(x^*)/\sqrt{\mathcal{N}_k}$. Using the definitions of $\mathsf{H}(x^*)$, $\hat{\alpha}$ and $\tilde{\mathsf{G}}_p(x^*)$, and the fact that $\mathcal{N}_k \geq 1$, it is easy to see that, in the bound of $M_{k+1}(x^*) - M_k(x^*)$ in Proposition 4, the sum of terms multiplying $\sqrt{\mathsf{B}/\mathsf{A}} \cdot \mathsf{H}(x^*)$ is less than or equal to $(1 + L\hat{\alpha})^2 + (3 + 2L\hat{\alpha})\sqrt{\mathsf{A}}\tilde{\mathsf{G}}_p + 2\mathsf{A}\tilde{\mathsf{G}}_p^2$. We use these bounds, the fact that $(|d_i|_p^2 + |d_i|_p + 1)^2 \leq 3(|d_i|_p^4 + |d_i|_p^2 + 1)$ and Definition (4.35) in order to obtain, for all $i \in \mathbb{N}_0$,

$$|M_{i+1}(x^*) - M_i(x^*)|_q^2 \leq \mathsf{B}_p(x^*)^2 \frac{1 + |d_i|_p^2 + |d_i|_p^4}{\mathcal{N}_i}. \tag{4.38}$$

The proof of (4.37) for $p \geq 4$ follows from (4.28), (4.29) with $\mathsf{A} \leq 2n$, $c > 1$ and the definition of $\mathsf{D}(x^*)$, as well as (4.38) and Proposition 6(i).

By Assumption 16, we can choose $k_0 \in \mathbb{N}_0$ and $\gamma > 0$ as in (4.36). In particular, $\sum_{i\geq k_0} \mathcal{N}_i^{-2} < \gamma^2$. Given an arbitrary $a > |d_{k_0}|_p$, define: $\tau_a := \inf\{k > k_0 : |d_k|_p \geq$

$a\}$. Suppose first that $\tau_a < \infty$ for all $a > |d_{k_0}|_p$. By (4.36), (4.37) and the definition of τ_a, we have

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

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

$$\leq |d_{k_0}|_p^2 + \mathsf{B}_p(x^*)\sqrt{\gamma}\left(1+a+a^2\right) + \mathsf{D}(x^*)\gamma\left(1+a^2\right) + \mathsf{D}(x^*)^2\gamma^2\left(1+a^2\right). \tag{4.39}$$

For $p = 2$, $\mathsf{B}_2(x^*) = 0$. Relation (4.39) and $\beta := \beta(x^*) \in (0,1)$ in (4.36) imply

$$a^2 \leq \frac{|d_{k_0}|_p^2 + 1}{1-\beta}. \tag{4.40}$$

For $p \geq 4$, (4.39) and $\beta := \beta(x^*)$ in (4.36) imply $\lambda a^2 \leq |d_{k_0}|_p^2 + a + 1$, with $\lambda := 1 - \beta$. This gives

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

and finally

$$a^2 \leq 4\frac{|d_{k_0}|_p^2 + 1}{(1-\beta)^2}. \tag{4.41}$$

Since (4.40)-(4.41) hold for an arbitrary $a > |d_{k_0}|_p$ and $\beta \in (0,1)$, it follows that $\sup_{k\geq k_0} |d_k|_p^2 \leq \mathsf{c}_p(x^*)\left[1 + |d_{k_0}|_p^2\right]$, with $\mathsf{c}_p(x^*)$ as in the statement of this proposition. This contradicts the initial assumption that $\tau_a < \infty$ for all $a > |d_{k_0}|_p$. Hence there exists $\bar{a} > |d_{k_0}|_p$ such that $\hat{a} := \sup_{k\geq k_0} |d_k|_p \leq \bar{a} < \infty$ by the definition of $\tau_{\bar{a}}$. For any $k > k_0$, we use that $|d_i|_p \leq \hat{a}$ for $k_0 \leq i < k$ in (4.37) obtaining

$$|d_k|_p^2 \leq |d_{k_0}|_p^2 + \mathsf{B}_p(x^*)\sqrt{\gamma}\left(1+\hat{a}+\hat{a}^2\right) + \mathsf{D}(x^*)\gamma\left(1+\hat{a}^2\right) + \mathsf{D}(x^*)^2\gamma^2\left(1+\hat{a}^2\right). \tag{4.42}$$

Note that (4.42) holds trivially for $k := k_0$. Thus, after taking the supremum over $k \geq k_0$ in (4.42), we proceed as done immediately after inequality (4.39), obtaining (4.40) and (4.41), respectively for $p = 2$ and $p \geq 4$, but with $\hat{a}$ substituting for a, which proves the claim, in view of the definition of $\mathsf{c}_p(x^*)$.

ii): The proof line is the same as for the case $p = 2$ in item (i), but summing the relation in Proposition 5 with the estimate (4.29), which gives the following uniform estimate: for all $k \geq 0$, $C_k\left(\Pi\left(\bar{x}^k\right)\right)\mathcal{N}_k^{-1} \leq 2c\hat{\alpha}^2 C_p^2 n\sigma^2 \mathcal{N}_k^{-1}\left(1 + 2\hat{\alpha}^2 C_p^2 n\sigma^2 \mathcal{N}_k^{-1}\right)$. We remark that we may replace $\beta(x^*)$ in (4.36) and (4.40) by $\beta := 2c\hat{\alpha}^2 C_p^2 n\sigma^2 + 4c\hat{\alpha}^4 C_p^4 n^2\sigma^4$. In this case, the definition of ϕ and k_0 imply $0 < 1 - \phi - \phi^2 \leq 1 - \beta$.

iii): Given $k \in \mathbb{N}$, we take total expectation in the relation of Proposition 6(iii) and sum from 0 to k obtaining

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

and the claim follows. □

Remark 6. In the statement of Proposition 7(i), for $p \geq 2$, it is sufficient to set $\phi \in (0, \frac{\sqrt{5}-1}{2})$ and $k_0 \in \mathbb{N}_0$ such that $\sum_{k\geq k_0} \mathcal{N}_k^{-1} \leq \phi \mathsf{D}(x^*)^{-1}$ in order to obtain that $\sup_{k\geq k_0} \mathbb{E}[\|x^k - x^*\|^2] \leq \frac{1+\mathbb{E}[\|x^{k_0}-x^*\|^2]}{1-\phi-\phi^2}$.

We now give explicit estimates on the convergence rate and oracle complexity. In the sequel we assume that the stepsize sequence is constant. Proposition 10.3.6 in [27] states that $\{r_a : a > 0\}$ is a family of equivalent merit functions of $\mathrm{VI}(T, X)$. Hence, the convergence rate analysis can be established for varying stepsizes satisfying Assumption 18 and constant stepsizes are assumed just for simplicity. We need now the following definitions: for $\ell \leq k$, $\mathsf{a}_0^k := \sum_{i=0}^k \frac{1}{\mathcal{N}_i}$, and $\mathsf{b}_0^k := \sum_{i=0}^k \frac{1}{\mathcal{N}_i^2}$. In the remainder of this section, we need the definitions of the constants used in Lemma 10, Propositions 3, 5 and 7 and Remark 5.

Theorem 10 (Convergence rate: non-uniform variance)**.** *Consider Assumptions 12-19(i) for some $x^* \in X^*$. Take $\alpha_k \equiv \alpha \in (0, 1/2L)^m$, $\phi \in (0, \frac{\sqrt{5}-1}{2})$ and $k_0 \in \mathbb{N}$ such that:*

$$\sum_{k\geq k_0} \frac{1}{\mathcal{N}_k} \leq \frac{\phi}{\mathsf{D}(x^*)}. \tag{4.43}$$

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 $\mathbb{E}[r_\alpha(x^{K_\epsilon})^2] \leq \epsilon \leq \frac{\mathsf{Q}_\infty(x^*)}{K_\epsilon}$, where for all $k \in \mathbb{N}_0 \cup \{\infty\}$,

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

Additionally, if Assumption 19(ii) holds then K_ϵ is independent of $x^* \in X^*$.

Proof. First note that finiteness of $\mathsf{a}_0^\infty, \mathsf{b}_0^\infty$ as defined in the statement of this theorem follows from Assumption 16, which also ensures existence of k_0 satisfying (4.43), because $\sum_{i\geq k} \mathcal{N}_i^{-1} \to 0$ as $k \to \infty$. We now invoke Proposition 5. Let $k \geq 0$. Given $0 \leq i \leq k$, we take total expectation in the relation of Proposition 5, with the estimate (4.29), using the facts that $\mathsf{A} \leq 2n$, $c > 1$ and the definition of $\mathsf{D}(x^*)$. We then sum with i running from 0 to k, obtaining:

$$\begin{aligned}
&\frac{\rho}{2}\sum_{i=0}^{k}\mathbb{E}[r_\alpha(x^i)^2] \\
&\leq \|x^0 - x^*\|^2 + \mathsf{D}(x^*)\sum_{i=0}^{k}\frac{1+\mathbb{E}\left[\|x^i - x^*\|^2\right]}{\mathcal{N}_i} + \mathsf{D}(x^*)^2\sum_{i=0}^{k}\frac{1+\mathbb{E}\left[\|x^i - x^*\|^2\right]}{\mathcal{N}_i^2} \\
&\leq \|x^0 - x^*\|^2 + \left(1 + \sup_{0\leq i\leq k}\mathbb{E}[\|x^i - x^*\|^2]\right)\left(\mathsf{D}(x^*)\sum_{i=0}^{k}\frac{1}{\mathcal{N}_i} + \mathsf{D}(x^*)^2\sum_{i=0}^{k}\frac{1}{\mathcal{N}_i^2}\right) \\
&\leq \|x^0 - x^*\|^2 + [1+\mathsf{J}(x^*)]\left[\mathsf{D}(x^*)\mathsf{a}_0^k + \mathsf{D}(x^*)^2\mathsf{b}_0^k\right] = \frac{\rho}{2}\mathsf{Q}_k(x^*).
\end{aligned} \tag{4.44}$$

The last inequality in (4.44) follows from (4.43), Proposition 7(i) for $p = 2$ and Remark 6, which imply

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

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

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

$$\frac{\rho}{2}\epsilon(k+1) \leq \frac{\rho}{2}\sum_{i=0}^{k}\mathbb{E}[r_\alpha(x^i)^2]. \tag{4.45}$$

We claim that K is finite. Indeed, if $K = \infty$, then (4.44) and (4.45) hold for all $k \in \mathbb{N}$. Hence, we arrive at a contradiction by letting $k \to \infty$ and using the facts that $\mathsf{a}_0^\infty < \infty$ and $\mathsf{b}_0^\infty < \infty$, which hold by Assumption 16. Since K is finite, we have that $\mathbb{E}[r_\alpha(x^K)^2] \leq \epsilon$ by definition. Setting $k := K-1$ in (4.44)-(4.45), we get $K \leq \frac{\mathsf{Q}_{K-1}(x^*)}{\epsilon} \leq \frac{\mathsf{Q}_\infty(x^*)}{\epsilon}$, using the definition of $\mathsf{Q}_k(x^*)$. We thus proved the claim. Under Assumption 19(ii), the proof is valid for any $x^* \in X^*$ and, hence, K is independent of $x^* \in X^*$. □

In Theorem 10 given $x^* \in X^*$, the constant $\mathsf{Q}_\infty(x^*) := \mathsf{Q}_\infty(x^*, k_0(x^*), \phi)$ depends on $n\sigma(x^*)^2$ and on the distance of the $k_0(x^*)$ initial iterates to x^*, where $k_0(x^*)$ and ϕ are chosen so that (4.43) is satisfied. Under Assumption 19(ii), since K_ϵ does not depend on $x^* \in X^*$, we get indeed the uniform estimate: $\sup_{\epsilon>0} \epsilon K_\epsilon \leq \inf_{x^* \in X^*} \mathsf{Q}_\infty(x^*, k_0(x^*), \phi)$. In the sense of the previous inequality and in the case of non-uniform variance, the performance of method (4.3)-(4.4) depends on the solution $x^* \in X^*$ such that $\mathsf{Q}_\infty(x^*, k_0(x^*), \phi)$ is minimal.

Proposition 8 (Rate and oracle complexity for $m = 1$: non-uniform variance)**.** *Suppose that the assumptions of Theorem 10 hold. Define $\mathcal{N}_k$ as*

$$\mathcal{N}_k = \left\lceil \theta n \sigma(x^*)^2 (k+\mu)(\ln(k+\mu))^{1+b} \right\rceil \tag{4.46}$$

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

$$k_0 \geq \exp\left[\left(\frac{2c\hat{\alpha}^2 C_p^2}{\phi b \theta}\right)^{1/b}\right] - \mu + 1. \tag{4.47}$$

Define $\lambda := 2c\hat{\alpha}^2 C_p^2$,

$$\mathcal{A} := \frac{\lambda}{b(\ln(\mu-1))^b},$$

$$\mathcal{B} := \frac{\lambda^2}{(\mu-1)(1+2b)[\ln(\mu-1)]^{1+2b}}.$$

Then Theorem 9 holds and for all $\epsilon > 0$, there exists $K := K_\epsilon \in \mathbb{N}$ such that $\mathbb{E}[r_\alpha(x^K)^2] \leq \epsilon$ and

$$\epsilon \leq \frac{2\rho^{-1}\max\{1, \theta^{-2}\}}{K} \cdot \left\{ \|x^0 - x^*\|^2 + (\mathcal{A} + \mathcal{B})\left[1 + \mathsf{J}(x^*)\right] \right\}, \tag{4.48}$$

$$\sum_{k=1}^{K} 2\mathcal{N}_k \leq 12 \max\{1, \theta^{-2}\} \max\{1, \theta n \sigma(x^*)^2\} \frac{\mathsf{P}(x^*)}{\epsilon^2} \mathsf{I}(x^*), \tag{4.49}$$

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

$$\mathsf{I}(x^*) := \rho^{-2}\|x^0 - x^*\|^4 + \rho^{-2}(\mathcal{A} + \mathcal{B})^2 \left[1 + \mathsf{J}(x^*)\right]^2 + 1.$$

Proof. For $\phi \in (0, \frac{\sqrt{5}-1}{2})$, we look for k_0 satisfying (4.43). We have

$$\begin{aligned}\sum_{k\geq k_0}\frac{1}{\mathcal{N}_k} &\leq \theta^{-1}n^{-1}\sigma(x^*)^{-2}\sum_{k\geq k_0}\frac{1}{(k+\mu)(\ln(k+\mu))^{1+b}}\\ &\leq \theta^{-1}n^{-1}\sigma(x^*)^{-2}\int_{k_0-1}^{\infty}\frac{dt}{(t+\mu)(\ln(t+\mu))^{1+b}}\\ &= \frac{\theta^{-1}n^{-1}\sigma(x^*)^{-2}}{b(\ln(k_0-1+\mu))^b}. \qquad (4.50)\end{aligned}$$

From (4.50) and (4.43), it is enough to choose k_0 as the minimum natural number such that the right hand side of (4.50) is less than $\phi/\mathsf{D}(x^*)$. From the definition of $\mathsf{D}(x^*)$, it follows that it is enough to choose k_0 as in (4.47).

We now give an estimate of $\mathsf{Q}_\infty(x^*)$. We have the bound

$$\begin{aligned}(4.51)\quad \mathsf{D}(x^*)\mathsf{a}_0^\infty + \mathsf{D}(x^*)^2\mathsf{b}_0^\infty &\leq \int_{-1}^{\infty}\frac{\lambda\theta^{-1}dt}{(t+\mu)(\ln(t+\mu))^{1+b}} +\\ +\int_{-1}^{\infty}\frac{\lambda^2\theta^{-2}dt}{(t+\mu)^2(\ln(t+\mu))^{2+2b}} &\leq \frac{\lambda\theta^{-1}}{b(\ln(\mu-1))^b}+\frac{\lambda^2\theta^{-2}}{(\mu-1)(1+2b)[\ln(\mu-1)]^{1+2b}}.\end{aligned}$$

From Theorem 10, (4.51) and the definitions of $\mathsf{Q}_\infty(x^*)$, $\mathsf{J}(x^*)$, $\mathcal{A}$ and $\mathcal{B}$ we get (4.48).

We now prove (4.49). Using $K := K_\epsilon \leq \mathsf{Q}_\infty(x^*)/\epsilon$, $\mu\epsilon \leq 1$ and $\mathcal{N}_k \leq \theta n\sigma(x^*)^2(k+\mu)(\ln(k+\mu))^{1+b}+1$, we have

$$\begin{aligned}\sum_{k=1}^{K}2\mathcal{N}_k &\leq \max\{\theta n\sigma(x^*)^2,1\}\sum_{k=1}^{K}2\left[(k+\mu)(\ln(k+\mu))^{1+b}+1\right]\\ &\leq \max\{\theta n\sigma(x^*)^2,1\}K(K+2\mu)\left[(\ln(K+\mu))^{1+b}+\frac{2}{K+2\mu}\right]\\ &\leq \max\{\theta n\sigma(x^*)^2,1\}\frac{\left\{\left[\ln\left(\mathsf{Q}_\infty(x^*)\epsilon^{-1}+\epsilon^{-1}\right)\right]^{1+b}+\mu^{-1}\right\}\mathsf{Q}_\infty(x^*)\left(\mathsf{Q}_\infty(x^*)+2\right)}{\epsilon^2}. \qquad (4.52)\end{aligned}$$

We now use (4.52) with $\mathsf{Q}_\infty(x^*)(\mathsf{Q}_\infty(x^*)+2) \leq (\mathsf{Q}_\infty(x^*)+2)^2$, the definitions of $\mathsf{Q}_\infty(x^*)$, $\mathsf{J}(x^*)$, $\mathcal{A}$, $\mathcal{B}$, equation (4.51) and the relation $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ in order to prove (4.49). □

We give next sharper estimates in the case in which the variance is *uniform* over X^* or X. We state them without proofs since they follow the same proof line of Theorem 10 and Proposition 8, but using Proposition 6(ii) and Proposition

7(ii), in the case in which the variance is uniform over X^*, or using Proposition 6(iii), in the case in which the variance is uniform over X.

Theorem 11 (Convergence rate: uniform variance)**.** *Consider Assumptions 12-19. Take $\alpha_k \equiv \alpha \in (0, 1/2L)^m$. Suppose first that Assumption 19(ii) holds and there exists $\sigma > 0$ such that $\sigma(x^*) \leq \sigma$ for all $x^* \in X^*$. Define $\mathsf{D} := 2c\hat{\alpha}^2 C_p^2 n\sigma^2$ and*

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

Take $\phi \in (0, \frac{\sqrt{5}-1}{2})$ and $k_0 \in \mathbb{N}$ such that $\sum_{k\geq k_0} \mathcal{N}_k^{-1} \leq \phi/\mathsf{D}$. Then, for all $\epsilon > 0$, there exists $K_\epsilon \in \mathbb{N}$, satisfying $\mathbb{E}[r_\alpha\left(x^{K_\epsilon}\right)^2] \leq \epsilon \leq \frac{\mathsf{Q}_\infty}{K_\epsilon}$, where, for all $k \in \mathbb{N}_0 \cup \{\infty\}$,

$$\mathsf{Q}_k := \frac{2}{\rho}\left\{\mathrm{d}(x^0, X^*)^2 + (1 + \mathsf{J})\left(\mathsf{D}\mathsf{a}_0^k + \mathsf{D}^2\mathsf{b}_0^k\right)\right\}.$$

Suppose now that Assumption 19(iii) holds. Then, for all $\epsilon > 0$, there exists $K_\epsilon \in \mathbb{N}$, satisfying $\mathbb{E}[r_\alpha\left(x^{K_\epsilon}\right)^2] \leq \epsilon \leq \frac{\widetilde{\mathsf{Q}}_\infty}{K_\epsilon}$, where, for all $k \in \mathbb{N}_0 \cup \{\infty\}$,

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

Proposition 9 (Rate and oracle complexity for $m = 1$: uniform variance)**.** *Suppose that the assumptions of Theorem 11 hold and that $\sup_{x^*\in X^*} \sigma(x^*) \leq \sigma$ for some $\sigma > 0$. Define $\mathcal{N}_k$ as*

$$\mathcal{N}_k = \left\lceil \theta\sigma^2(k + \mu)(\ln(k + \mu))^{1+b} \right\rceil$$

for any $\theta > 0$, $b > 0$, $\epsilon > 0$ and $2 < \mu \leq \epsilon^{-1}$. Suppose that either:

(i) Assumption 19(ii) holds, in which case we choose $\phi \in (0, \frac{\sqrt{5}-1}{2})$ and $k_0 \in \mathbb{N}$ as in (4.47), or that

(ii) Assumption 19(iii) holds.

Then Theorem 9 holds and for all $\epsilon > 0$, there exists $K_\epsilon \in \mathbb{N}$ such that $\mathbb{E}[r_\alpha(x^{K_\epsilon})^2] \leq \epsilon$ where:

(i) if Assumption 19(ii) holds, then

$$\epsilon \leq \frac{2\rho^{-1}\max\{1, \theta^{-2}\}}{K_\epsilon}\left\{\mathrm{d}(x^0, X^*)^2 + (\mathcal{A} + \mathcal{B})(1 + \mathsf{J})\right\},$$

$$\sum_{k=1}^{K_\epsilon} 2\mathcal{N}_k \leq 12\max\{1, \theta^{-2}\}\max\{1, \theta\sigma^2\}\frac{\left\{\ln\left[(\mathsf{Q}_\infty + 1)\,\epsilon^{-1}\right]\right\}^{1+b} + \mu^{-1}}{\epsilon^2}\mathsf{I},$$

$$\mathsf{I} := \rho^{-2}\,\mathrm{d}(x^0, X^*)^4 + \rho^{-2}(\mathcal{A} + \mathcal{B})^2\,(1 + \mathsf{J})^2 + 1.$$

(ii) if Assumption 19(iii) is satisfied then

$$\epsilon \leq \frac{2\rho^{-1}\max\{1,\theta^{-1}\}}{K_\epsilon}\left\{\mathrm{d}(x^0,X^*)^2+\frac{17C_p^2\hat{\alpha}^2}{b(\ln(\mu-1))^b}\right\},$$

$$\sum_{k=1}^{K_\epsilon}2\mathcal{N}_k \leq 12\max\{1,\theta^{-1}\}\max\{1,\theta\sigma^2\}\frac{\left\{\ln[(\widetilde{\mathsf{Q}}_\infty+1)\epsilon^{-1}]\right\}^{1+b}+\mu^{-1}}{\epsilon^2}\widetilde{\mathsf{I}},$$

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

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

Proposition 10 (Oracle complexity linear in the size of network)**.** *Under Assumptions 12-19(i) with Assumption 17(i) (centralized sampling), the results of Proposition 8 hold.*

Consider Assumption 17(ii) (decentralized sampling). Let $\{b_i\}_{i=1}^m$ be a positive sequence such that

$$N_{k,i} = \left\lceil \theta_i n_i \sigma(x^*)^2 (k+\mu_i)^{1+a}(\ln(k+\mu_i))^{1+b_i}\right\rceil, \tag{4.53}$$

$$b_1 \geq b_i + 2\ln(i+1) - \ln\mathsf{S} \quad \textit{and} \quad \theta_i \sim \theta m, \forall i \in [m], \tag{4.54}$$

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

$$k_0 \geq \left[\frac{2cC_p^2\hat{\alpha}^2 n}{\phi\theta_{\min}b_{\min}n_{\min}}\right]^{1/a} - \mu_{\min}+1. \tag{4.55}$$

Define $\lambda := 2c\hat{\alpha}^2C_p^2$, $\nu := 2+a$, $\mathcal{A}_m := \sum_{i=1}^m \frac{\lambda\theta}{\theta_i a(\mu_i-1)^a}$ and

$$\mathcal{B}_m := \frac{\lambda^2\theta^2}{(1+2b_{\min})(\mu_{\min}-1)^{1+2a}\ln(\mu_{\min}-1)}\left\{\sum_{i=1}^m\frac{1}{\theta_i[\ln(\mu_{\min}-1)]^{b_i}}\right\}^2.$$

Then Theorem 9 holds and for all $\epsilon > 0$, there exists $K := K_\epsilon \in \mathbb{N}$ such that $\mathbb{E}[r_\alpha(x^{K_\epsilon})^2] \le \epsilon$ and

$$\epsilon \le \frac{2\rho^{-1}\max\{1,\theta^{-2}\}}{K} \cdot \left\{\|x^0 - x^*\|^2 + (\mathcal{A}_m + \mathcal{B}_m)\left[1 + \mathsf{J}(x^*)\right]\right\}, \tag{4.56}$$

$$\sum_{k=1}^{K}\sum_{i=1}^{m} 2N_{k,i} \le \frac{4\cdot 3^{1+a}\mathsf{S}\max\left\{1,\theta^{-2\nu}\right\}\max\{\theta_{\max}n_{\max}\sigma(x^*)^2, 1\}\widehat{\mathsf{P}}(x^*)\widehat{\mathsf{I}}(x^*)}{\epsilon^{2+a}},$$

$$\widehat{\mathsf{P}}(x^*) := \left\{\ln\left[(\mathsf{Q}_\infty(x^*)+1)\epsilon^{-1}\right]\right\}^{1+b_1},$$

$$\widehat{\mathsf{I}}(x^*) := (2\rho^{-1})^\nu\|x^0 - x^*\|^{2\nu} + (2\rho^{-1})^\nu(\mathcal{A}_m + \mathcal{B}_m)^\nu\left[1 + \mathsf{J}(x^*)\right]^\nu + 1,$$

where the subscripts "$\min$*" and "*$\max$*" refer, respectively, to the minimal and maximal terms of the corresponding sequences.*

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

$$\int_k^\infty \frac{dt}{(t+\mu)^{1+a}(\ln(t+\mu))^{1+b}} \le \max\left\{\frac{1}{a(k+\mu)^a}, \frac{1}{(k+\mu)^a b[\ln(k+\mu)]^b}\right\}. \tag{4.57}$$

For $\phi \in (0, \frac{\sqrt{5}-1}{2})$ we look for k_0 satisfying (4.43). Since $\mathcal{N}_k$ is an harmonic average of $\{N_{k,i}\}_{i=1}^m$ and $N_{k,i} \ge \theta_{\min}n_{\min}\sigma(x^*)^2(k+\mu_{\min})^{1+a}[\ln(k+\mu_{\min})]^{1+b_{\min}}$ for all $i \in [m]$, we get from (4.57):

$$\sum_{k\ge k_0}\frac{1}{\mathcal{N}_k} \le \theta_{\min}^{-1}n_{\min}^{-1}\sigma(x^*)^{-2}\sum_{k\ge k_0}\frac{1}{(k+\mu_{\min})^{1+a}[\ln(k+\mu_{\min})]^{1+b_{\min}}}$$

$$\le \frac{\theta_{\min}^{-1}n_{\min}^{-1}\sigma(x^*)^{-2}}{(k_0-1+\mu_{\min})^a b_{\min}[\ln(k_0-1+\mu_{\min})^{b_{\min}}]} \le \frac{\theta_{\min}^{-1}n_{\min}^{-1}\sigma(x^*)^{-2}}{(k_0-1+\mu_{\min})^a b_{\min}}, \tag{4.58}$$

if $k_0 \ge e - \mu_{\min} + 1$. In view of (4.58) and (4.43), it is enough to choose k_0 as the minimum natural number greater than $e - \mu_{\min} + 1$ such that the rightmost expression of (4.58) is less than $\phi/\mathsf{D}(x^*)$. Taking into account the definition of $\mathsf{D}(x^*)$, it suffices to choose k_0 as in (4.55).

Next we estimate the value of $\mathsf{Q}_\infty(x^*)$. Recall that

$$\frac{1}{\mathcal{N}_k} = \sum_{i=1}^m \frac{n_i}{n}\cdot\frac{1}{N_{k,i}}.$$

The definitions of $\mathsf{D}(x^*)$, λ, a_0^∞ and b_0^∞ imply

$$\begin{aligned}\mathsf{D}(x^*)\mathsf{a}_0^\infty + \mathsf{D}(x^*)^2\mathsf{b}_0^\infty &\leq \lambda \sum_{k\geq 0}\sum_{i=1}^{m} \frac{\theta_i^{-1}}{(k+\mu_i)^{1+a}(\ln(k+\mu_i))^{1+b_i}} + \\ &\qquad \lambda^2 \sum_{k\geq 0}\left[\sum_{i=1}^{m} \frac{\theta_i^{-1}}{(k+\mu_i)^{1+a}(\ln(k+\mu_i))^{1+b_i}}\right]^2. \end{aligned} \tag{4.59}$$

The first summation in (4.59) is bounded by

$$\lambda \sum_{i=1}^{m}\sum_{k\geq 0} \frac{\theta_i^{-1}}{(k+\mu_i)^{1+a}} \leq \lambda \sum_{i=1}^{m} \int_{-1}^{\infty} \frac{\theta_i^{-1} dt}{(t+\mu_i)^{1+a}} \leq \sum_{i=1}^{m} \frac{\lambda}{\theta_i a(\mu_i - 1)^a} =: \frac{\mathcal{A}_m}{\theta}. \tag{4.60}$$

In view of (4.57), the second summation in (4.59) is bounded by

$$\begin{aligned}&\lambda^2 \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k\geq 0} \frac{\theta_i^{-1}\theta_j^{-1}}{(k+\mu_{\min})^{2+2a}[\ln(k+\mu_{\min})]^{2+b_i+b_j}} \\ &\leq \frac{\lambda^2}{\vartheta}\left\{\sum_{i=1}^{m} \frac{1}{\theta_i[\ln(\mu_{\min}-1)]^{b_i}}\right\}^2 =: \frac{\mathcal{B}_m}{\theta^2}, \end{aligned} \tag{4.61}$$

where $\vartheta := (1+2b_{\min})(\mu_{\min}-1)^{1+2a}\ln(\mu_{\min}-1)$. From Theorem 10, (4.59)-(4.61) and the definitions of $\mathsf{Q}_\infty(x^*)$, $\mathsf{J}(x^*)$, $\mathcal{A}_m$ and $\mathcal{B}_m$, we obtain (4.56).

We now prove the bound on the oracle complexity. Using the facts that $K \leq \mathsf{Q}_\infty(x^*)/\epsilon$ and $\mu_i \leq \epsilon^{-1}$, together with the definition of $N_{k,i}$ we obtain

$$\begin{aligned}\sum_{k=1}^{K}\sum_{i=1}^{m} 2N_{k,i} &\leq \sum_{k=1}^{K}\sum_{i=1}^{m} 2\left[\theta_i n_i \sigma(x^*)^2 (k+\mu_i)^{1+a}(\ln(k+\mu_i))^{1+b_i} + 1\right] \\ &\leq 2\max\{\theta_{\max} n_{\max}\sigma(x^*)^2, 1\} K \sum_{i=1}^{m}\left[(K+\mu_i)^{1+a}\left(\ln\left(K+\mu_i\right)\right)^{1+b_i} + 1\right] \\ &\leq 4\max\{\theta_{\max} n_{\max}\sigma(x^*)^2, 1\} K \sum_{i=1}^{m}\left[(K+\mu_i)^{1+a}\left(\ln\left(K+\mu_i\right)\right)^{1+b_i}\right] \\ &\leq 4\Phi \frac{(\mathsf{Q}_\infty(x^*)+1)^{2+a}}{\epsilon^{2+a}} \sum_{i=1}^{m}\left(\ln\left(\mathsf{Q}_\infty(x^*)\epsilon^{-1} + \epsilon^{-1}\right)\right)^{1+b_i}, \end{aligned} \tag{4.62}$$

using the fact that $1 \leq (K+\mu_i)^{1+a}\left(\ln\left(K+\mu_i\right)\right)^{1+b_i}$ for $i \in [m]$ in the third inequality and defining Φ as $\max\{\theta_{\max} n_{\max}\sigma(x^*)^2, 1\}$ in the rightmost expression of (4.62).

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

$$b_1 \geq b_i + 2\ln(i+1) - \ln \mathsf{S} \geq b_i + \frac{2\ln(i+1) - \ln \mathsf{S}}{\ln h} \Rightarrow h^{b_i} \leq \frac{\mathsf{S} h^{b_1}}{(i+1)^2}. \tag{4.63}$$

It follows from (4.63) that

$$\sum_{i=1}^m h^{b_i} \leq \mathsf{S} h^{b_1} \sum_{i=1}^m \frac{1}{(i+1)^2} \leq \mathsf{S} h^{b_1}. \tag{4.64}$$

From (4.62), the bounds (4.59)-(4.61), (4.64) the definitions of h, $\widehat{\mathsf{P}}(x^*)$, $\widehat{\mathsf{I}}(x^*)$, $\mathsf{Q}_\infty(x^*)$, $\mathsf{J}(x^*)$, $\mathcal{A}_m$ and $\mathcal{B}_m$ and the relation $(x+y+z)^{2+a} \leq 3^{1+a}(x^{2+a}+y^{2+a}+z^{2+a})$, we obtain the required bound on $\sum_{k=1}^K \sum_{i=1}^m 2N_{k,i}$. □

Remark 7 (Few initial iterates and complexity of $O(m)$)**.** We remark that for the choice of parameters (4.53)-(4.54), we have $\theta_{\min} \sim \theta m$ so that $\mathcal{A}_m \lesssim \frac{1}{a(\mu_{\min}-1)^a}$ and $\mathcal{B}_m \lesssim \frac{1}{(\mu_{\min}-1)^{1+2a}}$. Also, $b_{\min} \leq b_1 + \ln \mathsf{S} - 2\ln(m+1)$ so that it is enough to choose $b_1 > 2\ln(m+1) - \ln \mathsf{S}$, which is reasonably small in terms of m. Moreover, $\frac{n}{b_{\min}\theta_{\min} n_{\min}} \lesssim \frac{n_{\max}}{\theta n_{\min} m \ln m}$. Hence, in view of (4.55), k_0 is approximately constant and small for $m \gg 1$. Finally, the bound on the oracle complexity in Proposition 10 is of order $\max\{1, \theta^{-2\nu}\}\theta_{\max} n_{\max} \lesssim \max\{\theta, \theta^{-(3+2a)}\} m n_{\max}$, that is, it is linear in m. Moreover, the sampling is robust in the sense that the convergence rate is proportional to $\max\{1, \theta^{-2}\}$ and the oracle complexity is proportional to $\max\{\theta, \theta^{-(3+2a)}\}$. We remark that improvements can be achieved if a coordination $\mu_{\min} \sim \epsilon^{-1}$ is possible (given a prescribed tolerance $\epsilon > 0$).

For simplicity we do not present the analogous results of Proposition 9 for the case $m \gg 1$ under Assumption 19(iii). In that case, the estimates depend on $\mathrm{d}(x^0, X^*)$ and on smaller exponents of θ.

4.3.1 Comparison of complexity estimates

A merit function for $\mathrm{VI}(T, X)$ is a non-negative function f over X such that $X^* = X \cap f^{-1}(0)$. Next, we briefly compare our complexity results in terms of the quadratic natural residual, given in this section, with related results presented in the literature in terms of other merit functions for the stochastic variational inequality.

Given a compact feasible set X, the *dual gap-function* of $\mathrm{VI}(T, X)$ is defined as $G(x) := \sup_{y \in X} \langle T(y), x - y \rangle$ for $x \in X$. In [42, 20, 76, 77], a rate of convergence of $O(1/\sqrt{K})$ is given in terms of the expected value of G in the case in which X is compact or, in the case in which X is unbounded, in terms of the relaxed dual-gap function $\tilde{G}$, shown in (4.10). In the case in which X is compact, the dual gap-function is a modification of the *primal gap-function*, defined as $g(x) := \sup_{y \in X} \langle T(x), x - y \rangle$ for $x \in X$. Both the primal and dual gap-functions are continuous only if X is compact. A gap-function suitable for unbounded feasible sets is the *regularized gap-function*, defined, for fixed $a > 0$, as $g_a(x) := \sup_{y \in X} \{ \langle T(x), x - y \rangle - \frac{a}{2} \|x - y\|^2 \}$, for $x \in \mathbb{R}^n$. The regularized gap-function is continuous over $\mathbb{R}^n$. Another option is the so called *D-gap function*. It is defined, for fixed $b > a > 0$, as $g_{a,b}(x) := g_a(x) - g_b(x)$, for $x \in \mathbb{R}^n$. It is well known that $g_{a,b} : \mathbb{R}^n \to \mathbb{R}_+$ is a continuous unrestricted merit function of $\mathrm{VI}(T, X)$, i.e., $X^* = g_{a,b}^{-1}(0)$. Moreover, the quadratic natural residual and the D-gap function are equivalent merit functions in the sense that, given $b > a > 0$, for all $x \in \mathbb{R}^n$, $r_{b^{-1}}(x)^2 \lesssim g_{a,b}(x) \lesssim r_{a^{-1}}(x)^2$ (see [27], Theorems 10.2.3, 10.3.3 and Proposition 10.3.7). These properties hold independently of the compactness of X. An immediate consequence is that the previous complexity analysis, given in Theorems 10-11 and Propositions 8-10 in terms of the quadratic natural residual, are also valid in terms of the D-gap function. In this sense, our rate of convergence of $O(1/K)$ in terms of the D-gap function improves over the rate $O(1/\sqrt{K})$ in terms of the dual gap-functions analyzed in [42, 20, 76, 77].

From Proposition 8, if Assumption 19(ii) holds, then the algorithm performance, in terms of convergence rate and oracle complexity, depends on a $x^* \in X^*$ such that $\widehat{\mathsf{Q}}(x^*) := \sigma(x^*)^2 \max_{0 \leq k \leq k_0(x^*)} \mathbb{E}[\|x^k - x^*\|^2]^2$ is minimal, that is to say, we have a trade-off between variance of the oracle error and distance to initial iterates. We also remark that the sampling rate $\mathcal{N}_k$ possesses a *robust property*: a scaling in the sampling rate by a factor θ, keeps the algorithm running with a proportional scaling of $\max\{1, \theta^{-2}\}$ in the rate and $\max\{\theta, \theta^{-3}\}$ in the oracle complexity (see [60] for a discussion on robust algorithms). From Proposition 9, when the variance is bounded by σ^2 over X^*, the estimates depend on $\widehat{\mathsf{Q}} := \sigma^2 \max_{0 \leq k \leq k_0} \mathbb{E}[\mathrm{d}(x^k, X^*)^2]^2$ and k_0 is independent of any $x^* \in X^*$. When the variance is uniform over X, the estimates depend only on $\mathrm{d}(x^0, X^*)$ and a scaling factor θ in the sampling rate

implies a factor of $\max\{1, \theta^{-1}\}$ in the rate and of $\max\{\theta, \theta^{-1}\}$ in the oracle complexity. Interestingly, in the case of a compact feasible set, the estimates do not depend on $\operatorname{diam}(X)$, as in [42, 20], but rather on the distance of the initial iterates to X^*, which is a sharper result. In the case of networks the same conclusions hold, except that the dependence in the dimension is higher if a centralized sampling is used. From Proposition 10, if a distributed sampling is used and a coordination of a rapid decreasing sequence of positive numbers is implemented (in any order) then the oracle complexity depends linearly on the size of the network.

We briefly compare our convergence rate and complexity bounds presented in Propositions 8 and 9 with those in [20] (Corollaries 3.2 and 3.4). In [20], for a compact X and uniform variance over X, the convergence rate obtained in terms of the dual gap function is of order $L \operatorname{diam}(X)^2 K^{-1} + \sigma \operatorname{diam}(X) K^{-1/2}$, and the oracle complexity is of order $L \operatorname{diam}(X)^2 \epsilon^{-1} + \sigma^2 \operatorname{diam}(X)^2 \epsilon^{-2}$. For an unbounded X with uniform variance over X, the convergence rate in terms of gap function (4.10), is of order $L\|x^0 - x^*\|^2 K^{-1} + \sigma \|x^0 - x^*\|^2 K^{-1/2}$, while the oracle complexity is of order $L\|x^0 - x^*\|^2 \epsilon^{-1} + \sigma^2 \|x^0 - x^*\|^4 \epsilon^{-2}$. In the estimates given in Propositions 8-9, the "coercivity" modulus ρ^{-1} introduced by the extragradient step behaves qualitatively as L. We improve on the rate of convergence to $O(1/K)$ with respect to the stochastic term $\sigma/\sqrt{K}$ by reducing iteratively the variance, while preserving the complexity performance in terms of σ and ϵ (up to first order logarithm term). Differently from [20], our analysis is the same for a compact or unbounded X, in the sense that the same merit function is used. For the case of a compact X, our bounds depend on $\mathrm{d}(x^0, X^*)$ rather $\operatorname{diam}(X)$ as in [20], which is a sharper result. In the case the variance is uniform over an unbounded X, our bounds depend on $\mathrm{d}(x^0, X^*)$ instead of $\|x^0 - x^*\|$ for a given $x^* \in X^*$ as in [20], which is also a sharper bound. We analyze the new case of non-uniform variance, which has a similar performance, except that the estimates depend on a point $x^* \in X^*$ with a minimum trade-off between variance $\sigma(x^*)^2$ and distances to a few initial iterates. Moreover, we include asymptotic convergence, which it is not the case in [20] (see Example 1).

Finally, we discuss error bounds on the solution set. It is well known that important classes of variational inequalities admit the natural residual as an error bound for the solution set, i.e., for all $\alpha > 0$, there exists $\delta > 0$ such that for

all $x \in \mathbb{R}^n$ with $r_\alpha(x) \leq \delta$, there holds $\mathrm{d}(x, X^*) \lesssim r_\alpha(x)$. This property holds, for example, for (i) semi-stable VIs, (ii) composite strongly monotone VIs such that X is a polyhedron, (ii) VIs such that T is linear and X is cone (see [27]). Item (ii) includes affine VIs and strongly monotone VIs. Item (iii) includes linear homogeneous complementarity problems and linear system of equations. When such property holds, the results of Theorems 10-11 and Propositions 8-10 provide other classes of SVI for which convergence of $O(1/K)$ holds in terms of the mean-squared distance to the solution set. In the previous literature, such property was shown only for strongly pseudo-monotone or weak-sharp SVIs on a compact set.

4.4 Appendix of Chapter 4

Proof of Lemma 3

Proof. Set $A := D(\alpha)^{-1}$ for $\alpha = (\alpha_i)_{i=1}^m \in \mathbb{R}^m_{>0}$. We first prove that $\Pi_C \equiv \Pi_{C,A}$. Indeed, let $x = (x_i)_{i=1}^m$ and set $\hat{x} := \Pi_C(x)$ with $\hat{x} = (\hat{x}_i)_{i=1}^m$. It is not difficult to check, using Lemma 1(i), that $\hat{x}_i = \Pi_{C_i}(x_i)$ for $i \in [m]$. Given $y = (y_i)_{i=1}^m \in C$, we use the fact that $\alpha_i > 0$ and Lemma 1(i) with $\hat{x}_i = \Pi_{C_i}(x_i)$, $y_i \in C_i$ for every $i \in [m]$, in order to obtain $\langle x - \hat{x}, A(y - \hat{x})\rangle = \sum_{i=1}^m \alpha_i^{-1}\langle x_i - \hat{x}_i, y_i - \hat{x}_i\rangle \leq 0$. Again by Lemma 1(i), we conclude that $\hat{x} = \Pi_{C,A}(x)$ as claimed.

The required statement follows immediately from Lemma 1(iv) and the fact that $\Pi_C \equiv \Pi_{C,A}$. □

Proof of Lemma 9

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

$$\begin{aligned}
\|x^{k+1} - x\|^2 &= \|\Pi(y^k) - x\|^2 \\
&\leq \|y^k - x\|^2 - \|y^k - \Pi(y^k)\|^2 \\
&= \|(x^k - x) - D(\alpha_k)\widehat{F}(\epsilon_2^k, z^k)\|^2 - \|(x^k - x^{k+1}) - D(\alpha_k)\widehat{F}(\epsilon_2^k, z^k)\|^2 \\
&= \|x^k - x\|^2 - \|x^k - x^{k+1}\|^2 + 2\langle x - x^{k+1}, D(\alpha_k)\widehat{F}(\epsilon_2^k, z^k)\rangle \\
&= \|x^k - x\|^2 - \|x^k - x^{k+1}\|^2 + 2\langle x - z^k, D(\alpha_k)\widehat{F}(\epsilon_2^k, z^k)\rangle +
\end{aligned}$$

$$2\langle z^k - x^{k+1}, D(\alpha_k)\widehat{F}(\epsilon_2^k, z^k)\rangle = \|x^k - x\|^2 - \|(x^k - z^k) + (z^k - x^{k+1})\|^2 +$$
$$2\langle z^k - x^{k+1}, D(\alpha_k)\widehat{F}(\epsilon_2^k, z^k)\rangle + 2\langle x - z^k, D(\alpha_k)\widehat{F}(\epsilon_2^k, z^k)\rangle$$
$$= \|x^k - x\|^2 - \|x^k - z^k\|^2 - \|z^k - x^{k+1}\|^2$$
$$-2\langle x^k - z^k, z^k - x^{k+1}\rangle + 2\langle z^k - x^{k+1}, D(\alpha_k)\widehat{F}(\epsilon_2^k, z^k)\rangle +$$
$$2\langle x - z^k, D(\alpha_k)\widehat{F}(\epsilon_2^k, z^k)\rangle$$
$$= \|x^k - x\|^2 - \|x^k - z^k\|^2 - \|z^k - x^{k+1}\|^2 +$$

$$2\langle x^{k+1} - z^k, x^k - D(\alpha_k)\widehat{F}(\epsilon_2^k, z^k) - z^k\rangle + 2\langle x - z^k, D(\alpha_k)\widehat{F}(\epsilon_2^k, z^k)\rangle, \tag{4.65}$$

using Lemma 1(ii) in the inequality and simple algebra in the equalities.

Looking at the fourth term $\mathsf{I} := 2\langle x^{k+1} - z^k, x^k - D(\alpha_k)\widehat{F}(\epsilon_2^k, z^k) - z^k\rangle$ in the rightmost expression of (4.65), we take into account (4.7) and the fact that $\widehat{F}(\epsilon_2^k, z^k) = T(z^k) + \epsilon_2^k$, and then we apply Lemma 1(i) with $C = X$, $x = x^k - D(\alpha_k)(T(x^k) + \epsilon_1^k)$ and $y = x^{k+1} \in X$, obtaining:

$$\begin{aligned} \mathsf{I} &= 2\langle x^{k+1} - z^k, x^k - D(\alpha_k)(T(x^k) + \epsilon_1^k) - z^k\rangle + \\ &\quad 2\langle x^{k+1} - z^k, D(\alpha_k)\cdot\Big[(T(x^k) + \epsilon_1^k) - (T(z^k) + \epsilon_2^k)\Big]\rangle \\ &\leq 2\|D(\alpha_k)\|\|x^{k+1} - z^k\|\|(T(z^k) + \epsilon_2^k) - (T(x^k) + \epsilon_1^k)\|, \end{aligned} \tag{4.66}$$

using Cauchy-Schwartz inequality. Next we apply Lemma 1(iii) to (4.7)-(4.8), obtaining

$$\begin{aligned} \|x^{k+1} - z^k\| &= \|\Pi[x^k - D(\alpha_k)(T(z^k) + \epsilon_2^k)] - \Pi[x^k - D(\alpha_k)(T(x^k) + \epsilon_1^k)]\| \\ &\leq \|D(\alpha_k)\|\|(T(z^k) + \epsilon_2^k) - (T(x^k) + \epsilon_1^k)\|. \end{aligned} \tag{4.67}$$

Combining (4.66) and (4.67) we get

$$\begin{aligned} \mathsf{I} &\leq 2\|D(\alpha_k)\|^2\|(T(z^k) + \epsilon_2^k) - (T(x^k) + \epsilon_1^k)\|^2 \\ &\leq 4\|D(\alpha_k)\|^2\|T(z^k) - T(x^k)\|^2 + 4\|D(\alpha_k)\|^2\|\epsilon_2^k - \epsilon_1^k\|^2 \\ &\leq 4L^2\|D(\alpha_k)\|^2\|z^k - x^k\|^2 + 4\|D(\alpha_k)\|^2\|\epsilon_2^k - \epsilon_1^k\|^2, \end{aligned} \tag{4.68}$$

using the fact that $(a+b)^2 \leq 2a^2 + 2b^2$ in the second inequality and the Lipschitz continuity of T in the last one. We set $x := x^*$ in (4.65). Looking now at the last

term in the rightmost expression of (4.65), we get

$$\begin{aligned} 2\langle x^* - z^k, D(\alpha_k)\widehat{F}(\epsilon_2^k, z^k)\rangle &= 2\langle x^* - z^k, D(\alpha_k)(T(z^k) + \epsilon_2^k)\rangle \\ &= 2\langle x^* - z^k, D(\alpha_k)T(z^k)\rangle + 2\langle x^* - z^k, D(\alpha_k)\epsilon_2^k\rangle \\ &\leq 2\langle x^* - z^k, D(\alpha_k)\epsilon_2^k\rangle =: \mathsf{J}_k, \end{aligned} \tag{4.69}$$

using the fact that $\langle x^* - z^k, D(\alpha_k)T(z^k)\rangle \leq 0$, (which follows from Assumption 15, the facts that $x^* \in X^*$ and $z^k \in X$ and Lemma 3) in the last inequality of (4.69). Combining (4.65), (4.68) and (4.69), we get

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &\leq \|x^k - x\|^2 - \|z^k - x^k\|^2 - \|z^k - x^{k+1}\|^2 + \\ &\quad 4L^2\|D(\alpha_k)\|^2\|z^k - x^k\|^2 + 4\|D(\alpha_k)\|^2\|\epsilon_2^k - \epsilon_1^k\|^2 + \mathsf{J}_k \\ &\leq \|x^k - x\|^2 - \rho_k\|z^k - x^k\|^2 + 8\|D(\alpha_k)\|^2(\|\epsilon_2^k\|^2 + \|\epsilon_1^k\|^2) + \mathsf{J}_k, \end{aligned} \tag{4.70}$$

using the facts that $\|D(\alpha_k)\| = \alpha_{k,\max}$, $\rho_k = 1 - 4L^2\alpha_{k,\max}^2$ and $(a+b)^2 \leq 2a^2 + 2b^2$.

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

$$\begin{aligned} r_{\alpha_k}(x^k)^2 &= \|x^k - \Pi[x^k - D(\alpha_k)T(x^k)]\|^2 \\ &\leq 2\|x^k - z^k\|^2 + 2\|\Pi[x^k - D(\alpha_k)(T(x^k) + \epsilon_1^k)] - \Pi[x^k - D(\alpha_k)T(x^k)]\|^2 \\ &\leq 2\|x^k - z^k\|^2 + 2\|D(\alpha_k)\|^2\|\epsilon_1^k\|^2, \end{aligned} \tag{4.71}$$

using Lemma 1(iii) in the second inequality. From (4.70), (4.71), $\|D(\alpha_k)\| = \alpha_{k,\max}$ and definitions (4.13)-(4.14) and $\mathsf{J}_k = M_{k+1}(x^*) - M_k(x^*)$, we get the claimed relation. □

Proof of Lemma 10

Proof. We first prove the result under Assumption 19(i)-(ii). Consider first item 1). Assume first that $m > 1$ and take $i \in [m]$. For $1 \leq t \leq N_i$, define $U_i^t \in \mathbb{R}^{n_i}$ by

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

with real components $U_{1,i}^t, \ldots, U_{n_i,i}^t$. Defining $U_i^0 = 0$ and the natural filtration $\mathcal{G}_t := \sigma(\xi_{1,i}, \ldots, \xi_{t,i})$ for $0 \leq t \leq N_i$, then each $\{U_{l,i}^t, \mathcal{G}_t\}_{t=0}^{N_i}$ for $l \in [n_i]$ defines

a real valued martingale (since it is a sum of N_i independent mean-zero random variables) whose increments satisfy

$$\left|U_{l,i}^t - U_{l,i}^{t-1}\right|_p = \left|\frac{F_i(\xi,x) - T_i(x)}{N_i}\right|_p \leq \frac{|||F(\xi,x) - T(x)|||_p}{N_i} \leq \frac{\sigma(x^*)\,(1 + \|x - x^*\|)}{N_i},$$

by Assumption 19. Hence, for $l \in [n_i]$

$$\left|U_{l,i}^{N_i}\right|_p \leq \frac{C_p\,\sigma(x^*)\,(1 + \|x - x^*\|)}{\sqrt{N_i}}, \tag{4.72}$$

which follows from the BDG inequality (3). For each $i \in [m]$, $U_{1,i}^{N_i}, \ldots, U_{n_i,i}^{N_i}$ are the real components of $\epsilon_i(x) \in \mathbb{R}^{n_i}$. Hence, since $q \geq 1$, from Minkovski's inequality and (4.72) we get:

$$|||\epsilon(x)|||_p^2 = \left|\|\epsilon(x)\|^2\right|_q \leq \sum_{i=1}^{m}\sum_{l=1}^{n_i}\left|\left|U_{l,i}^{N_i}\right|^2\right|_q \leq C_p^2\left(\sum_{i=1}^{m}\frac{2n_i}{N_i}\right)\sigma(x^*)^2\,(1 + \|x - x^*\|^2), \tag{4.73}$$

using $(a+b)^2 \leq 2a^2 + 2b^2$. The first claim follows from (4.73) with $\mathsf{A} = 2n$. If $m = 1$, the same proof line holds with $\mathsf{A} = n$, since relation $(a+b)^2 \leq 2a^2 + 2b^2$ is not required.

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

$$|\langle v, D(\alpha)\epsilon(x)\rangle|_p \leq \|v\|\|D(\alpha)\|||\epsilon(x)|||_p, \tag{4.74}$$

by Cauchy-Schwarz inequality. The claim follows from (4.73) and (4.74) with $\mathsf{B} = 2n$.

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

$$\begin{aligned}\left|\left\langle v, D(\alpha)\frac{F(\xi_t,x) - T(x)}{N}\right\rangle\right|_p &\leq \frac{\|v\|\|D(\alpha)\|\,|||F(\xi,x) - T(x)|||_p}{N} \\ &\leq \frac{\|v\|\|D(\alpha)\|\sigma(x^*)(1 + \|x - x^*\|)}{N},\end{aligned} \tag{4.75}$$

using Cauchy-Schwarz inequality in the first inequality and Assumption 19 in the last one. Hence, from (4.75) and the BDG-inequality (3), we get the claim with $\mathsf{B} = 1$ (in this case $\mathcal{N} = N$).

The proof of the bounds under the stronger Assumption 19(iii) is essentially the same with sharper bounds on the increments, and so we omit it. □

Chapter 5

Stochastic extragradient methods with line search

The estimation in SA methods is measured by the *oracle error*. This is the map $\epsilon : \Xi \times X \to \mathbb{R}^d$ defined by

$$\epsilon(\xi, x) := F(\xi, x) - T(x), \qquad (\xi \in \Xi, x \in X). \tag{5.1}$$

For $p \geq 2$, the *oracle error's p-moment* function is defined by

$$\sigma_p(x) := \sqrt[p]{\mathbb{E}\left[\|\epsilon(\xi, x)\|^p\right]} \qquad (x \in X). \tag{5.2}$$

In the deterministic case, assumptions on the operator T provide local surrogate models to establish the convergence of methods which solve $\mathrm{VI}(T, X)$. In order to define and analyze SA methods, assumptions on the variance $\sigma(\cdot)^2 := \sigma_2(\cdot)^2$ (or even higher order moments) are as important as assumptions on T. This is because local surrogate models also need the estimation of T from the SO. In that respect, we will consider Lemma 1 which is a consequence of the following assumption.

Assumption 1 (Heavy-tailed Hölder continuous operators)**.** *Consider definition* (1)*. There exist* $\delta \in (0, 1]$ *and nonnegative random variable* $\mathsf{L} : \Xi \to \mathbb{R}_+$ *such that, for almost every* $\xi \in \Xi$, $\mathsf{L}(\xi) \geq 1$ *and, for all* $x, y \in X$,

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

Define $\mathsf{a} := 1$ *if* X *is compact and* $\mathsf{a} := 2$ *for a general* X*. We assume there exist* $x_* \in X$ *and* $p \geq 2$ *such that* $\mathbf{P}\left[\|F(\cdot, x_*)\|^{\mathsf{a}p}\right] < \infty$ *and* $\mathbf{P}\left[\mathsf{L}(\cdot)^{\mathsf{a}p}\right] < \infty$*. We define* $L := \mathbf{P}\mathsf{L}(\cdot)$ *and* $L_q := \sqrt[q]{\mathbf{P}[\mathsf{L}(\cdot)^q]} + L$ *for any* $q > 0$*.*

Lemma 1 (Hölder continuity of the mean and the standard deviation)**.** *Consider definitions* (5.1)-(5.2)*, suppose Assumption 1 holds and take* $q \in [p, 2p]$ *such that the integrability conditions of Assumption 1 are satisfied. Then* T *is* (L, δ)*-Hölder continuous on* X *and* $\sigma_q(\cdot)$ *is* (L_q, δ)*-Hölder continuous*[1] *on* X *with respect to the norm* $\|\cdot\|$*.*

From a practical point of view, our statistical analysis will be built upon the standard assumption of an *unbiased oracle with i.i.d. sampling* (UO). In the rest of the paper, it will be convenient to define the following quantities associated to an i.i.d. sample $\xi^N := \{\xi_j\}_{j=1}^N$ drawn from $\mathbf{P}$. Recall definitions (1) and (5.1). We define the *empirical mean operator* and the *oracle's empirical mean error* associated to ξ^N, respectively, by

$$\widehat{F}(\xi^N, x) := \frac{1}{N}\sum_{j=1}^{N} F(\xi_j, x), \qquad \widehat{\epsilon}(\xi^N, x) := \frac{1}{N}\sum_{j=1}^{N} \epsilon(\xi_j, x), \qquad (x \in X). \tag{5.3}$$

The main purpose of this chapter is the introduction of an extragradient method with a line search for determining the stepsizes, as was done by Khobotov [47] and by Iusem and Svaiter [39] for the deterministic case. The introduction of such a line 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 significant detrimental effect on the convergence rate. It also improves over the alternative of "small" exogenous stepsizes, (i.e., a summable sequence $\{\alpha_k\}$), considered in Chapter 3, which has also a very detrimental effect on the convergence. The intuition is that a line 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 line search is the Armijo search applied to the steepest descent method for unconstrained optimization problems, adapted to the VI problem in [47] and [39]. It is widely recognized that the Armijo search substantially enhances the numerical performance of the steepest descent method, compared with the variants which use exogenous stepsizes, be it summable ones, or dependent on the Lipschitz constant.

[1] We say T is (L, δ)-Hölder continuous if $\|T(x) - T(y)\| \leq L\|x - y\|^\delta$ for all $x, y \in X$.

All these nice properties make the extragradient methods with line search we propose more implementable.

Algorithm 5 (The stochastic extragradient method with line search)**.**

1. **Initialization:** *Choose the initial iterate* $x^0 \in \mathbb{R}^d$*, parameters* $\hat{\alpha}, \theta \in (0,1]$*,* $\lambda \in \left(0, \frac{1}{2\sqrt{2}}\right)$ *and* $\beta \in (0, \hat{\alpha}^{-1}]$*, the sample rate* $\{N_k\} \subset \mathbb{N}$ *and the sequence* $\{\delta_k\} \subset (0,\infty)$*.*

2. **Iterative step:** *Given iterate* x^k*, generate sample* $\xi^k := \{\xi_j^k\}_{j=1}^{N_k}$ *from* $\mathbf{P}$*. Then compute* $\widehat{F}(\xi^k, x^k) := N_k^{-1} \sum_{j=1}^{N_k} F(\xi_j^k, x^k)$ *and* $r^k := x^k - \Pi[x^k - \hat{\alpha}\widehat{F}(\xi^k, x^k)]$*. Set*

$$d^k := \begin{cases} 0, & \text{if } \|r^k\| > 0, \\ \delta_k \frac{\hat{x}^k - x^k}{\|\hat{x}^k - x^k\|}, \text{ for any } \hat{x}^k \in X \text{ such that } \hat{x}^k \neq x^k, & \text{if } \|r^k\| = 0. \end{cases} \tag{5.4}$$

Line search rule: define α_k *as the maximum* $\alpha \in \{\theta^\ell \hat{\alpha} : \ell \in \{0\} \cup \mathbb{N}\}$ *such that*

$$\alpha \left\| \widehat{F}\left(\xi^k, z^k(\alpha)\right) - \widehat{F}\left(\xi^k, x^k + d^k\right) \right\| \leq \lambda \|z^k(\alpha) - (x^k + d^k)\|, \tag{5.5}$$

where, for all $\alpha \in (0, \hat{\alpha}]$*,*

$$z^k(\alpha) := \Pi\left[x^k + d^k - \alpha\left(\widehat{F}(\xi^k, x^k) + \beta d^k\right)\right], \tag{5.6}$$

and $\widehat{F}\left(\xi^k, z^k(\alpha)\right) := N_k^{-1} \sum_{j=1}^{N_k} F(\xi_j^k, z^k(\alpha))$*.*

Extragradient step: Generate sample $\eta^k := \{\eta_j^k\}_{j=1}^{N_k}$ *from* $\mathbf{P}$ *and set*

$$z^k = \Pi\left[x^k - \alpha_k \widehat{F}(\xi^k, x^k)\right], \tag{5.7}$$

$$x^{k+1} = \Pi\left[x^k - \alpha_k \widehat{F}(\eta^k, z^k)\right]. \tag{5.8}$$

Note that if T is Lipschitz continuous with constant L, Algorithm 5 recovers Algorithm 3 in Chapter 4 if we set $0 < \inf_k \alpha_k \leq \sup_k \alpha_k = \hat{\alpha} < 1/2L$ (i.e., the line search rule (5.5) is satisfied in the first iteration with $\alpha_k := \hat{\alpha}$).

Remark 1 (Initialization of the line search rule)**.** *We make a remark regarding the exogenous parameters β and $\{\delta_k\}$ and the endogenous sequence $\{d^k\}$ defined in* Algorithm *5. By the definition of d^k in* (5.4) *and convexity of X, we have that, for all $k \in \mathbb{N}$,*

$$\|d^k\| \leq \delta_k, \quad x^k + d^k \in X. \tag{5.9}$$

Moreover, it can be shown that, if $\beta \in (0, \hat{\alpha}^{-1}]$, then, for all $\alpha \in (0, \hat{\alpha}]$ and $k \in \mathbb{N}$,

$$\|z^k(\alpha) - (x^k + d^k)\| > 0, \tag{5.10}$$

where $z^k(\alpha)$ is defined in (5.6) *(see the proof of Lemma 16 in the next section). In fact, the rule* (5.4) *chosen to update d^k could be replaced by any rule satisfying* (5.47)-(5.48)*.*

The purpose of β, $\{\delta_k\}$ and d^k is solely to initialize the line search rule with a well defined direction. In deterministic regimes, this is not needed since if $r^k = 0$ (see Algorithm *5), x^k is an exact solution and we can stop the algorithm. In our framework, we use a sampled-based line search scheme so that the termination criteria is generally not clear. By choosing β, $\{\delta_k\}$ and d^k as above, the sampled-based line search rule* (5.5)-(5.6) *is always clearly specified and terminates in a finitely number of iterations. The direction d^k serves merely as a small perturbation to address the case $r^k = 0$. Since $\|d^k\| \leq \delta_k$ holds for all k, we can set $\delta_k \to 0$ in any desired rate so to correct iteratively such small perturbations. In this way, the optimality of the iteration and oracle complexities of* Algorithm *5 are unaltered. We refer to the convergence analysis in the next section for further details.*

Remark 2 (Intuition for the line search scheme)**.** *The stochastic approximated line search* (5.5) *is motivated by [47]. We make some comments for the case $d^k = 0$ (see Remark 10). Using definition* (2.9)*,* (5.5) *can be rewritten as*

$$\left\|\widehat{F}\left(\xi^k, z^k(\alpha)\right) - \widehat{F}\left(\xi^k, x^k\right)\right\| \leq \lambda \frac{r_\alpha(H_k; x^k)}{\alpha}, \tag{5.11}$$

where $H_k := \widehat{F}(\xi^k, \cdot)$. Provided that $r_{\hat{\alpha}}(H_k; x^k) \neq 0$, the line search tests (5.49) *for decreasing $\alpha \in (0, \hat{\alpha}]$. The idea is that the right hand side of* (5.49) *does not increase by Lemma 4 while the left hand side tends to 0 by continuity of the operator. Hence,* (5.49) *will hold eventually.*

We now present the stochastic hyperplane projection method.

Algorithm 6 (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* $\{N_k\}$*.*

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:*

   Line search rule: *Find the maximum* $\alpha \in \{\theta^j \hat{\alpha} : j \in \mathbb{N}_0\}$ *such that*

   $$\left\langle \widehat{F}\left(\xi^k, \bar{z}^k(\alpha)\right), x^k - \Pi(g^k)\right\rangle \geq \frac{\lambda}{\beta_k}\|x^k - \Pi(g^k)\|^2, \tag{5.12}$$

   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*

   $$z^k := \bar{z}^k(\alpha_k) = \alpha_k \Pi\left[x^k - \beta_k \widehat{F}(\xi^k, x^k)\right] + (1-\alpha_k)x^k, \tag{5.13}$$

   $$x^{k+1} := \Pi\left[x^k - \gamma_k \widehat{F}(\xi_k, z^k)\right], \tag{5.14}$$

   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 [39], 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 6, 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 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 20).

Remark 8. *For simplicity of presentation, we have not introduced a initialization for the line search rule in method 6. A similar adaptation can be introduced in method 6 so that a more implementable stopping rule is used than the one presented in such method.*

A significant difference between Algorithms 6, 5 and Algorithm 3 in Chapter 4 is that the stepsize α_k obtained in the line search *depends on the sample* ξ^k. The inevitable consequence is that the errors $\{\bar{\epsilon}_2^k, \bar{\epsilon}_3^k\}$ in Algorithm 6 and ϵ_3^k in Algorithm 5 do *not* induce martingales. This complicates considerably the convergence analysis requiring other statistical tools (see Theorem 12).

5.1 An empirical process theory for DS-SA line search schemes

If L in Assumption 1 is known then the analysis of SA methods with the CSP can exploit the fact that the oracle error's define a *martingale difference*. This type of errors can be controlled in a relatively straightforward way (see Lemma 15 in Section 5.1.3). The main objective of this section is to prove the following theorem. This is will the most sensitive part of our analysis and it is the cornerstone tool to handle *nonmartingale-like* oracle errors obtained when stepsize DS-SA line search schemes are used to estimate an unknown L.

Theorem 12 (Local bound for the $\mathcal{L}^p$-norm of the correlated error in DS-SA line search schemes)**.** *Consider the* SVI *given by* (1.2) *and Definition 1 with solution set* X^*. *Let* $\xi^N := \{\xi_j\}_{j=1}^N$ *be an i.i.d sample drawn from* $\mathbf{P}$ *and let* $\alpha_N : \Xi \to [0, \hat{\alpha}]$ *be a random variable for some* $0 < \hat{\alpha} \leq 1$. *Suppose that Assumption 1 holds, recall definitions* (5.1)-(5.3) *and define* $\delta_1 := 0$ *if* $\delta = 1$ *and* $\delta_1 := 1$ *if* $\delta \in (0, 1)$.

Given $(\alpha, x) \in [0, \hat{\alpha}] \times X$, *we define*

$$z\left(\xi^N; \alpha, x\right) := \Pi\left[x - \alpha \widehat{F}\left(\xi^N, x\right)\right],$$

and $\overline{z}_\beta(\xi^N; \alpha, x) := \alpha z(\xi^N; \beta, x) + (1-\alpha)x$, *given* $\beta > 0$. *Then the following holds:*

(i) There exist positive constants $\{\mathsf{c}_i\}_{i=1}^4$ *(depending on* d, δ, p *and* $L_{2p}\hat{\alpha}$*) such*

that, for any $x \in X$ *and* $x^* \in X^*$,

$$\left|\left|\left|\widehat{\epsilon}\left(\xi^N, z(\xi^N; \alpha_N, x)\right)\right|\right|\right|_p \;\leq\; \frac{\mathsf{c}_1\sigma_{2p}(x^*) + \overline{L}_{2p}\left[\delta_1 \vee \|x - x^*\|^\delta\right]}{\sqrt{N}},$$

where $\overline{L}_{2p} := \mathsf{c}_2 L_2 + \mathsf{c}_3 L_p + \mathsf{c}_4 L_{2p}$.

(ii) If X *is compact, there exist positive constants* d_2 *and* C_p *(depending on* d, δ *and* p*) such that, for any* $x \in X$ *and* $x^* \in X^*$,

$$\left|\left|\left|\widehat{\epsilon}\left(\xi^N, z(\xi^N; \alpha_N, x)\right)\right|\right|\right|_p \;\leq\; \frac{C_p\sigma_p(x^*) + L_p^*\,\mathrm{diam}(X)^\delta}{\sqrt{N}},$$

where $L_p^* := \mathsf{d}_2 L_2 + pL_p$.

Up to universal constants, the same bounds above holds for $\left|\left|\left|\widehat{\epsilon}\left(\xi^N, \overline{z}_\beta(\xi^N; \alpha_N, x)\right)\right|\right|\right|_p$.

For further detail on the constants of Theorem 12, see Remark 9 in Section 5.1.3. To prove Theorem 12, we will crucially require intermediate results which rely on a branch of statistics called *Empirical Process Theory*. Let $\{X_j\}_{j=1}^N$ be a sequence of *independent* stochastic processes $X_j := (X_{j,t})_{t\in\mathcal{T}}$ indexed by a countable set $\mathcal{T}$ with real-valued random components $X_{j,t}$. The associated *empirical process* (EP) is the stochastic process $\mathcal{T} \in t \mapsto Z_t := \sum_{j=1}^N X_{j,t}$. An essential quantity in this theory is $Z := \sup_{t\in\mathcal{T}} Z_t$. If $\mathcal{T} = \{t\}$, then Z is simply a sum of independent random variables. Otherwise, Z is a much more complicated object. To understand Z, it is important to bound its expectation and variance. EPs arise in many different settings in mathematical statistics [11].

We apply EP theory as a novel way to successfully analyze stochastic approximated line search schemes. Referring to Algorithm 5 and Theorem 12, we have $z^k = z(\xi^k; \alpha_k, x^k)$ and must control the correlated error $\widehat{\epsilon}(\xi^k, z(\xi^k; \alpha_k, x^k))$. Our strategy is to construct an EP that *locally decouples* the dependence in $\widehat{\epsilon}(\xi^k, z^k)$ between ξ^k and z^k at the k-th iteration.[2] The intuition behind our decoupling technique is that, although z^k is a function of (ξ^k, x^k), z^k *lies at a ball* $\mathbb{B}_k$ *centered at any given* $x^* \in X^*$ *with radius of* $\mathcal{O}(\|x^k - x^*\| + \|\widehat{\epsilon}(\xi^k, x^k)\|)$. Based on this fact and that, by i.i.d. sampling, $\xi^k \perp\!\!\!\perp x^k$, we can decouple ξ^k and z^k using the following guidelines:

[2]Recall that such dependence is produced by the need to evaluate $\widehat{F}(\xi^k, \cdot)$ along the path $\alpha \mapsto z^k(\alpha)$ in order to choose the stepsize α_k. Analogous observations hold for (5.12): $z^k = \overline{z}_{\beta_k}(\xi^k; \alpha_k, x^k)$.

(i) we *condition* on the past information $\mathcal{F}_k$, noting that $x^k \in \mathcal{F}_k$ and $\xi^k \perp\!\!\!\perp \mathcal{F}_k$,

(ii) we then *control an* EP indexed by the ball $\mathbb{B}_k$,

(iii) we further note that in item (ii) we must also control $\widehat{\epsilon}(\xi^k, x^k)$ which affects the radius of the ball $\mathbb{B}_k$. Nevertheless, since $x^k \in \mathcal{F}_k$ and $\xi^k \perp\!\!\!\perp \mathcal{F}_k$, $\widehat{\epsilon}(\xi^k, x^k)$ is a *martingale difference* and, hence, easier to estimate.

The developed theory is presented in consecutive sections. The statistical preliminaries used outside the proofs are carefully introduced so to make the presentation as self contained as possible. We refer to the excelent book [11] by S. Boucheron, G. Lugosi and P. Massart, a standard reference in the area. A global outline is as follows. Typically, if $Z := \sup_{t\in\mathcal{T}} Z_t$ for a stochastic process $(Z_t)_{t\in\mathcal{T}}$, an upper bound on $\mathbb{E}[Z]$ is derived under a suitable tail property on the increments of $(Z_t)_{t\in\mathcal{T}}$ and chaining arguments [25]. In Section 5.1.1, we derive instead an upper bound on $|Z|_2 \geq \mathbb{E}[Z]$ in Lemma 13. The main reason to do so is that we assume *heavy-tailed* random operators satisfying Assumption 1. As a consequence, we will work with the *square* of *sub-Gaussian* random variables (see Definition 2). In Section 5.1.2, we apply Lemma 13 derived in Section 5.1.1 to obtain the general Lemma 14. This lemma provides an *uniform bound over a ball* on the $\mathcal{L}^p$-norm of *empirical error increments of heavy-tailed Hölder continuous operators*, the main stochastic object in this work. *Self-normalization* (see [62] and Theorem 15), variance bounds (Theorem 14) and a simple decoupling argument based on Hölder's inequality are also needed for that purpose. Finally, the proof of Theorem 12 is given in Section 5.1.3. It relies on Lemma 14, the Burkholder-Davis-Gundy's moment inequality for martingales in Hilbert spaces [13, 54] and the ideas of items (i)-(iii) above.

5.1.1 The $\mathcal{L}^2$-norm of suprema of sub-Gaussian processes

In order to bound the expectation or the $\mathcal{L}^2$-norm of $\sup_{t\in\mathcal{T}} Z_t$ for a stochastic process $(Z_t)_{t\in\mathcal{T}}$, it is important to understand the tail behavior of its increments $(Z_t - Z_{t'})_{(t,t')\in\mathcal{T}\times\mathcal{T}}$. We will thus need the definitions of *sub-Gaussian* and *sub-Gamma* random variables.

Definition 2 (sub-Gaussian and sub-Gamma random variables)**.** *A random variable $Y \in \mathbb{R}$ is called* sub-Gaussian with variance factor $\sigma^2 > 0$ *if, for all $s \in \mathbb{R}$,* $\ln \mathbb{E}\left[e^{sY}\right] \leq \frac{\sigma^2 s^2}{2}$. *A random variable $Y \in \mathbb{R}$ is called* sub-Gamma on the right tail with variance factor $\sigma^2 > 0$ and scale parameter $c > 0$ *if, for all $0 < s < \frac{1}{c}$,* $\ln \mathbb{E}\left[e^{sY}\right] \leq \frac{\sigma^2 s^2}{2(1-cs)}$.

Hence, a random variable Y is sub-Gaussian if Y and $-Y$ are sub-Gamma on the right tail with scale parameter $c = 0$. In order to compute $\mathcal{L}^2$-norms under heavier tails, we will need also the following result which establishes that the centered *square* of a sub-Gaussian random variable is sub-Gamma on the right tail. It follows, e.g., as a corollary of Theorem 2.1 and Remark 2.3 in [33] in the one dimensional setting.

Theorem 13 (Square of sub-Gaussian random variables)**.** *Suppose that $Y \in \mathbb{R}$ is a sub-Gaussian random variable with variance factor σ^2. Then, for all $0 \leq s < \frac{1}{2\sigma^2}$,* $\ln \mathbb{E}\left[e^{sY^2}\right] \leq \sigma^2 s + \frac{\sigma^4 s^2}{1-2\sigma^2 s}$.

One celebrated technique to understand $\sup_{t\in\mathcal{T}} Z_t$ for a stochastic process $(Z_t)_{t\in\mathcal{T}}$ is the so called *chaining method* (see e.g. [25]). This consists in approximating $\mathcal{T}$ by a increasing chain of finer discrete subsets. In this quest, the "complexity" of the index set $\mathcal{T}$ plays an important role. This is formalized in the next definition.

Definition 3 (Metric entropy)**.** *Let $(\mathcal{T}, d)$ be a totally bounded metric space. Given $\theta > 0$, a* θ-net *for $\mathcal{T}$ is a finite set $\mathcal{T}_\theta \subset \mathcal{T}$ of maximal cardinality $N(\theta, \mathcal{T})$ such that for all $s, t \in \mathcal{T}_\theta$ with $s \neq t$, one has $\mathrm{d}(s,t) > \theta$. The* θ-entropy number *is $H(\theta, \mathcal{T}) := \ln N(\theta, \mathcal{T})$. The function $H(\cdot, \mathcal{T})$ is called the* metric entropy *of $\mathcal{T}$.*

In particular, for all $t \in \mathcal{T}$, there is $s \in \mathcal{T}_\theta$ such that $\mathrm{d}(s,t) \leq \theta$. Note that the metric entropy is a nonincreasing real-valued function. The next lemma establishes the metric entropy of the Euclidean unit ball $\mathbb{B}$ of $\mathbb{R}^d$ (see Lemma 13.11 of [11]).

Lemma 11 (Metric entropy of Euclidean balls)**.** *Let $\mathbb{B}$ be the Euclidean unit ball of $\mathbb{R}^d$. For all $\theta \in (0, 1]$, $H(\theta, \mathbb{B}) \leq d \ln\left(1 + \frac{1}{\theta}\right)$.*

Hence, the "complexity" of $\mathbb{B}$ is proportional to d, an effect perceived in high-dimensional problems. However, note that $H(\theta, \mathbb{B})$ grows slowly when the dis-

cretization precision θ diminishes. This is a key property in order for the chaining method to work.

Before proving the main Lemma 13 in this section, we state one more needed preliminary result. It bounds the expectation of the maximum of a *finite* number of sub-Gamma random variables (see, e.g., Corollary 2.6 of [11]). It is an essential lemma while using discretization arguments.

Lemma 12 (Expectation of maxima of sub-Gamma random variables)**.** *Let $\{Y_i\}_{i=1}^N$ be real-valued sub-Gamma random variables on the right tail with variance factor $\sigma^2 > 0$ and scale parameter $c > 0$. Then*

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

Lemma 13 ($\mathcal{L}^2$-norm of suprema of sub-Gaussian processes)**.** *Let $(\mathcal{T}, d)$ be a totally bounded metric space and $\theta := \sup_{t\in\mathcal{T}} \mathrm{d}(t, t_0)$ for some $t_0 \in \mathcal{T}$. Suppose $(Z_t)_{t\in\mathcal{T}}$ is a continuous stochastic process for which there exist $a, v > 0$ and $\delta \in (0, 1]$ such that, for all $t, t' \in \mathcal{T}$ and all $\lambda > 0$,*

$$\ln \mathbb{E}[\exp\{\lambda(Z_t - Z_{t'})\}] \leq a\, \mathrm{d}(t, t')^\delta \lambda + \frac{v\, \mathrm{d}(t, t')^{2\delta}\lambda^2}{2}. \tag{5.15}$$

Then

$$\left|\sup_{t\in\mathcal{T}} Z_t - Z_{t_0}\right|_2 \leq (3\theta)^\delta \sqrt{2(a^2 + v)} \left[\frac{1}{2^\delta - 1} + \sum_{i=1}^{\infty} \frac{\sqrt[4]{8H\left(\theta 2^{-i}, \mathcal{T}\right)} + 2\sqrt{H\left(\theta 2^{-i}, \mathcal{T}\right)}}{2^{i\delta}}\right].$$

Proof. We first note that the continuity of $t \mapsto Z_t$ and separability of $\mathcal{T}$ imply that, for any continuous function f, $\sup_{t\in\mathcal{T}} f(Z_t)$ is measurable since it equals $\sup_{t\in\mathcal{T}'} f(Z_t)$ for a countable dense subset $\mathcal{T}'$ of $\mathcal{T}$.

Set $\mathcal{T}_0 := \{t_0\}$. Given $i \in \mathbb{N}$, we set $\theta_i := \theta 2^{-i}$ and denote by $\mathcal{T}_i$ a θ_i-net for $\mathcal{T}$ with maximal cardinality $N(\theta_i, \mathcal{T})$. We also denote by $\Pi_i : \mathcal{T} \to \mathcal{T}_i$ 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 that, for all $t \in \mathcal{T}$ and $i \in \mathbb{N}$, $\mathrm{d}(t, \Pi_i(t)) \leq \theta_i$. By the triangular inequality, this implies that for all $t \in \mathcal{T}$ and $i \in \mathbb{N}$,

$$\mathrm{d}(\Pi_i(t), \Pi_{i+1}(t)) \leq \theta_i + \theta_{i+1} = 3\theta_{i+1}. \tag{5.16}$$

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

$$Z_t = Z_{t_0} + \sum_{j=0}^{\infty} (Z_{\Pi_{i+1}(t)} - Z_{\Pi_i(t)}).$$

In the following, we denote $\Delta_i(t) := Z_{\Pi_{i+1}(t)} - Z_{\Pi_i(t)}$ for all $i \in \mathbb{N}$ and $t \in \mathcal{T}$. The above equality implies that $(Z_t - Z_{t_0})^2 = \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} \Delta_i(t)\Delta_k(t)$. Hence,

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in\mathcal{T}}(Z_t - Z_{t_0})^2\right] &\leq \sum_{i=0}^{\infty}\sum_{k=0}^{\infty}\mathbb{E}\left[\sup_{t\in\mathcal{T}}\{\Delta_i(t)\Delta_k(t)\}\right] \\
&\leq \sum_{i=0}^{\infty}\sum_{k=0}^{\infty}\left|\sup_{t\in\mathcal{T}}|\Delta_i(t)|\right|_2 \cdot \left|\sup_{t\in\mathcal{T}}|\Delta_k(t)|\right|_2 \\
&= \left[\sum_{i=0}^{\infty}\left|\sup_{t\in\mathcal{T}}|\Delta_i(t)|\right|_2\right]^2, \qquad (5.17)
\end{aligned}$$

using Hölder's inequality in the second inequality.

Fix $i \in \mathbb{N}$. Since $N(\theta_i, \mathcal{T}) \leq N(\theta_{i+1}, \mathcal{T})$, we have that

$$|\{(\Pi_i(t), \Pi_{i+1}(t)) : t \in \mathcal{T}\}| \leq N(\theta_{i+1}, \mathcal{T})^2 = e^{2H(\theta_{i+1})}. \qquad (5.18)$$

Relations (5.15) and (5.16) imply that, for all $t \in \mathcal{T}$,

$$\ln \mathbb{E}\left[e^{\lambda\Delta_i(t)}\right] \leq a\,\mathrm{d}\left(\Pi_i(t), \Pi_{i+1}(t)\right)^{\delta}\lambda + \frac{v\,\mathrm{d}\left(\Pi_i(t), \Pi_{i+1}(t)\right)^{2\delta}\lambda^2}{2} \leq a_i\lambda + \frac{v_i\lambda^2}{2},$$

where we have defined $a_i := a(3\theta_{i+1})^{\delta}$ and $v_i := v(3\theta_{i+1})^{2\delta}$. The above relation implies that, for all $t \in \mathcal{T}$, $\Delta_i(t) - a_i$ is sub-Gaussian with variance factor v_i. This, Theorem 13, the bound $\Delta_i(t)^2 \leq 2[\Delta_i(t) - a_i]^2 + 2a_i^2$ and the change of variables $\lambda \mapsto 2\lambda$ imply that, for all $t \in \mathcal{T}$ and $0 < \lambda < \frac{1}{4v_i}$,

$$\ln \mathbb{E}\left[e^{\lambda\Delta_i(t)^2}\right] \leq 2(a_i^2 + v_i)\lambda + \frac{4v_i^2\lambda^2}{(1 - 4v_i\lambda)}, \qquad (5.19)$$

that is, for all $t \in \mathcal{T}$, $\Delta_i(t)^2 - 2(a_i^2 + v_i)$ is sub-Gamma on the right tail with variance factor $8v_i^2$ and scale parameter $4v_i$. Relations (5.18)-(5.19) and Lemma 12 imply further that

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in\mathcal{T}}\Delta_i(t)^2\right] &\leq 2(a_i^2 + v_i) + \sqrt{2 \cdot 8v_i^2 \cdot 2H(\theta_{i+1}, \mathcal{T})} + 4v_i \cdot 2H(\theta_{i+1}, \mathcal{T}) \\
&\leq 2 \cdot 9^{\delta}(a^2 + v)\left[\theta_{i+1}^{2\delta} + \theta_{i+1}^{2\delta}\sqrt{8H(\theta_{i+1}, \mathcal{T})} + 4\theta_{i+1}^{2\delta}H(\theta_{i+1}, \mathcal{T})\right].
\end{aligned}$$

Taking the square root in the above relation we get

$$\left|\sup_{t\in\mathcal{T}}|\Delta_i(t)|\right|_2 \leq 3^{\delta}\sqrt{2(a^2 + v)}\left[\theta_{i+1}^{\delta} + \theta_{i+1}^{\delta}\sqrt[4]{8H(\theta_{i+1}, \mathcal{T})} + 2\theta_{i+1}^{\delta}\sqrt{H(\theta_{i+1}, \mathcal{T})}\right]. \qquad (5.20)$$

We now take the square root in (5.17) and use (5.20), valid for any $i \in \mathbb{N}$, obtaining

$$\left|\sup_{t\in\mathcal{T}} Z_t - Z_{t_0}\right|_2 \leq 3^\delta \sqrt{2(a^2+v)} \left[\sum_{i=1}^{\infty} \theta_i^\delta + \sum_{i=1}^{\infty} \theta_i^\delta \sqrt[4]{8H(\theta_i,\mathcal{T})} + 2\sum_{i=1}^{\infty} \theta_i^\delta \sqrt{H(\theta_i,\mathcal{T})}\right].$$

To finish the proof, we use $\theta_i = \theta 2^{-i}$ and $\sum_{i=1}^{\infty} \theta_i^\delta = \frac{\theta^\delta}{2^\delta - 1}$ in the above inequality. □

5.1.2 Heavy-tailed Hölder continuous operators: self-normalization and sup-norms

We will now focus on bounds of EPs associated to sums of the form $x \mapsto \sum_{j=1}^{N} \frac{F(\xi_j,x)-T(x)}{N}$, where $\{\xi_j\}_{j=1}^N$ is an i.i.d. sample of $\mathbf{P}$ and $F : \Xi \times X \to \mathbb{R}^d$ satisfies Assumption 1. The main result proved in this section is Lemma 14. Its proof will need Lemma 13 and the following theorem (see Theorem 15.14 in [11]).

Theorem 14 ($\mathcal{L}^q$-norm for suprema of EPs)**.** *Let $\{X_j\}_{j=1}^N$ be an independent sequence of stochastic processes $X_j := (X_{j,t})_{t\in\mathcal{T}}$ indexed by a countable set $\mathcal{T}$ with real-valued random components $X_{j,t}$ such that $\mathbb{E}[X_{j,t}] = 0$ and $\mathbb{E}[X_{j,t}^2] < \infty$ for all $t \in \mathcal{T}$ and $j \in [N]$. Define $Z := \sup_{t\in\mathcal{T}} \left|\sum_{j=1}^N X_{j,t}\right|$ and*

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

Set $\kappa := \frac{\sqrt{e}}{2(\sqrt{e}-1)} < 1.271$. Then, for all $q \geq 2$,

$$|Z|_q \leq 2\mathbb{E}[Z] + 2\sqrt{2\kappa q}\widehat{\sigma} + 4\sqrt{\kappa q}|M|_2 + 20\kappa q|M|_q.$$

In order to cope with a heavy-tailed $\mathsf{L}(\xi)$ in Assumption 1, we will need Theorem 15, a result due to Panchenko (see Theorem 1 in [62] or Theorem 12.3 in [11]). It establishes a sub-Gaussian tail for the deviation of an EP around its mean after a proper *normalization* with respect to a *random* quantity V. In our set-up, the standard Hölder continuous assumption turns out to be sufficient to estimate this quantity.

Theorem 15 (Panchenko's inequality for self-normalized EPs)**.** *Consider a countable family $\mathcal{G}$ of measurable functions $f : \Xi \to \mathbb{R}$ such that $\mathbf{P}f(\cdot)^2 < \infty$. Let*

$\{\xi_j\}_{j=1}^N$ *and* $\{\eta_j\}_{j=1}^N$ *be i.i.d. samples of* $\mathbf{P}$ *independent of each other. Set*

$$Y := \sup_{f\in\mathcal{G}} \sum_{j=1}^N f(\xi_j), \qquad \textit{and} \qquad V := \mathbb{E}\left\{ \sup_{f\in\mathcal{G}} \sum_{j=1}^N \left[f(\xi_j) - f(\eta_j)\right]^2 \middle| \xi_1, \dots, \xi_N \right\}.$$

Then there exists an universal constant $\mathsf{c} > 0$ *such that, for all* $t > 0$*,*

$$\mathbb{P}\left\{ Y - \mathbb{E}[Y] \geq \mathsf{c}\sqrt{V(1+t)} \right\} \bigvee \mathbb{P}\left\{ Y - \mathbb{E}[Y] \leq -\mathsf{c}\sqrt{V(1+t)} \right\} \leq e^{-t}.$$

Finally, before proving Lemma 14, we will need Theorem 16 which is a standard tail characterization of sub-Gaussian random variables. Theorem 2.1 in [11] gives a proof for the case $\mathbb{E}[\tilde{Y}] = 0$. The adaptation for the general case is immediate using the facts that $\mathbb{E}[e^{-t\tilde{Y}}] \geq e^{-t\mathbb{E}[\tilde{Y}]}$ by Jensen's inequality, the integral formula $\mathbb{E}[\tilde{Y}] \leq \mathbb{E}[|\tilde{Y}|] = \int_0^\infty \mathbb{P}(|\tilde{Y}| > t)\,\mathrm{d}\,t$ and $\int_0^\infty e^{-\frac{t^2}{2}}\,\mathrm{d}\,t = \sqrt{\frac{\pi}{2}}$.

Theorem 16 (Tail characterization of sub-Gaussian random variables)**.** *If* $\tilde{Y} \in \mathbb{R}$ *is a random variable such that, for some* $v > 0$ *and for all* $t > 0$*,*

$$\mathbb{P}\left\{ \tilde{Y} \geq \sqrt{2vt} \right\} \bigvee \mathbb{P}\left\{ \tilde{Y} \leq -\sqrt{2vt} \right\} \leq e^{-t},$$

then, for all $t > 0$*, we have* $\ln \mathbb{E}\left[e^{t\tilde{Y}}\right] \leq e^{\sqrt{\frac{v\pi}{2}}t + 8vt^2}$.

We now prove the main lemma of this section. It uses Lemma 13 and Theorems 14-16.

Lemma 14 (Local uniform bound for the $\mathcal{L}^p$-norm of empirical error increments)**.** *Consider* (1.2) *and let* $\xi^N := \{\xi_j\}_{j=1}^N$ *be an i.i.d. sample from* $\mathbf{P}$*. Suppose that Assumption 1 holds and recall definitions* (5.1)-(5.3)*. Given* $x_* \in X$ *and* $R > 0$*, we define*

$$Z := \sup_{x\in\mathbb{B}[x_*,R]\cap X} \left\| \widehat{\epsilon}(\xi^N, x) - \widehat{\epsilon}(\xi^N, x_*) \right\|. \tag{5.21}$$

Then

$$|Z|_p \lesssim \left[\frac{3^\delta \sqrt{d} L_2}{\sqrt{\delta}\left(\sqrt{2}^\delta - 1\right)} + \sqrt{p} L_2 + p L_p \right] \frac{R^\delta}{\sqrt{N}}.$$

Proof. A first step is to rewrite Z as the supremum of a suitable EP and use Theorem 14. In the following, we define the set $\mathbb{B}_X := \{u \in \mathbb{B} : x_* + Ru \in X\}$ for $x_* \in X$ and $R > 0$ as stated in the theorem. Note that

$$\begin{aligned} Z &= \sup_{u\in\mathbb{B}_X} \frac{1}{N} \left\| \sum_{j=1}^{N} \epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*) \right\| \\ &= \sup_{u\in\mathbb{B}_X} \frac{1}{N} \sup_{y\in\mathbb{B}} \left\langle \sum_{j=1}^{N} \epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*), y \right\rangle \qquad (5.22) \\ &= \sup_{(u,y)\in\mathbb{B}_X\times\mathbb{B}} \frac{1}{N} \sum_{j=1}^{N} \langle \epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*), y \rangle , \end{aligned}$$

where the second equality uses the fact that $\|\cdot\| = \sup_{y\in\mathbb{B}}\langle y, \cdot\rangle$. Next, we define the index set $\mathcal{T} := \mathbb{B}_X \times \mathbb{B}$ and, for every $j \in [N]$ and $t := (u, y) \in \mathbb{B}_X \times \mathbb{B}$, we define the random variables

$$X_{j,t} := \frac{1}{N} \langle \epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*), y \rangle , \qquad (5.23)$$

$$\tilde{Z}_t := \sum_{j=1}^{N} X_{j,t}. \qquad (5.24)$$

From Assumption 1, it is not difficult to show that, for every $j \in [N]$, the process $\mathcal{T} \ni t \mapsto X_{j,t}$ is Hölder continuous with respect to the metric

$$\mathrm{d}(t, t') := \|u - u'\| + \|y - y'\|. \qquad (5.25)$$

This fact, the separability of $\mathcal{T}$ and (5.22), imply that $(\tilde{Z}_t)_{t\in\mathcal{T}}$ is a continuous process and $Z = \sup_{t\in\mathcal{T}_0} \tilde{Z}_t = \sup_{t\in\mathcal{T}_0} \left|\tilde{Z}_t\right|$ is measurable, where $\mathcal{T}_0$ is a dense countable subset of $\mathcal{T}$. Hence, we may assume next that $\mathcal{T}$ is countable without loss on generality. Our next objective is to use Theorem 14, bounding $|Z|_p$ in terms of $\mathbb{E}[Z]$, M and $\widehat{\sigma}^2$.

PART 1 (An upper bound on $\mathbb{E}[Z]$): To bound $\mathbb{E}[Z]$ we will need Lemma 13 and Theorems 15-16. At this point, let's fix $t = (u, y) \in \mathcal{T}$ and $t' = (u', y') \in \mathcal{T}$ and define the measurable function

$$f(\cdot) := \frac{1}{N} \langle \epsilon(\cdot, x_* + Ru) - \epsilon(\cdot, x_*), y \rangle - \frac{1}{N} \langle \epsilon(\cdot, x_* + Ru') - \epsilon(\cdot, x_*), y' \rangle .$$

We have that $\mathbf{P}f(\cdot)^2 < \infty$ since $\||F(\xi, \cdot)\||_2 < \infty$ on X (Assumption 1). By construction and (5.23)-(5.24), we have $f(\xi_j) = X_{j,t} - X_{j,t'}$ for all $j \in [N]$ and

$\tilde{Z}_t - \tilde{Z}_{t'} = \sum_{j=1}^N f(\xi_j)$. Note also that $\mathbb{E}\left[\sum_{j=1}^N f(\xi_j)\right] = 0$, using (1.2), (5.1) and that $\{\xi_j\}_{j\in[N]}$ is an i.i.d. sample of $\mathbf{P}$.

The previous observations allow us to claim Theorem 15 with $\mathcal{G} := \{f\}$ and $Y := \sum_{j=1}^N f(\xi_j)$. Precisely, if $\{\eta_j\}_{j=1}^N$ is an i.i.d. sample from $\mathbf{P}$ which is independent of $\{\xi_j\}_{j=1}^N$, then Theorem 15 and $\mathbb{E}\left[\sum_{j=1}^N f(\xi_j)\right] = 0$ imply that, for all $\lambda > 0$,

$$\mathbb{P}\left\{\sum_{j=1}^N f(\xi_j) \geq \mathsf{c}\sqrt{V(1+\lambda)}\right\} \bigvee \mathbb{P}\left\{\sum_{j=1}^N f(\xi_j) \leq -\mathsf{c}\sqrt{V(1+\lambda)}\right\} \leq e^{-\lambda}, \tag{5.26}$$

for some universal constant $\mathsf{c} > 0$ and

$$V := \mathbb{E}\left[\sum_{j=1}^N [f(\xi_j) - f(\eta_j)]^2 \middle| \xi_1, \ldots, \xi_N\right].$$

We will now give an upper bound on V. Given $\xi \in \Xi$, (1.2), (5.1) and Hölder continuity of $F(\xi,\cdot)$ and T (Assumption 1 and Lemma 1) imply that $\epsilon(\xi,\cdot)$ is $(\mathsf{L}(\xi) + L, \delta)$-Hölder continuous on X. This, definition of f and the facts that $y, y, u, u' \in \mathbb{B}$ and $x_* + Ru, x_* + Ru' \in X$ imply that, for all $j \in [N]$ and $\Delta f_j := N\,|[f(\xi_j) - f(\eta_j)]|$,

$$\begin{aligned}
\Delta f_j &\leq |\langle \epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*) - \epsilon(\eta_j, x_* + Ru) + \epsilon(\eta_j, x_*), y - y'\rangle| \\
&+ |\langle \epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_* + Ru') - \epsilon(\eta_j, x_* + Ru) + \epsilon(\eta_j, x_* + Ru'), y'\rangle| \\
&\leq [\mathsf{L}(\xi_j) + \mathsf{L}(\eta_j) + 2L]\, R^\delta \left[\|y - y'\| + \|u - u'\|^\delta\right] \\
&\leq [\mathsf{L}(\xi_j) + \mathsf{L}(\eta_j) + 2L]\, R^\delta 2^{1-\delta} \left[\|y - y'\|^{\frac{1}{\delta}} + \|u - u'\|\right]^\delta \\
&\leq [\mathsf{L}(\xi_j) + \mathsf{L}(\eta_j) + 2L]\, R^\delta 2^{(1-\delta)} \left[\|y - y'\| + \|u - u'\|\right]^\delta,
\end{aligned}$$

where we used concavity of $\mathbb{R}_+ \ni x \mapsto x^\delta$ in third inequality and the fact that $\|y - y'\|^{\frac{1}{\delta}} \leq 2^{\frac{(1-\delta)}{\delta}} \|y - y'\|$ for $y, y' \in \mathbb{B}$ in last inequality. We take squares in the above inequality, use relation $(\sum_{i=1}^3 a_i)^2 \leq 3\sum_{i=1}^3 a_i^2$ and definitions of V and (5.25). We thus obtain

$$\begin{aligned}
V &\leq \frac{3 \cdot 4^{1-\delta} R^{2\delta}\, \mathrm{d}(t,t')^{2\delta}}{N} \left\{\sum_{j=1}^N \frac{\mathsf{L}(\xi_j)^2}{N} + \sum_{j=1}^N \frac{\mathbb{E}\left[\mathsf{L}(\eta_j)^2 | \xi_1, \ldots, \xi_N\right]}{N} + 4L^2\right\} \\
&= \frac{3 \cdot 4^{1-\delta} R^{2\delta}\, \mathrm{d}(t,t')^{2\delta} W_N^2}{N}, \qquad (5.27)
\end{aligned}$$

where we have defined

$$W_N := \sqrt{\frac{1}{N}\sum_{j=1}^{N} \mathsf{L}(\xi_j)^2 + |\mathsf{L}(\xi)|_2^2 + 4L^2}, \tag{5.28}$$

and used that $\{\eta_j\}_{j\in[N]}$ is an i.i.d. sample of $\mathbf{P}$ independent of $\{\xi_j\}_{j\in[N]}$.

Set $\tilde{Y} := \frac{\tilde{Z}_t - \tilde{Z}_{t'}}{W_N} - \frac{\sqrt{3}\mathsf{c}2^{1-\delta}R^\delta \,\mathrm{d}(t,t')^\delta}{\sqrt{N}}$. Relations (5.26)-(5.27) and $\sum_{j=1}^N f(\xi_j) = \tilde{Z}_t - \tilde{Z}_{t'}$, together with $\sqrt{1+\lambda} \le 1 + \sqrt{\lambda}$ for $\lambda > 0$, imply that

$$\mathbb{P}\left\{\tilde{Y} \ge \frac{\sqrt{3}\mathsf{c}2^{1-\delta}R^\delta \,\mathrm{d}(t,t')^\delta}{\sqrt{N}}\sqrt{\lambda}\right\} \bigvee \mathbb{P}\left\{\tilde{Y} \le -\frac{\sqrt{3}\mathsf{c}2^{1-\delta}R^\delta \,\mathrm{d}(t,t')^\delta}{\sqrt{N}}\sqrt{\lambda}\right\} \le e^{-\lambda}.$$

The above relation and Theorem 16 imply that for some universal constants $C_1, C_2 > 0$ and for all $\lambda > 0$,

$$\ln \mathbb{E}\left[\exp\left\{\frac{(\tilde{Z}_t - \tilde{Z}_{t'})}{W_N}\lambda\right\}\right] \le \frac{C_1 2^{1-\delta}R^\delta \,\mathrm{d}(t,t')^\delta}{\sqrt{N}}\lambda + \frac{C_2^2 4^{1-\delta}R^{2\delta}\,\mathrm{d}(t,t')^{2\delta}}{2N}\lambda^2. \tag{5.29}$$

We now observe that (5.29) holds for any $t, t' \in \mathcal{T}$. Inequality (5.29) and Lemma 13 with $(\mathcal{T}, \mathrm{d})$ as defined in (5.25), the continuous process $\mathcal{T} \ni t \mapsto Z_t := \frac{\tilde{Z}_t}{W_N}$, $t_0 := (0,0)$, $\theta := \sup_{t\in\mathcal{T}} \mathrm{d}(t,0) \le 2$, $a := \frac{C_1 2^{1-\delta}R^\delta}{\sqrt{N}}$ and $v := \frac{C_2^2 4^{1-\delta}R^{2\delta}}{N}$ imply that

$$\left|\sup_{t\in\mathcal{T}} Z_t\right|_2 \le \frac{\sqrt{2}C2^{1-\delta}(6R)^\delta}{\sqrt{N}}\left[\frac{1}{2^\delta - 1} + \sum_{i=1}^{\infty}\frac{\sqrt[4]{8H\left(2^{-i+1},\mathcal{T}\right)} + 2\sqrt{H\left(2^{-i+1},\mathcal{T}\right)}}{2^{i\delta}}\right], \tag{5.30}$$

where we defined $C = \sqrt{C_1^2 + C_2^2}$ and used the fact that $Z_{t_0} = \frac{\tilde{Z}_{t_0}}{W_N} = 0$. From Lemma 11 and the fact that, for any $\theta > 0$, $H(\theta, \mathbb{B}_X \times \mathbb{B}) \le H(\theta, \mathbb{B}_X) + H(\theta, \mathbb{B}) \le 2H(\theta, \mathbb{B})$, we also have that

$$\begin{aligned}\sum_{i=1}^{\infty}\frac{\sqrt[4]{8H\left(2^{-i+1},\mathcal{T}\right)} + 2\sqrt{H\left(2^{-i+1},\mathcal{T}\right)}}{2^{i\delta}} &\lesssim \sqrt{d}\sum_{i=1}^{\infty}\frac{\sqrt{\ln(1+2^{i+1})}}{2^{i\delta}} \\ &\lesssim \sqrt{d}\sum_{i=1}^{\infty}\frac{\sqrt{i+1}}{2^{i\delta}} \lesssim \frac{\sqrt{d/\delta}}{2^{\frac{\delta}{2}}-1},\end{aligned} \tag{5.31}$$

where we used the facts that $\ln(1+x) \le x$, $\sqrt{i+1} \le \frac{2^{\frac{i\delta}{2}}}{\sqrt{\delta}\ln 2}$ and[3] $\sum_{i=1}^{\infty} 2^{-\frac{i\delta}{2}} = \frac{1}{2^{\frac{\delta}{2}}-1}$.

[3]The previous fact can be derived from the inequality $2^x \ge 1 + (\ln 2)x$.

Hölder's inequality implies that

$$\mathbb{E}[Z] = \mathbb{E}\left[\sup_{t\in\mathcal{T}} |\tilde{Z}_t|\right] = \mathbb{E}\left[\sup_{t\in\mathcal{T}} |Z_t| \cdot W_N\right] \leq \left|\sup_{t\in\mathcal{T}} |Z_t|\right|_2 \cdot |W_N|_2. \tag{5.32}$$

Since $\{\xi_j\}_{j\in[N]}$ is an i.i.d. sample from $\mathbf{P}$, we also obtain from (5.28) that $|W_N|_2 \lesssim |\mathsf{L}(\xi)|_2 + L = L_2$. Finally, this, relations (5.30)-(5.32) and the facts that $2^{1-\delta}6^\delta = 2\cdot 3^\delta$ and $2^\delta - 1 \geq 2^{\frac{\delta}{2}} - 1$ imply that

$$\mathbb{E}[Z] \lesssim \frac{\sqrt{d}(3R)^\delta L_2}{\left(2^{\frac{\delta}{2}} - 1\right)\sqrt{\delta N}}. \tag{5.33}$$

PART 2 (An upper bound on M and $\widehat{\sigma}^2$): From the definition of $\widehat{\sigma}^2$ in Theorem 14 and (5.23), we get

$$\begin{aligned}
\widehat{\sigma} &= \sqrt{\sup_{(u,y)\in\mathcal{T}} \frac{1}{N^2}\sum_{j=1}^N \mathbb{E}\left[\langle \epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*), y\rangle^2\right]} \\
&\leq \sqrt{\frac{1}{N}\sup_{(u,y)\in\mathcal{T}} \mathbb{E}\left[\sum_{j=1}^N \frac{(\mathsf{L}(\xi_j)+L)^2}{N} R^{2\delta}\|u\|^{2\delta}\|y\|^2\right]} \\
&\leq \frac{R^\delta(|\mathsf{L}(\xi)|_2 + L)}{\sqrt{N}}, \qquad (5.34)
\end{aligned}$$

where we used the fact that $\|\epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*)\| \leq [\mathsf{L}(\xi_j) + L]R^\delta$ for $u \in \mathbb{B}_X$ (Assumption 1 and Lemma 1) in first inequality and the fact that $\{\xi_j\}_{j\in[N]}$ is an i.i.d. sample of $\mathbf{P}$ in the last inequality.

From the definition of M in Theorem 14 and (5.23), we get

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

where, again, we used the fact that $\|\epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*)\| \leq [\mathsf{L}(\xi_j) + L]R^\delta$ for $u \in \mathbb{B}_X$ in second inequality and the fact that $\{\xi_j\}_{j\in[N]}$ is an i.i.d. sample of $\mathbf{P}$ in

the last inequality. We take the p-th root in the above inequality and note that for $p \geq 2$ we have $N^{\frac{p-1}{p}} \geq \sqrt{N}$, obtaining

$$|M|_p \leq \frac{(|\mathsf{L}(\xi)|_p + L)R^\delta}{\sqrt{N}}. \tag{5.35}$$

From (5.33)-(5.35) and definitions of L_2 and L_p in Assumption 1, we obtain the required claim. □

5.1.3 The proof of Theorem 12

With the theory developed in Sections 5.1.1-5.1.2, we are now ready to prove Theorem 12. We shall use Lemma 14 and follow the ideas of items (i)-(iii) presented in the introduction of Section 5.1. We will also need the next Lemma 15 which controls the oracle's empirical error. Its control is easier than the oracle's correlated error, since it defines a martingale difference. Its proof uses Assumption 1 and a version of Burkholder-Davis-Gundy's inequality in Hilbert spaces (see [13, 54]), which we recall here for convenience.

Theorem 17 (Burkholder-Davis-Gundy inequality in $\mathbb{R}^d$)**.** *Let $\|\cdot\|$ be the Euclidean norm in $\mathbb{R}^d$. Then, for all $q \geq 2$, there exists $C_q > 0$ such that for any vector-valued martingale $\{y_j\}_{j=0}^N$ adapted to the filtration $\{\mathcal{G}_j\}_{j=1}^N$ with $y_0 = 0$, it holds that*

$$\left|\sup_{j\leq N}\|y_j\|\right|_q \leq C_q \left|\sqrt{\sum_{j=1}^N \|y_j - y_{j-1}\|^2}\right|_q \leq C_q \sqrt{\sum_{j=1}^N |||y_j - y_{j-1}|||_q^2}.$$

Lemma 15 (Local bound for the $\mathcal{L}^q$-norm of the empirical error)**.** *Consider* (1.2) *and let $\xi^N := \{\xi_j\}_{j=1}^N$ be an i.i.d. sample from $\mathbf{P}$. Suppose that Assumption 1 holds and take $q \in [p, 2p]$ such that the integrability conditions of Assumption 1 are satisfied. Recall definitions in* (5.1)-(5.3) *and definition of C_q in Theorem 17. Set $C_2 := 1$ if $q = p = 2$. Then, for any $x, x_* \in X$,*

$$\left|\left\|\widehat{\epsilon}(\xi^N, x)\right\|\right|_q \leq C_q \frac{\sigma_q(x^*) + L_q\|x - x^*\|^\delta}{\sqrt{N}}.$$

Proof. We define the $\mathbb{R}^d$-valued process $\{y_t\}_{t=0}^N$ by $y_0 = 0$ and $y_t := \sum_{j=1}^t \frac{\epsilon(\xi_j, x)}{N}$ for $t \in [N]$ and the filtration $\mathcal{G}_t := \sigma(y_0, \ldots, y_t)$ for $t \in \{0\} \cup [N]$. Since $\{\xi_j\}_{j=1}^N$ is an

i.i.d. sample of $\mathbf{P}$, $\{y_t, \mathcal{G}_t\}_{t=0}^N$ is a $\mathbb{R}^d$-valued martingale whose increments satisfy

$$|||y_t - y_{t-1}|||_q = \left| \frac{\|\epsilon(\xi, x)\|}{N} \right|_q \leq \frac{|||\epsilon(\xi, x_*)|||_q + L_q \|x - y\|^\delta}{N},$$

using that $|||\epsilon(\xi, \cdot)|||_q$ is Hölder continuous with modulus $L_q = |\mathsf{L}(\xi)|_q + L$ and exponent δ (Lemma 1) in the inequality. The required claim follows from the above relation, Theorem 17 and $\widehat{\epsilon}(\xi^N, x) = y_N$. We note that if $q = 2$, then the linearity of the expectation, the Pythagorean identity (valid for the Euclidean norm) and independence imply the sharper equality $\left| \left\| \widehat{\epsilon}(\xi^N, x) \right\| \right|_2 = \frac{|||\epsilon(\xi,x)|||_2}{\sqrt{N}}$. This fact and Lemma 1 imply the claim of the lemma with $C_2 = 1$. □

Proof of Theorem 12. We fix $x \in X$ and $x^* \in X^*$ as stated in the theorem and set $z^N := z(\xi^N; \alpha_N, x)$ and $\overline{z}^N := \overline{z}(\xi^N; \alpha_N, x)$. In the following, we only give a proof for $\widehat{\epsilon}(\xi^N, z^N)$. The proof for $\widehat{\epsilon}(\xi^N, \overline{z}^N)$ requires only minor changes. For reasons to be shown in the following, it will be convenient to define $\Delta(x, x^*) := \|x - x^*\| \vee \|x - x^*\|^\delta$ and, for any $s > 0$, $\mathsf{R}(s) := (1 + L\hat{\alpha})\Delta(x, x^*) + \hat{\alpha}s$ and the ball $\mathbb{B}(s) := \mathbb{B}[x^*, \mathsf{R}(s)]$.

Example 14.29 of [68] and Assumption 1 imply that the map $\Xi \times X \ni (\omega, x) \mapsto \|\widehat{\epsilon}(\xi^N(\omega), x)\|$ is a *normal integrand*, that is,

$$\omega \mapsto \operatorname{epi} \left\| \widehat{\epsilon}(\xi^N(\omega), \cdot) \right\| := \{(x, y) \in X \times \mathbb{R} : \left\| \widehat{\epsilon}(\xi^N(\omega), x) \right\| \leq y\}$$

is a set-valued measurable function. This fact and Theorem 14.37 in [68] imply further that, for any measurable function $\epsilon : \Omega \to [0, \infty)$ and $R > 0$,

$$\omega \mapsto \sup_{x' \in \mathbb{B}(\epsilon(\omega)) \cap X} \left\| \widehat{\epsilon}(\xi^N(\omega), x') \right\| \quad \text{and} \quad \omega \mapsto \sup_{x' \in \mathbb{B}[x^*, R] \cap X} \left\| \widehat{\epsilon}(\xi^N(\omega), x') \right\| \tag{5.36}$$

are measurable functions.

We first prove item (ii) for the easier case when X is compact. We set $R :=$

$\mathrm{diam}(X)$ and note that $z^N \in \mathbb{B}[x^*, R] \cap X$. This and (5.36) imply that

$$\begin{aligned}
\left|\|\widehat{\epsilon}(\xi^N, z^N)\|\right|_p &\leq \left|\sup_{x' \in \mathbb{B}[x^*,R]\cap X} \|\widehat{\epsilon}(\xi^N, x')\|\right|_p \\
&\leq \left|\sup_{x' \in \mathbb{B}[x^*,R]\cap X} \|\widehat{\epsilon}(\xi^N, x') - \widehat{\epsilon}(\xi^N, x^*)\|\right|_p + \left|\|\widehat{\epsilon}(\xi^N, x^*)\|\right|_p \\
&\leq c\left[\frac{3^\delta \sqrt{d} L_2}{\sqrt{\delta}\left(\sqrt{2}^\delta - 1\right)} + \sqrt{p} L_2 + p L_p\right] \frac{\mathrm{diam}(X)^\delta}{\sqrt{N}} + \frac{C_p |\|\epsilon(\xi, x^*)\||_p}{\sqrt{N}},
\end{aligned}$$

for some universal constant $c > 0$, where we used Lemmas 14 and 15 with $q = p$ in the last inequality. The above inequality and definition (5.2) prove item (ii).

We now prove item (i) in the case X may be unbounded. Given $\alpha \in [0, \hat{\alpha}]$, Lemma 1(iv) implies that $x^* = \Pi[x^* - \alpha T(x^*)]$. Taking into account this fact, Lemma 1(iii) and definitions of $z(\xi^N; \alpha, x)$, (5.1) and (5.3), we get that, for any $\alpha \in [0, \hat{\alpha}]$,

$$\begin{aligned}
\left\|x^* - z(\xi^N; \alpha, x)\right\| &= \left\|\Pi\left[x^* - \alpha T(x^*)\right] - \Pi\left[x - \alpha\left(T(x) + \widehat{\epsilon}(\xi^N, x)\right)\right]\right\| \\
&\leq \|x^* - x\| + \alpha \|T(x) - T(x^*)\| + \alpha \left\|\widehat{\epsilon}(\xi^N, x)\right\| \\
&\leq (1 + L\hat{\alpha})\left[\|x - x^*\| \vee \|x - x^*\|^\delta\right] + \hat{\alpha}\left\|\widehat{\epsilon}(\xi^N, x)\right\|, \qquad (5.37)
\end{aligned}$$

where, in last inequality, we used Hölder continuity of T (Lemma 1).

In the sequel we define the quantities

$$s_* := L_{2p}\Delta(x, x^*) \quad \text{and} \quad \epsilon_N := \left\|\widehat{\epsilon}(\xi^N, x)\right\|. \qquad (5.38)$$

Setting $\alpha := \alpha_N$ in (5.37), we have that[4] $z^N \in \mathbb{B}(\epsilon_N) \cap X$. We now make the following decomposition

$$\left|\left\|\widehat{\epsilon}(\xi^N, z^N)\right\|\right|_p = I_1 + I_2, \qquad (5.39)$$

using the definitions

$$I_1 := \left|\left\|\widehat{\epsilon}(\xi^N, z^N)\right\| \mathbb{I}_{\{\epsilon_N \leq s_*\}}\right|_p \quad \text{and} \quad I_2 := \left|\left\|\widehat{\epsilon}(\xi^N, z^N)\right\| \mathbb{I}_{\{\epsilon_N > s_*\}}\right|_p.$$

[4]Note that from $\alpha_N \in [0, 1]$ and convexity of X and $\mathbb{B}(\epsilon_N)$, we also have that $\overline{z}^N \in \mathbb{B}(\epsilon_N) \cap X$.

PART 1 (Upper bound on I_1): from the fact that $z^N \in \mathbb{B}(\epsilon_N) \cap X$ and (5.36), we may bound I_1 by

$$\begin{aligned} I_1 &= \left| \|\widehat{\epsilon}(\xi^N, z^N)\| \mathbb{I}_{\{\epsilon_N \leq s_*\}} \right|_p \\ &\leq \left| \sup_{x' \in \mathbb{B}(s_*) \cap X} \|\widehat{\epsilon}(\xi^N, x')\| \right|_p \\ &\leq \left| \sup_{x' \in \mathbb{B}(s_*) \cap X} \|\widehat{\epsilon}(\xi^N, x') - \widehat{\epsilon}(\xi^N, x^*)\| \right|_p + \left| \|\widehat{\epsilon}(\xi^N, x^*)\| \right|_p \\ &\leq c \left[\frac{3^\delta \sqrt{d} L_2}{\sqrt{\delta}\left(\sqrt{2}^\delta - 1\right)} + \sqrt{p} L_2 + p L_p \right] \frac{\mathsf{R}(s_*)^\delta}{\sqrt{N}} + \frac{C_p |\|\epsilon(\xi, x^*)\||_p}{\sqrt{N}}, \end{aligned}$$

where we used Lemmas 14 and 15 with $q = p$ in the last inequality. Using the fact that $\mathsf{R}(s_*) = (1 + L\hat{\alpha} + L_{2p}\hat{\alpha})\,\Delta(x, x^*)$ and setting $c_\delta := \frac{c 3^\delta}{\sqrt{\delta}(\sqrt{2}^\delta - 1)}$, we get from the above chain of inequalities that

$$I_1 \leq \left[\left(c_\delta \sqrt{d} + c\sqrt{p} \right) L_2 + cpL_p \right] C_{\mathsf{L}\hat{\alpha},p}^\delta \frac{\Delta(x, x^*)^\delta}{\sqrt{N}} + \frac{C_p |\|\epsilon(\xi, x^*)\||_p}{\sqrt{N}}, \tag{5.40}$$

with $C_{\mathsf{L}\hat{\alpha},p} := 1 + L\hat{\alpha} + L_{2p}\hat{\alpha}$.

PART 2 (Upper bound on I_2): Defining $\widehat{L}_N := N^{-1} \sum_{j=1}^N \mathsf{L}(\xi_j)$, we note that

$$\begin{aligned} \left\| \widehat{\epsilon}(\xi^N, z^N) \right\| &\leq \left\| \widehat{\epsilon}(\xi^N, z^N) - \widehat{\epsilon}(\xi^N, x^*) \right\| + \left\| \widehat{\epsilon}(\xi^N, x^*) \right\| \\ &\leq \left\| \frac{1}{N} \sum_{j=1}^N \left[F(\xi_j, z^N) - F(\xi_j, x^*) \right] \right\| + \left\| T(z^N) - T(x^*) \right\| + \left\| \widehat{\epsilon}(\xi^N, x^*) \right\| \\ &\leq \left(\widehat{L}_N + L \right) \left\| z^N - x^* \right\|^\delta + \left\| \widehat{\epsilon}(\xi^N, x^*) \right\| \\ &\leq \left(\widehat{L}_N + L \right) (1 + L\hat{\alpha}) \Delta(x, x^*) + \hat{\alpha} \left(\widehat{L}_N + L \right) \epsilon_N + \epsilon_N^*, \end{aligned}$$

using Assumption 1 and Lemma 1 in the third inequality and (5.37) with $\alpha := \alpha_N$, (5.38) and the definition $\epsilon_N^* := \left\| \widehat{\epsilon}(\xi^N, x^*) \right\|$ in the last inequality. The inequality above and definition of I_2 imply that

$$\begin{aligned} I_2 &= \left| \left\| \widehat{\epsilon}(\xi^N, z^N) \right\| \mathbb{I}_{\{\epsilon_N > s_*\}} \right|_p \\ &\leq (1 + L\hat{\alpha}) \Delta(x, x^*) \left| \left(\widehat{L}_N + L \right) \mathbb{I}_{\{\epsilon_N > s_*\}} \right|_p + \hat{\alpha} \left| \left(\widehat{L}_N + L \right) \epsilon_N \right|_p + |\epsilon_N^*|_p \\ &\leq (1 + L\hat{\alpha}) \Delta(x, x^*) \left| \widehat{L}_N + L \right|_{2p} \left| \mathbb{I}_{\{\epsilon_N > s_*\}} \right|_{2p} + \hat{\alpha} \left| \widehat{L}_N + L \right|_{2p} |\epsilon_N|_{2p} + |\epsilon_N^*|_p, \end{aligned} \tag{5.41}$$

where we used Hölder's inequality.

With respect to the last term in the rightmost expression of (5.41), we have, in view of Lemma 15 with $q = p$,

$$|\epsilon_N^*|_p = \left| \|\widehat{\epsilon}(\xi^N, x^*)\| \right|_p \leq \frac{C_p |||\epsilon(\xi, x^*)|||_p}{\sqrt{N}}. \tag{5.42}$$

Concerning the second term in the rightmost expression of (5.41), Lemma 15 with $q = 2p$ implies that

$$|\epsilon_N|_{2p} = \left| \left\| \widehat{\epsilon}(\xi^N, x) \right\| \right|_{2p} \leq C_{2p} \frac{|||\epsilon(\xi, x^*)|||_{2p} + L_{2p}\|x - x^*\|^\delta}{\sqrt{N}}. \tag{5.43}$$

From Markov's inequality and (5.43) we obtain

$$\begin{aligned} \left| \mathbb{I}_{\{\epsilon_N > s_*\}} \right|_{2p} &= \sqrt[2p]{\mathbb{E}\left[\mathbb{I}_{\{\epsilon_N > s_*\}}\right]} = \sqrt[2p]{\mathbb{P}\left(\|\widehat{\epsilon}(\xi^N, x)\| > s_*\right)} \\ &\leq \sqrt[2p]{\frac{\mathbb{E}\left[\|\widehat{\epsilon}(\xi^N, x)\|^{2p}\right]}{s_*^{2p}}} \\ &= \frac{\left\| \|\widehat{\epsilon}(\xi^N, x)\| \right\|_{2p}}{s_*} \\ &\leq C_{2p} \frac{|||\epsilon(\xi, x^*)|||_{2p} + L_{2p}\|x - x^*\|^\delta}{s_* \sqrt{N}}. \end{aligned} \tag{5.44}$$

The convexity of $t \mapsto t^{2p}$ and the fact that $\{\xi_j\}_{j \in [N]}$ is an i.i.d. sample of $\mathbf{P}$ imply that $\left| \widehat{L}_N + L \right|_{2p} \leq |\mathsf{L}(\xi)|_{2p} + L = L_{2p}$. Using this fact and putting together relations (5.41)-(5.44) we get

$$\begin{aligned} I_2 &\leq (1 + L\hat{\alpha}) \frac{\Delta(x, x^*) L_{2p}}{s_*} C_{2p} \frac{|||\epsilon(\xi, x^*)|||_{2p} + L_{2p}\|x - x^*\|^\delta}{\sqrt{N}} \\ &\quad + L_{2p}\hat{\alpha} C_{2p} \frac{|||\epsilon(\xi, x^*)|||_{2p} + L_{2p}\|x - x^*\|^\delta}{\sqrt{N}} + \frac{C_p |||\epsilon(\xi, x^*)|||_p}{\sqrt{N}} \\ &= C_{2p} \left(1 + L\hat{\alpha} + L_{2p}\hat{\alpha}\right) \frac{|||\epsilon(\xi, x^*)|||_{2p}}{\sqrt{N}} + \frac{C_p |||\epsilon(\xi, x^*)|||_p}{\sqrt{N}} \\ &\quad + C_{2p} \left(1 + L\hat{\alpha} + L_{2p}\hat{\alpha}\right) \frac{L_{2p}\|x - x^*\|^\delta}{\sqrt{N}}, \end{aligned} \tag{5.45}$$

where we used the fact that[5] $s_* = L_{2p}\Delta(x, x^*)$.

[5]Note that $[\Delta(x, x^*)\|x - x^*\|^\delta]^2 \lesssim \|x - x^*\|^{4\delta}$ with $4\delta > 2$ in the Lipschitz continuous case. The geometry of projection methods implies the derivation of a recursion in terms of $\{\|x^k - x^*\|^2\}$. It is then crucial for the convergence analysis that follows that we can choose a s_* that balances the bounds $\mathsf{R}(s_*)^\delta \lesssim \|x - x^*\|^{\beta_1}$ in I_1 and $\frac{\Delta(x, x^*)}{s_*}\|x - x^*\|^\delta \lesssim \|x - x^*\|^{\beta_2}$ in I_2 with $\beta_1, \beta_2 \in (0, 1]$.

Relations (5.39)-(5.40) and (5.45), definition (5.2) and the facts that $\Delta(x,x^*)^\delta \leq \delta_1 \vee \|x-x^*\|^\delta$ and $|||\epsilon(\xi,x^*)|||_p \leq |||\epsilon(\xi,x^*)|||_{2p}$ prove item (i). □

Remark 9 (Constants)**.** *In Theorem 12, the constants satisfy*

$$\mathsf{c}_1 := 2C_p + C_{2p}C_{\mathsf{L}\hat{\alpha},p}, \qquad \mathsf{c}_3 \lesssim pC^\delta_{\mathsf{L}\hat{\alpha},p}, \qquad \mathsf{c}_4 := C_{2p}C_{\mathsf{L}\hat{\alpha},p},$$
$$\mathsf{c}_2 \lesssim \left[\frac{3^\delta\sqrt{d}}{\sqrt{\delta}\left(\sqrt{2}^\delta - 1\right)} + \sqrt{p}\right] C^\delta_{\mathsf{L}\hat{\alpha},p}, \qquad \mathsf{d}_2 \lesssim \left[\frac{3^\delta\sqrt{d}}{\sqrt{\delta}\left(\sqrt{2}^\delta - 1\right)} + \sqrt{p}\right],$$

where $C_{\mathsf{L}\hat{\alpha},p} := 1 + 2L\hat{\alpha} + |\mathsf{L}(\xi)|_{2p}\hat{\alpha}$ *and* C_p *and* C_{2p} *are defined in Lemma 15.*

5.2 Analysis of Algorithm 5 for Lipschitz continuous operators

We state next additional assumptions needed for the convergence analysis of Algorithm 5. In this section we always assume that in Assumption 1 we have $\delta = 1$. For brevity, we will not mention it any further.

Assumption 2 (Consistency)**.** *The solution set* X^* *of VI*(T,X) *is non-empty.*

Assumption 20 (Pseudo-monotonicity)**.** *We assume that* $T : X \to \mathbb{R}^d$ *as defined in* (1.2) *is pseudo-monotone: for all* $z, x \in X$, $\langle T(x), z-x\rangle \geq 0 \implies \langle T(z), z-x\rangle \geq 0$.

Pseudo-monotonicity includes monotonicity as a special class [42, 20]. Pseudo-monotone SVIs were also considered in [44]. In these works knowledge of parameters such as the Lipschitz constant are still assumed and no variance reduction schemes are presented in [44]. Recall that the gradient of a smooth convex function is monotone and the quotient of a positive smooth convex function with a positive smooth concave function has a pseudo-monotone gradient. Recall the notation $[N_k] := \{1,\ldots,N_k\}$.

Assumption 21 (I.I.D. sampling)**.** *In Algorithm 5, the sequences* $\{\xi^k_j : k \in \mathbb{N}_0, j \in [N_k]\}$ *and* $\{\eta^k_j : k \in \mathbb{N}_0, j \in [N_k]\}$ *are i.i.d. samples drawn from* $\mathbf{P}$ *independent of each other. Moreover,* $\sum_{k=0}^\infty N_k^{-1} < \infty$.

We set $\xi^k := \{\xi_j^k\}_{j=1}^{N_k}$ and $\eta^k := \{\eta_j^k\}_{j=1}^{N_k}$. Concerning Algorithm 5, we shall study the stochastic process $\{x^k\}$ with respect to the filtrations

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

Recalling (5.1), (5.3) and Algorithm 5, we will define the following oracle errors:

$$\epsilon_1^k := \widehat{\epsilon}(\xi^k, x^k), \qquad \epsilon_2^k := \widehat{\epsilon}(\eta^k, z^k), \qquad \epsilon_3^k := \widehat{\epsilon}(\xi^k, z^k). \tag{5.46}$$

Assumption 21 implies that the processes $[N_k] \ni t \mapsto N_k^{-1} \sum_{j=1}^t \epsilon(\xi_j^k, x^k)$, $[N_k] \ni t \mapsto N_k^{-1} \sum_{j=1}^t \epsilon(\eta_j^k, z^k)$, $k \mapsto \epsilon_1^k$ and $k \mapsto \epsilon_2^k$ define martingale differences. Such property does not hold for the *correlated error* ϵ_3^k since α_k and z^k are measurable functions of ξ^k. It is also important to note that the stepsize α_k is a random variable satisfying $\alpha_k \notin \mathcal{F}_k$ and $\alpha_k \in \widehat{\mathcal{F}}_k$.

Remark 10 (Initialization of the line search rule)**.** *We make a remark regarding the exogenous parameters β and $\{\delta_k\}$ and the endogenous sequence $\{d^k\}$ defined in Algorithm 5. By the definition of d^k in* (5.4) *and convexity of X, we have that, for all $k \in \mathbb{N}$,*

$$\|d^k\| \leq \delta_k, \quad x^k + d^k \in X. \tag{5.47}$$

Moreover, it can be shown that, if $\beta \in (0, \hat{\alpha}^{-1}]$, then, for all $\alpha \in (0, \hat{\alpha}]$ and $k \in \mathbb{N}$,

$$\|z^k(\alpha) - (x^k + d^k)\| > 0, \tag{5.48}$$

where $z^k(\alpha)$ is defined in (5.6) *(see the proof of Lemma 16 in the next section). In fact, the rule* (5.4) *chosen to update d^k could be replaced by any rule satisfying* (5.47)-(5.48)*.*

The purpose of β, $\{\delta_k\}$ and d^k is solely to initialize the line search rule with a well defined direction. In deterministic regimes, this is not needed since if $r^k = 0$ (see Algorithm 5), x^k is an exact solution and we can stop the algorithm. In our framework, we use a sampled-based line search scheme so that the termination criteria is generally not clear. By choosing β, $\{\delta_k\}$ and d^k as above, the sampled-based line search rule (5.5)-(5.6) *is always clearly specified and terminates in a finitely number of iterations. The direction d^k serves merely as a small perturbation to address the case $r^k = 0$. Since $\|d^k\| \leq \delta_k$ holds for all k, we can set $\delta_k \to 0$*

in any desired rate so to correct iteratively such small perturbations. In this way, the optimality of the iteration and oracle complexities of Algorithm *5 are unaltered. We refer to the convergence analysis in the next section for further details.*

Remark 11 (Intuition for the line search scheme)**.** *The stochastic approximated line search* (5.5) *is motivated by [47]. We make some comments for the case* $d^k = 0$ *(see Remark 10). Using* (2.9), (5.5) *can be rewritten as*

$$\left\|\widehat{F}\left(\xi^k, z^k(\alpha)\right) - \widehat{F}\left(\xi^k, x^k\right)\right\| \leq \lambda \frac{r_\alpha(H_k; x^k)}{\alpha}, \tag{5.49}$$

where $H_k := \widehat{F}(\xi^k, \cdot)$*. Provided that* $r_{\hat{\alpha}}(H_k; x^k) \neq 0$*, the line search tests* (5.49) *for decreasing* $\alpha \in (0, \hat{\alpha}]$*. The idea is that the right hand side of* (5.49) *does not increase by Lemma 4 while the left hand side tends to 0 by continuity of the operator. Hence,* (5.49) *will hold eventually.*

5.2.1 Derivation of an error bound

In this section we show that Algorithm 5 is well defined and, given some $x^* \in X^*$, we derive a recursive bound for the iteration error sequence $\{\|x^k - x^*\|^2\}$.

Lemma 16 (Finite termination of the line search)**.** *Consider Assumption 1. Then the line search* (5.5) *in the iteration* k *of* Algorithm 5 *terminates after a finite number* ℓ_k *of steps.*

Proof. Set $H_k(x) := \widehat{F}(\xi^k, x - d^k) + \beta d^k$ for every $x \in X$. In particular, $\widehat{F}(\xi^k, x^k) + \beta d^k = H_k(x^k + d^k)$. Note that, from (5.6) and definition (2.9), we have that, for all $\alpha \in (0, \hat{\alpha}]$,

$$\|z^k(\alpha) - (x^k + d^k)\| = \left\|\Pi\left[x^k + d^k - \alpha H_k(x^k + d^k)\right] - (x^k + d^k)\right\| = r_\alpha(H_k; x^k + d^k).$$

We first show that $r_{\hat{\alpha}}(H_k; x^k + d^k) > 0$. From (5.4), if $\|r^k\| > 0$, we immediately have that $d^k := 0$ and $r_{\hat{\alpha}}(H_k; x^k) = \|r^k\| > 0$. If $r^k = 0$, again by (5.4), we have

that $d^k \neq 0$. Hence,

$$\begin{aligned} r_{\hat{\alpha}}(H_k; x^k + d^k) &= \left\| (x^k + d^k) - \Pi\left[x^k + d^k - \hat{\alpha} H_k(x^k + d^k)\right] - r^k \right\| \\ &= \left\| d^k + \Pi\left[x^k - \hat{\alpha}\widehat{F}(\xi^k, x^k)\right] - \Pi\left[x^k + d^k - \hat{\alpha} H_k(x^k + d^k)\right] \right\| \\ &\geq \|d^k\| - \left\| \Pi\left[x^k - \hat{\alpha}\widehat{F}(\xi^k, x^k)\right] - \Pi\left[x^k + d^k - \hat{\alpha} H_k(x^k + d^k)\right] \right\| \\ &\geq \|d^k\| - \left\| -\hat{\alpha}\widehat{F}(\xi^k, x^k) - d^k + \hat{\alpha} H_k(x^k + d^k) \right\| \\ &= \|d^k\| - \left\| (\hat{\alpha}\beta - 1) d^k \right\| = \hat{\alpha}\beta \|d^k\| > 0, \end{aligned}$$

using Lemma 1(iii) in last inequality and $0 < \hat{\alpha}\beta \leq 1$ in last equality.

We now conclude the proof of the lemma. Set $\gamma_\ell := \theta^{-\ell}\hat{\alpha}$. Assuming by contradiction that the line search (5.5) does not terminate after a finite number of iterations, for every $\ell \in \mathbb{N}_0$,

$$\left\| \widehat{F}\left(\xi^k, z^k(\gamma_\ell)\right) - \widehat{F}\left(\xi^k, x^k + d^k\right) \right\| \;>\; \lambda \frac{r_{\gamma_\ell}(H_k; x^k + d^k)}{\gamma_\ell} \geq \lambda \cdot r_{\hat{\alpha}}(H_k; x^k + d^k),$$

using definition of $r_\alpha(H_k; \cdot)$ in (2.9), the fact that $\gamma_\ell \in (0, \hat{\alpha}]$ and Lemma 4 in the last inequality. The contradiction follows by letting $\ell \to \infty$ in the above inequality and invoking the continuity of $\widehat{F}(\xi^k, \cdot)$, resulting from Assumption 1, the fact that $\lim_{\ell \to \infty} z^k(\gamma_\ell) = x^k + d^k$, which follows from the continuity of Π and $x^k + d^k \in X$, and the fact that $r_{\hat{\alpha}}(H_k; x^k + d^k) > 0$, which follows from the previous paragraph. □

The next lemma shows that the DS-SA line search scheme (5.5) either chooses the initial stepsize $\hat{\alpha}$ or it is an UO for a *lower bound* of the Lipschitz constant $L = \mathbb{E}[\mathsf{L}(\xi)]$ (using the *same samples* generated by the operator's SO): if $\hat{\alpha}$ is not chosen, then $\frac{(\lambda\theta)\wedge\hat{\alpha}}{\alpha_k}$ is a.s. a lower bound for $\widehat{L}(\xi^k) = \frac{1}{N_k}\sum_{j=1}^{N_k} \mathsf{L}(\xi_j^k)$.

Lemma 17 (Unbiased lower estimation of the Lipschitz constant)**.** *Consider Assumptions 1 and 21. Then* $\alpha_k \geq \left(\frac{\lambda\theta}{\widehat{L}(\xi^k)}\right) \wedge \hat{\alpha}$, *a.s. and* $\left|\alpha_k\right|\mathcal{F}_k\Big|_2 \cdot |\mathsf{L}(\xi)|_2 \geq (\lambda\theta) \wedge \hat{\alpha}$.

Proof. If $\hat{\alpha}$ satisfies (5.5), then $\alpha_k = \hat{\alpha}$. Otherwise, we have

$$\theta^{-1}\alpha_k \left\| \widehat{F}\left(\xi^k, z^k(\theta^{-1}\alpha_k)\right) - \widehat{F}(\xi^k, x^k + d^k) \right\| > \lambda \left\| z^k\left(\theta^{-1}\alpha_k\right) - (x^k + d^k) \right\|. \tag{5.50}$$

Assumption 1 and definition of $\widehat{F}(\xi^k, \cdot)$ in (5.3) imply that

$$\left\| \widehat{F}\left(\xi^k, z^k(\theta^{-1}\alpha_k)\right) - \widehat{F}(\xi^k, x^k + d^k) \right\| \leq \widehat{L}(\xi^k) \left\| z^k\left(\theta^{-1}\alpha_k\right) - (x^k + d^k) \right\|. \tag{5.51}$$

The fact that $z^k\left(\theta^{-1}\alpha_k\right) \neq x^k + d^k$ and (5.50)-(5.51) imply that $\alpha_k \geq \frac{\lambda\theta}{\widehat{L}(\xi^k)}$. We have thus proved the first statement.

Since a.s. $\mathsf{L}(\xi) \geq 1$, we also have a.s. $\widehat{L}(\xi^k)\alpha_k \geq (\lambda\theta)\wedge\hat{\alpha}$. The second statement follows from this fact and

$$\begin{aligned}
(\lambda\theta)\wedge\hat{\alpha} &\leq \mathbb{E}\left[\alpha_k\widehat{L}(\xi^k)\middle|\mathcal{F}_k\right] \\
\text{(by Hölder's inequality)} \qquad &\leq \left|\alpha_k\middle|\mathcal{F}_k\right|_2 \cdot \left|\widehat{L}(\xi^k)\middle|\mathcal{F}_k\right|_2 \\
\text{(by convexity of } t\mapsto t^2) \qquad &\leq \left|\alpha_k\middle|\mathcal{F}_k\right|_2 \sqrt{\frac{1}{N_k}\sum_{j=1}^{N_k}\mathbb{E}\left[\mathsf{L}(\xi_j^k)^2\middle|\mathcal{F}_k\right]} = \left|\alpha_k\middle|\mathcal{F}_k\right|_2 \cdot |\mathsf{L}(\xi)|_2,
\end{aligned}$$

using Assumption 21 in last equality. □

Recall (5.3) and (5.46). We define, for $k \in \mathbb{N}_0$ and for $x^* \in X^*$,

$$\Delta A_k := (1-8\lambda^2)\hat{\alpha}^2\|\epsilon_1^k\|^2 + 8\hat{\alpha}^2\|\epsilon_2^k\|^2 + 8\hat{\alpha}^2\|\epsilon_3^k\|^2, \tag{5.52}$$

$$\Delta M_k(x^*) := 2\alpha_k\langle x^* - z^k, \epsilon_2^k\rangle, \tag{5.53}$$

$$\Delta P_k := 8(2-\alpha_k\beta)^2[\lambda + \alpha_k\widehat{L}(\xi^k)]^2\delta_k^2. \tag{5.54}$$

We recall the reader to the definition $r := r_1(T;\cdot)$ in (2.9).

Lemma 18 (A recursive error bound for Algorithm 5)**.** *Consider Assumptions 1 and 2-20. The sequence generated by* Algorithm 5 *satisfies, for all $x^* \in X^*$ and $k \in \mathbb{N}_0$,*

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \frac{(1-8\lambda^2)\alpha_k^2}{2}r^2(x^k) + \Delta M_k(x^*) + \Delta A_k + \Delta P_k.$$

Proof of Lemma 18. We divide the proof in two parts. The first uses the extragradient step (5.7)-(5.8). The second uses the line search (5.5)-(5.6) with some judicious error bounds.

PART 1 (Extragradient step): by (5.7)-(5.8), we invoke twice Lemma 1(i) with $v := \alpha_k\widehat{F}(\xi^k, x^k)$, $x := x^k$ and $z := z^k$ and with $v := \alpha_k\widehat{F}(\eta^k, z^k)$, $x := x^k$ and $z := x^{k+1}$, obtaining, for all $x \in X$,

$$2\langle\alpha_k\widehat{F}(\xi^k, x^k), z^k - x\rangle \leq \|x^k - x\|^2 - \|z^k - x\|^2 - \|z^k - x^k\|^2, \tag{5.55}$$

$$2\langle\alpha_k\widehat{F}(\eta^k, z^k), x^{k+1} - x\rangle \leq \|x^k - x\|^2 - \|x^{k+1} - x\|^2 - \|x^{k+1} - x^k\|^2. \tag{5.56}$$

We now set $x := x^{k+1}$ in (5.55) and sum the obtained relation with (5.56) eliminating $\|x^k - x^{k+1}\|^2$. We thus get, for all $x \in X$,

$$\begin{aligned} \mathsf{I} &:= 2\langle \alpha_k \widehat{F}(\xi^k, x^k), z^k - x^{k+1}\rangle + 2\langle \alpha_k \widehat{F}(\eta^k, z^k), x^{k+1} - x\rangle \\ &\leq \|x^k - x\|^2 - \|x^{k+1} - x\|^2 - \|z^k - x^{k+1}\|^2 - \|z^k - x^k\|^2. \end{aligned}$$

Using definitions (5.1), (5.3) and (5.46), we have

$$\begin{aligned} \mathsf{I} &= 2\alpha_k \langle \widehat{F}(\xi^k, x^k) - \widehat{F}(\eta^k, z^k), z^k - x^{k+1}\rangle + 2\langle \alpha_k \widehat{F}(\eta^k, z^k), z^k - x\rangle \\ &= 2\alpha_k \langle \widehat{F}(\xi^k, x^k) - \widehat{F}(\eta^k, z^k), z^k - x^{k+1}\rangle + 2\alpha_k \langle T(z^k), z^k - x\rangle + 2\alpha_k \langle \epsilon_2^k, z^k - x\rangle. \end{aligned}$$

The two previous relations imply that, for all $\in X$,

$$\begin{aligned} 2\alpha_k \langle T(z^k), z^k - x\rangle &\leq 2\alpha_k \langle \widehat{F}(\eta^k, z^k) - \widehat{F}(\xi^k, x^k), z^k - x^{k+1}\rangle + 2\alpha_k \langle \epsilon_2^k, x - z^k\rangle \\ &\quad + \|x^k - x\|^2 - \|x^{k+1} - x\|^2 - \|z^k - x^{k+1}\|^2 - \|z^k - x^k\|^2 \\ &\leq 2\alpha_k \|\widehat{F}(\eta^k, z^k) - \widehat{F}(\xi^k, x^k)\| \|z^k - x^{k+1}\| + 2\alpha_k \langle \epsilon_2^k, x - z^k\rangle \\ &\quad + \|x^k - x\|^2 - \|x^{k+1} - x\|^2 - \|z^k - x^{k+1}\|^2 - \|z^k - x^k\|^2 \\ &\leq 2\alpha_k^2 \|\widehat{F}(\eta^k, z^k) - \widehat{F}(\xi^k, x^k)\|^2 + 2\alpha_k \langle \epsilon_2^k, x - z^k\rangle \\ &\quad + \|x^k - x\|^2 - \|x^{k+1} - x\|^2 - \|z^k - x^k\|^2, \end{aligned} \tag{5.57}$$

where we used Cauchy-Schwartz in second inequality and Lemma 1(iii) with (5.7)-(5.8) in the third inequality.

PART 2 (Line search rule): For simplicity, we set $\tilde{z}^k := z^k(\alpha_k)$ and $\tilde{x}^k := x^k + d^k$ as defined in (5.6). We first note that, by (5.6)-(5.7), Lemma 1(iii), $0 < \alpha_k \beta \leq \hat{\alpha}\beta \leq 1$ and $\|d^k\| \leq \delta_k$,

$$\|\tilde{z}^k - z^k\| \leq \|d^k - \alpha_k \beta d^k\| \leq (1 - \alpha_k \beta)\delta_k, \quad \|\tilde{x}^k - x^k\| \leq \delta_k. \tag{5.58}$$

Recall that, according to (5.3), $\widehat{L}(\xi^k) = \frac{1}{N_k} \sum_{j=1}^{N_k} \mathsf{L}(\xi_j^k)$. Concerning the first term in the rightmost expression in (5.57), we have by the triangle inequality,

$$\begin{aligned} \alpha_k \|\widehat{F}(\eta^k, z^k) - \widehat{F}(\xi^k, x^k)\| &\leq \alpha_k \|\widehat{F}(\eta^k, z^k) - \widehat{F}(\xi^k, \tilde{z}^k)\| + \alpha_k \|\widehat{F}(\xi^k, \tilde{z}^k) - \widehat{F}(\xi^k, \tilde{x}^k)\| \\ &\quad + \alpha_k \|\widehat{F}(\xi^k, \tilde{x}^k) - \widehat{F}(\xi^k, x^k)\|. \end{aligned} \tag{5.59}$$

The first term above can be bounded as

$$\begin{aligned}\alpha_k\|\widehat{F}(\eta^k,z^k)-\widehat{F}(\xi^k,\tilde{z}^k)\| &\leq \alpha_k\|\widehat{F}(\eta^k,z^k)-T(z^k)\|+\alpha_k\|\widehat{F}(\xi^k,z^k)-T(z^k)\| \\ &\quad +\alpha_k\|\widehat{F}(\xi^k,z^k)-\widehat{F}(\xi^k,\tilde{z}^k)\| \\ &\leq \alpha_k\|\epsilon_2^k\|+\alpha_k\|\epsilon_3^k\|+\alpha_k\widehat{L}(\xi^k)\|z^k-\tilde{z}^k\| \\ &\leq \alpha_k\|\epsilon_2^k\|+\alpha_k\|\epsilon_3^k\|+\alpha_k\widehat{L}(\xi^k)(1-\alpha_k\beta)\delta_k, \end{aligned} \tag{5.60}$$

using the triangle inequality in first inequality, Assumption 1 and definitions in (5.1), (5.3) and (5.46) in second inequality and (5.58) in the last inequality. Similarly, the third term in (5.59) satisfies

$$\alpha_k\|\widehat{F}(\xi^k,\tilde{x}^k)-\widehat{F}(\xi^k,x^k)\|\leq\alpha_k\widehat{L}(\xi^k)\|\tilde{x}^k-x^k\|\leq\alpha_k\widehat{L}(\xi^k)\delta_k. \tag{5.61}$$

Finally, from the line search (5.5)-(5.6) and (5.58), the second term in (5.59) satisfies

$$\begin{aligned}\alpha_k\|\widehat{F}(\xi^k,\tilde{z}^k)-\widehat{F}(\xi^k,\tilde{x}^k)\| &\leq \lambda\|\tilde{z}^k-\tilde{x}^k\| \\ &\leq \lambda\|\tilde{z}^k-z^k\|+\lambda\|z^k-x^k\|+\lambda\|x^k-\tilde{x}^k\| \\ &\leq \lambda\|z^k-x^k\|+\lambda(2-\alpha_k\beta)\delta_k. \end{aligned} \tag{5.62}$$

Putting together (5.59)-(5.62), squaring, using the fact that $(\sum_{i=1}^4 a_i)^2\leq 4\sum_{i=1}^4 a_i^2$ and using definition (5.54), we obtain

$$2\alpha_k^2\|\widehat{F}(\eta^k,z^k)-\widehat{F}(\xi^k,x^k)\|^2\leq 8\lambda^2\|z^k-x^k\|^2+8\hat{\alpha}^2(\|\epsilon_2^k\|^2+\|\epsilon_3^k\|^2)+\Delta P_k. \tag{5.63}$$

From $z^k=\Pi[x^k-\alpha_k(T(x^k)+\epsilon_1^k)]$ and Lemma 4 with $\alpha_k\in(0,1]$, we also have

$$\begin{aligned}\alpha_k^2 r^2(x^k) &\leq r_{\alpha_k}^2(x^k) \\ &= \|x^k-\Pi[x^k-\alpha_k T(x^k)]\|^2 \\ &\leq 2\|x^k-z^k\|^2+2\|\Pi[x^k-\alpha_k(T(x^k)+\epsilon_1^k)]-\Pi[x^k-\alpha_k T(x^k)]\|^2 \\ &\leq 2\|x^k-z^k\|^2+2\hat{\alpha}^2\|\epsilon_1^k\|^2, \end{aligned} \tag{5.64}$$

where we used Lemma 1(iii) in the second inequality. The claim is proved using relations (5.57) and (5.63)-(5.64) with $x:=x^*$, for a given $x^*\in X^*$, definitions (5.52)-(5.53) and the facts that $0<1-8\lambda^2<1$ (see Algorithm 5) and $\langle T(z^k),z^k-x^*\rangle\geq 0$, which follows from $\langle T(x^*),z^k-x^*\rangle\geq 0$ (since $x^*\in X^*$) and Assumption 20. □

5.2.2 Bound on oracle error

This section is devoted to the control of the oracle errors in (5.52)-(5.53). Since $\{\epsilon_1^k\}$, $\{\epsilon_2^k\}$ and $\{\Delta M_k(x^*)\}$ define martingale difference sequences, their control is simpler and uses Lemma 15.

As mentioned in the introduction, one of the significant challenges in analyzing our sampled-based line search scheme is to control the correlated error $\|\epsilon_3^k\|^2$ in (5.52). Indeed, $z^k = z(\xi^k; \alpha_k, x^k)$ is a function of the sample ξ^k so that $\epsilon_3^k = \widehat{\epsilon}(\xi^k, z^k)$ is not a martingale. In order to bound it, we will use Theorem 12. We are now ready to obtain the following result.

Proposition 11 (Bound on oracle error)**.** *Consider Assumptions 1, 2 and 21. Recall definitions in* (5.1), (5.52)*, Lemma 15 and Theorem 12. Then there exist positive constants* C_p *and* $\overline{\mathsf{C}}_p$ *(depending only on d, p,* $\mathsf{L}(\xi)\hat{\alpha}$ *and* $\{N_k\}$*) such that, for all* $x^* \in X^*$*,*

$$|\Delta A_k|\mathcal{F}_k|_{\frac{p}{2}} \quad \leq \quad \frac{\mathsf{C}_p \left[\hat{\alpha}\sigma_{\mathsf{a}p}(x^*)\right]^2 + \overline{\mathsf{C}}_p \left(\hat{\alpha}\widetilde{L}_p\right)^2 D_k^2}{N_k}.$$

In above, for X compact, we have $\mathsf{a} = 1$*,* $\widetilde{L}_p := (C_p L_p) \vee L_p^*$ *and* $D_k :\equiv \mathrm{diam}(X)$*. For a general X, we have* $\mathsf{a} = 2$*,* $\widetilde{L}_p := \overline{L}_{2p}$ *and* $D_k := \|x^k - x^*\|$*.*

Proof of Proposition 11. First, we obtain a bound on $\|z^k - x^*\|$. Recall that $z^k = \Pi[x^k - \alpha_k(T(x^k) + \epsilon_1^k)]$, $x^* = \Pi[x^* - \alpha_k T(x^*)]$ (Lemma 1(iv)), $\epsilon_1^k = \widehat{\epsilon}(\xi^k, x^k)$ and $x^k \in \mathcal{F}_k$. From these facts, Lemma 1(ii) and Lipschitz continuity of T, we obtain

$$\left|\|z^k - x^*\||\mathcal{F}_k\right|_p \leq (1 + L\hat{\alpha})\|x^k - x^*\| + \hat{\alpha}\left|\|\epsilon_1^k\||\mathcal{F}_k\right|_p. \tag{5.65}$$

Lemma 15 with $q = p$, (5.46) and the facts that $x^k \in \mathcal{F}_k$ and $\xi^k \perp\!\!\!\perp \mathcal{F}_k$ imply that

$$\left|\|\epsilon_1^k\||\mathcal{F}_k\right|_p \leq C_p \frac{\sigma_p(x^*) + L_p\|x^k - x^*\|}{\sqrt{N_k}}. \tag{5.66}$$

Lemma 15 with $q = p$, (5.46) and the facts that $z^k \in \widehat{\mathcal{F}}_k$, $\eta^k \perp\!\!\!\perp \widehat{\mathcal{F}}_k$ and $\left|\left|\cdot|\widehat{\mathcal{F}}_k\right|_p|\mathcal{F}_k\right|_p = |\cdot|\mathcal{F}_k|_p$ imply that

$$\left|\|\epsilon_2^k\||\mathcal{F}_k\right|_p = \left|\left|\|\epsilon_2^k\|\left|\widehat{\mathcal{F}}_k\right|_p\right|\mathcal{F}_k\right|_p \leq C_p \frac{\sigma_p(x^*) + L_p\left|\|z^k - x^*\||\mathcal{F}_k\right|_p}{\sqrt{N_k}}. \tag{5.67}$$

Finally, Theorem 12(i), (5.46), Assumption 21, $0 < \alpha_k \leq \hat{\alpha} \leq 1$ and the facts that $z^k = z(\xi^k; \alpha_k, x^k)$, $x^k \in \mathcal{F}_k$ and $\xi^k \perp\!\!\!\perp \mathcal{F}_k$ imply that

$$\left\| \|\epsilon_3^k\| | \mathcal{F}_k \right|_p = \left\| \widehat{\epsilon}\left(\xi^k, z(\xi^k; \alpha_k, x^k)\right) \right\| \left| \mathcal{F}_k \right|_p \leq \frac{\mathsf{c}_1 \sigma_{2p}(x^*) + \overline{L}_{2p} \|x^k - x^*\|}{\sqrt{N_k}}. \tag{5.68}$$

The required claim is proved by putting together relations (5.52), (5.65)-(5.68) and using the facts that $|a^2|\mathcal{F}_k|_{\frac{p}{2}} = |a|\mathcal{F}_k|_p^2$, $(a+b)^2 \leq 2a^2 + 2b^2$, $\overline{L}_{2p} > L_p C_p$, $\mathsf{c}_1 > C_p$ (as defined in Assumption 1, Lemma 15, Theorem 12 and Remark 9) and $\sigma_{2p}(x^*) \geq \sigma_p(x^*)$. The proof for the case X is *compact* is analogous but replacing (5.65) by the facts that $\|x^k - x^*\| \leq \mathrm{diam}(X)$ and $\|z^k - x^*\| \leq \mathrm{diam}(X)$ and replacing (5.68) by the bound of Theorem 12(ii). □

Remark 12 (Constants of Proposition 11)**.** *Recall definitions in Assumption 1, Algorithm 5, Lemma 15, Theorem 12 and Remark 9. Let* $\mathsf{G}_p := \sup_k \frac{C_p L_p \hat{\alpha}}{\sqrt{N_k}}$. *The constants in Proposition 11 are given, for a general* X*, by*

$$\mathsf{C}_p := 2\mathsf{c}_1^2 \left[8\left(1 + \mathsf{G}_p\right)^2 + 9 - 8\lambda^2\right], \qquad \overline{\mathsf{C}}_p := 2\left[8\left(1 + L\hat{\alpha} + \mathsf{G}_p\right)^2 + 9 - 8\lambda^2\right].$$

For a compact X*, the constants are* $\mathsf{C}_p := (34 - 16\lambda^2)C_p^2$ *and* $\overline{\mathsf{C}}_p := 34 - 16\lambda^2$.

5.2.3 Asymptotic convergence, convergence rate and oracle complexity

In this section, we establish the asymptotic convergence of Algorithm 5 and give bounds on its iteration and oracle complexities. In the following, we set $p = 2$ (see Remark 15 for the interest in $p > 2$).

Proposition 12 (Stochastic quasi-Fejér property)**.** *Consider Assumptions 1 and 2-21 and definitions in Proposition 11 with* $p = 2$*. Set* $\nu := \frac{(1-8\lambda^2)[(\lambda\theta)\wedge\hat{\alpha}]^2}{2|\mathsf{L}(\xi)|_2^2}$ *and* $\mathsf{C}_0 := 64\lambda^2 + 64\hat{\alpha}^2\mathbb{E}[\mathsf{L}(\xi)^2]$*. The sequence generated by* Algorithm 5 *satisfies, for all* $x^* \in X^*$ *and* $k \in \mathbb{N}_0$*,*

$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2 | \mathcal{F}_k\right] \;\leq\; \|x^k - x^*\|^2 - \nu r^2(x^k) + \frac{\mathsf{C}_2\left[\hat{\alpha}\sigma_{2\mathsf{a}}(x^*)\right]^2 + \overline{\mathsf{C}}_2(\hat{\alpha}\widetilde{L}_2)^2 D_k^2}{N_k} + \mathsf{C}_0\delta_k^2.$$

Proof. We first show that $\{\Delta M_k(x^*), \mathcal{F}_k\}$ defines a martingale difference even if $\alpha_k \notin \mathcal{F}_k$. Indeed, the facts that $z^k \in \widehat{\mathcal{F}}_k$ and $\eta^k \perp\!\!\!\perp \widehat{\mathcal{F}}_k$ imply that $\mathbb{E}[\epsilon_2^k | \widehat{\mathcal{F}}_k] =$

0, where ϵ_2^k is defined in (5.46). This fact, $z^k \in \widehat{\mathcal{F}}_k$ and $\alpha_k \in \widehat{\mathcal{F}}_k$ imply that $\mathbb{E}[\Delta M_k(x^*)|\widehat{\mathcal{F}}_k] = 0$ and, hence, $\mathbb{E}[\Delta M_k(x^*)|\mathcal{F}_k] = \mathbb{E}[\mathbb{E}[\Delta M_k(x^*)|\widehat{\mathcal{F}}_k]|\mathcal{F}_k] = 0$ as claimed.

From definition (5.54), $\alpha_k \leq \hat{\alpha}$ and the fact that $(\sum_{i=1}^2 a_i)^2 \leq 2\sum_{i=1}^2 a_i^2$, it follows that $\Delta P_k \leq 32(2\lambda^2 + 2\hat{\alpha}^2\widehat{L}(\xi^k)^2)\delta_k^2$. This fact and $\mathbb{E}[\widehat{L}(\xi^k)^2|\mathcal{F}_k] = \mathbb{E}[\mathsf{L}(\xi)^2]$ imply that $\mathbb{E}[\Delta P_k|\mathcal{F}_k] \leq \mathsf{C}_0\delta_k^2$.

After we take $\mathbb{E}[\cdot|\mathcal{F}_k]$ in Lemma 18, the recursion follows from the facts that $\{\Delta M_k(x^*), \mathcal{F}_k\}$ is a martingale difference and $\mathbb{E}[\Delta P_k|\mathcal{F}_k] \leq \mathsf{C}_0\delta_k^2$, the facts that $\mathbb{E}\left[\alpha_k^2\middle|\mathcal{F}_k\right] \geq \frac{[(\lambda\theta)\wedge\hat{\alpha}]^2}{|\mathsf{L}(\xi)|_2^2}$ (Lemma 17) and $x^k \in \mathcal{F}_k$ and Proposition 11 with $p = 2$. □

Theorem 1 (Asymptotic convergence)**.** *Consider Assumptions 1 and 2-21. Suppose that* $\sum_k \delta_k^2 < \infty$. *Then* Algorithm 5 *generates an infinite sequence* $\{x^k\}$ *such that a.s. it is bounded,* $\lim_{k\to\infty} \mathrm{d}(x^k, X^*) = 0$, *and* $r(x^k) \to 0$ *a.s. and in* $\mathcal{L}^2$. *In particular, a.s. every cluster point of* $\{x^k\}$ *belongs to* X^*.

Proof. Take some $x^* \in X^*$. Taking into account $\sum_k N_k^{-1} < \infty$ and $\sum_k \delta_k^2 < \infty$, Proposition 12 for a general X ($\mathsf{a} := 2$) and the fact that $x^k \in \mathcal{F}_k$, we apply Theorem 1 with $y_k := \|x^k - x^*\|^2$, $a_k := \frac{\overline{\mathsf{C}}_2(\hat{\alpha}\overline{L}_4)^2}{N_k}$, $b_k := \frac{\mathsf{C}_2[\hat{\alpha}\sigma_4(x^*)]^2}{N_k} + \mathsf{C}_0\delta_k^2$ and $u_k := \nu r^2(x^k)$, in order to conclude that a.s. $\{\|x^k - x^*\|^2\}$ converges and $\sum_{k=0}^\infty r^2(x^k) < \infty$. In particular, a.s. $\{x^k\}$ is bounded and $0 = \lim_{k\to\infty} r^2(x^k) = \lim_{k\to\infty} \left\|x^k - \Pi\left[x^k - T(x^k)\right]\right\|^2$. This fact and the continuity of T (Lemma 1) and Π (Lemma 1(ii)) 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 Lemma 1(iv), we conclude that $\bar{x} \in X^*$. On an event of probability 1, the boundedness of $\{x^k\}$ and the fact that every cluster point of $\{x^k\}$ belongs to X^* imply that $\lim_{k\to\infty} \mathrm{d}(x^k, X^*) = 0$. The fact that $\lim_{k\to\infty} \mathbb{E}[r^2(x^k)] = 0$ is proved in a similar way, taking expectation in the recursion of Proposition 12. □

Under lack of boundedness of X or $\sigma_2(\cdot)$ we cannot infer a priori the boundedness of the sequence $\left\{\left|\|x^k\|\right|_2\right\}$. This is obtained next and later used to obtain an iteration complexity and oracle complexity in terms of local parameters.

Proposition 13 ($\mathcal{L}^2$-boundedness of the iterates: unbounded case)**.** *Let Assumptions 1 and 2-21 hold and suppose that* $\sum_k \delta_k^2 < \infty$. *Recall definitions in* Algorithm 5, (5.1), *Theorem 12 and Propositions 11-12 with* $p = 2$. *Let* $x^* \in X^*$ *and choose*

$k_0 := k_0(\overline{\mathsf{C}}_2, \hat{\alpha}\overline{L}_4, \mathsf{C}_0) \in \mathbb{N}$ *and* $\phi \in (0,1)$ *such that*

$$\sum_{i=k_0}^{\infty} \frac{1}{N_i} \leq \frac{\phi}{\overline{\mathsf{C}}_2 \left(\hat{\alpha}\overline{L}_4\right)^2} \qquad \textit{and} \qquad \sum_{i=k_0}^{\infty} \delta_i^2 \leq \frac{1}{\mathsf{C}_0}. \tag{5.69}$$

Then $\sup_{k\geq k_0} \left|\left\|x^k - x^*\right\|\right|_2^2 \leq \frac{\left|\left\|x^{k_0} - x^*\right\|\right|_2^2 + \frac{\phi \mathsf{C}_2 \sigma_4(x^*)^2}{\overline{\mathsf{C}}_2 \overline{L}_4^2} + 1}{1-\phi}$.

Proof. In the following, we set $d_i := \|x^i - x^*\|^2$ for $i \in \mathbb{N}_0$. Let $k > k_0$ in $\mathbb{N}_0$ with k_0 as stated in (5.69). Note that such k_0 always exists since $\sum_k N_k^{-1} < \infty$ by Assumption 21. Consider the recursion of Proposition 12 for the case X is unbounded ($\mathsf{a} := 2$). We take expectation, use $\mathbb{E}[\mathbb{E}[\cdot|\mathcal{F}_i]] = \mathbb{E}[\cdot]$ and drop the negative term in the right hand side. We then sum recursively the obtained inequality from $i := k_0$ to $i := k-1$, obtaining

$$|d_k|_2^2 \leq |d_{k_0}|_2^2 + \overline{\mathsf{C}}_2 \left(\hat{\alpha}\overline{L}_4\right)^2 \sum_{i=k_0}^{k-1} \frac{|d_i|_2^2}{N_i} + \mathsf{C}_2 \left[\hat{\alpha}\sigma_4(x^*)\right]^2 \sum_{i=k_0}^{k-1} \frac{1}{N_i} + \mathsf{C}_0 \sum_{i=k_0}^{k-1} \delta_i^2. \tag{5.70}$$

For any $a > 0$, we define the stopping time $\tau_a := \inf\{k \geq k_0 : |d_k|_2 > a\}$. From (5.69)-(5.70) and definition of τ_a, we have that, for any $a > 0$ such that $\tau_a < \infty$,

$$\begin{aligned} a^2 < |d_{\tau_a}|_2^2 &\leq |d_{k_0}|_2^2 + \overline{\mathsf{C}}_2 \left(\hat{\alpha}\overline{L}_4\right)^2 \sum_{i=k_0}^{\tau_a - 1} \frac{|d_i|_2^2}{N_i} + \mathsf{C}_2 \left[\hat{\alpha}\sigma_4(x^*)\right]^2 \sum_{i=k_0}^{\tau_a - 1} \frac{1}{N_i} + \mathsf{C}_0 \sum_{i=k_0}^{\tau_a - 1} \delta_i^2 \\ &\leq |d_{k_0}|_2^2 + \phi a^2 + \frac{\phi \mathsf{C}_2 \sigma_4(x^*)^2}{\overline{\mathsf{C}}_2 \overline{L}_4^2} + 1, \end{aligned}$$

and hence, $a^2 < \frac{\left|d_{k_0}\right|_2^2 + \frac{\phi \mathsf{C}_2 \sigma_4(x^*)^2}{\overline{\mathsf{C}}_2 \overline{L}_4^2} + 1}{1-\phi} =: B$, where we used that $\phi \in (0,1)$. By definition of τ_a for any $a > 0$, the argument above implies that any threshold a^2 which the sequence $\{|d_k|_2^2\}_{k\geq k_0}$ eventually exceeds is bounded above by B. Hence $\{|d_k|_2^2\}_{k\geq k_0}$ is bounded and it satisfies the statement of the proposition. □

Theorem 18 (Rate of convergence)**.** *Consider Assumptions 1 and 2-21. Take any positive sequence* $\{\delta_k\}$ *such that* $\Delta := \sum_k \delta_k^2 < \infty$. *Recall definitions in* Algorithm 5, (5.1) *and Propositions 11-12 with* $p = 2$. *Set*

$$N_k := N \left\lceil (k+\mu)(\ln(k+\mu))^{1+b} \right\rceil, \tag{5.71}$$

for any $N \in \mathbb{N}$, $b > 0$ *and* $\mu > 2$. *Then Theorem 1 holds and the sequence* $\{x^k\}$ *generated by* Algorithm 5 *is bounded in* $\mathcal{L}^2$. *Moreover, for any* $x^* \in X^*$, *if* $\mathsf{J} > 0$ *is*

such that $\sup_{k\geq 0}\left|\|x^k - x^*\|\right|_2^2 \leq \mathsf{J}$, *the following bound holds for all* $k \in \mathbb{N}_0$*:*

$$\min_{i\in\{0,\dots,k\}} \mathbb{E}[r^2(x^i)] \leq \frac{\nu^{-1}}{k+1}\left\{\|x^0 - x^*\|^2 + \frac{\mathsf{C}_2[\hat{\alpha}\sigma_{2\mathsf{a}}(x^*)]^2 + \overline{\mathsf{C}}_2\left(\hat{\alpha}\widetilde{L}_2\right)^2\mathsf{J}}{Nb[\ln(\mu-1)]^b} + \mathsf{C}_0\Delta\right\}.$$

Proof. Clearly, $\{N_k\}$ satisfies Assumption 21 and $\sum_k \delta_k^2 < \infty$. Hence, Theorem 1 and Proposition 13 hold. In particular, $\{x^k\}$ is bounded in $\mathcal{L}^2$. Let $x^* \in X^*$ and J as stated in the theorem. Hence, $\sup_k \mathbb{E}[D_k^2] \leq \mathsf{J}$. In the recursion of Proposition 12, we take expectation, use $\mathbb{E}[\mathbb{E}[\cdot|\mathcal{F}_i]] = \mathbb{E}[\cdot]$ and sum recursively the obtained inequality from $i := 0$ to $i := k$. We then obtain

$$\nu\sum_{i=0}^{k}\mathbb{E}\left[r^2(x^i)\right] \leq \|x^0 - x^*\|^2 + \left\{\mathsf{C}_2[\hat{\alpha}\sigma_{2\mathsf{a}}(x^*)]^2 + \overline{\mathsf{C}}_2\left(\hat{\alpha}\widetilde{L}_2\right)^2\mathsf{J}\right\}\mathsf{S}_k + \mathsf{C}_0\Delta,$$

where $\mathsf{S}_k := \sum_{i=0}^k N_i^{-1}$. The proof of the statement follows from the above inequality, the bound

$$\mathsf{S}_k \leq \sum_{i=0}^{\infty}\frac{1}{N_i} \leq \int_{-1}^{\infty}\frac{\mathrm{d}\,t}{N(t+\mu)[\ln(t+\mu)]^{1+b}} = \frac{1}{Nb[\ln(\mu-1)]^b},$$

and $\min_{i\in\{0,\dots,k\}}\mathbb{E}\left[r^2(x^i)\right] \leq \frac{1}{k+1}\sum_{i=0}^k \mathbb{E}\left[r^2(x^i)\right]$. □

We end this section with an estimate on the iteration and oracle complexities. Unlike SA methods with endogenous stepsizes, the number of oracle calls in Algorithm 5 is a *random variable*. In order to compute the first operator step (5.7) of iteration k, the oracle is called $\ell_k N_k$ times using a sampled-based line search (which terminates in ℓ_k random iterations).[6] For the second operator step (5.8) of iteration k, the oracle is called N_k times. We thus present two types of oracle complexities for which $\min_{i\in\{0,\dots,K\}}\mathbb{E}[r^2(x^i)] \leq \epsilon$ after Algorithm 5 is run K times. The first is an upper bound of $\sum_{i=0}^K(1+\ell_i)N_i$ *with probability 1*. This bound will depend on the logarithm of the *largest empirical mean Lipschitz constant* of previous iterations. The second result is an upper bound on the *mean oracle complexity* $\sum_{i=0}^K(1+\mathbb{E}[\ell_i])N_i$. This will depend on the logarithm of the *mean Lipschitz constant* L.

[6] During one step of the line search testing a stepsize α, we count all N_k oracle calls $\{F(\xi_j^k, z^k(\alpha))\}_{j=1}^{N_k}$ used to compute step (5.5). In all such ℓ_k steps, the same sample $\xi^k = \{\xi_j^k\}_{j=1}^{N_k}$ is used.

Corollary 5 (Iteration and oracle complexities)**.** *Let the assumptions of Theorem 18 hold and set $N := \mathcal{O}(d)$. Given $\epsilon > 0$,* Algorithm 5 *achieves the tolerance*

$$\min_{i\in\{0,\dots,K\}} \mathbb{E}[r^2(x^i)] \leq \epsilon, \tag{5.72}$$

after $K = b^{-1}\mathcal{O}(\epsilon^{-1})$ iterations.

Additionally, with probability 1, (5.72) *is ensured with an oracle complexity $\sum_{i=0}^{K}(1+\ell_i)N_i$ upper bounded by*

$$b^{-2}\cdot \log_{\frac{1}{\theta}}\left(\frac{\hat{\alpha}\max_{i\in\{0,\dots,K\}}\widehat{L}(\xi^i)}{(\lambda\theta)\wedge\hat{\alpha}}\right)\cdot\left[\ln\left(b^{-1}\epsilon^{-1}\right)\right]^{1+b}\cdot\mathcal{O}(d\epsilon^{-2}),$$

where ℓ_k is the number of oracle calls used in the line search scheme (5.5) *at iteration k and $\widehat{L}(\xi^k) = \frac{1}{N_k}\sum_{j=1}^{N_k}\mathsf{L}(\xi_j^i)$.*

Moreover, (5.72) *is ensured with a mean oracle complexity $\sum_{i=0}^{K}(1+\mathbb{E}[\ell_i])N_i$ upper bounded by*

$$b^{-2}\cdot \log_{\frac{1}{\theta}}\left(\frac{\hat{\alpha}L}{(\lambda\theta)\wedge\hat{\alpha}}\right)\cdot\left[\ln\left(b^{-1}\epsilon^{-1}\right)\right]^{1+b}\cdot\mathcal{O}(d\epsilon^{-2}).$$

Proof. We recall the definitions in Assumption 1, Lemma 15, Theorem 12, Remark 9, Propositions 11 and 12 and Remark 12 with $p = 2$. The definitions of $\widetilde{L}_2$, $\overline{L}_4$, L_2^*, c_2 and d_2 (which depend on d) and Theorem 18 imply that, up to a constant $B > 0$, for every $k \in \mathbb{N}$, $\min_{i\in\{0,\dots,k\}}\mathbb{E}[r(x^i)^2] \leq Bd(Nbk)^{-1}$. Hence, given $\epsilon > 0$, we obtain $\min_{i\in\{0,\dots,K\}}\mathbb{E}[r^2(x^i)] \leq \epsilon$ after $K = \mathcal{O}(dN^{-1}b^{-1}\epsilon^{-1})$ iterations.

The total number of oracle calls after K iterations is upper bounded by

$$\begin{aligned}\sum_{i=0}^{K}(1+\ell_i)N_i &\lesssim \left(\max_{i\in\{0,\dots,K\}}\ell_i\right)\sum_{i=1}^{K}Ni(\ln i)^{1+b} \lesssim \left(\max_{i\in\{0,\dots,K\}}\ell_i\right)NK^2(\ln K)^{1+b}\\ &\lesssim \left(\max_{i\in\{0,\dots,K\}}\ell_i\right)N^{-1}d^2b^{-2}\epsilon^{-2}\left[\ln\left(dN^{-1}b^{-1}\epsilon^{-1}\right)\right]^{1+b}. \qquad (5.73)\end{aligned}$$

Moreover, Lemma 17 implies that $\ell_k \leq \log_{\frac{1}{\theta}}\left(\frac{\hat{\alpha}\widehat{L}(\xi^k)}{(\lambda\theta)\wedge\hat{\alpha}}\right)$. This fact, (5.73) and $N = \mathcal{O}(d)$ imply the claimed bound on $\sum_{i=0}^{K}(1+\ell_i)N_i$.

The concavity of $t \mapsto \log_{\frac{1}{\theta}} t$ and Jensen's inequality imply

$$\mathbb{E}[\ell_k] \leq \mathbb{E}\left[\log_{\frac{1}{\theta}}\left(\frac{\hat{\alpha}\widehat{L}(\xi^k)}{(\lambda\theta)\wedge\hat{\alpha}}\right)\right] \leq \log_{\frac{1}{\theta}}\left(\frac{\hat{\alpha}L}{(\lambda\theta)\wedge\hat{\alpha}}\right),$$

where we used that $\mathbb{E}[\widehat{L}(\xi^k)] = L$ by definitions of $\widehat{L}(\xi^k)$ and L and Assumption 21. We take expectation in (5.73), use the above relation and the fact that $N := \mathcal{O}(d)$. This implies the claimed bound on $\sum_{i=0}^{K}(1+\mathbb{E}[\ell_i])N_i$. □

Remark 13 (Linear memory budget per operation)**.** *The policy in Corollary 5 requires the computation of a sum of size $N_k \sim dk$ (up to logs) of d-dimensional vectors at iteration k of the* Algorithm 5*. For large d, such computation is still cheap in terms of memory budget* per operation*: it can be computed in parallel or serially in k steps, each one requiring memory of $\mathcal{O}(d)$ per operation.*

Remark 14. *Recall the constant definitions in Assumption 1, Lemma 15, Theorem 12, Remark 9 and Remark 12 with $p = 2$. Recall also the definition of C_0 in Proposition 12. By Proposition 13 (X unbounded), the constant J in Theorem 18 satisfies*

(5.74)

$$\mathsf{J} \leq \frac{\max_{k\in\{0,\dots,k_0\}} \left|\left\|x^k - x^*\right\|\right|_2^2 + \frac{\phi \mathsf{C}_2 \sigma_4(x^*)^2}{\overline{\mathsf{C}}_2 \overline{L}_4^2} + 1}{1-\phi} \lesssim \max_{k\in\{0,\dots,k_0\}} \left|\left\|x^k - x^*\right\|\right|_2^2 + \frac{\sigma_4(x^*)^2}{|\mathsf{L}(\xi)|_4^2}.$$

Moreover, if we choose the sequence $\{\delta_k\}$ in Algorithm 5 *such that, for some $\Delta_0 > 0$,*

$$\delta_k := \frac{\Delta_0}{(k+\mu)^{1/2}(\ln(k+\mu))^{\frac{1+b}{2}}},$$

then, from (5.69) *and* (5.71)*, k_0 in* (5.74) *can be estimated by*

$$k_0 := \left\lceil \exp\left\{ \sqrt[b]{\frac{\overline{\mathsf{C}}_2 \left(\hat{\alpha}\overline{L}_4\right)^2}{\phi b N}} \right\} - \mu + 1 \right\rceil \bigvee \left\lceil \exp\left\{ \sqrt[b]{\frac{\mathsf{C}_0 \Delta_0^2}{b}} \right\} - \mu + 1 \right\rceil. \tag{5.75}$$

Remark 15 (Boundedness in $\mathcal{L}^p$)**.** *Adapting the proofs of Propositions 11 and 13, it is possible to prove, in case X is unbounded, that* the sequence $\{x^k\}$ is $\mathcal{L}^p$-bounded *for any given $p \geq 4$ satisfying Assumption 1. This is a significant statistical stability property. The proof exploits that $\Delta M_k(x^*)$ in* (5.53) *is still a martingale difference even if ϵ_3^k in* (5.46) *is not.*

5.3 Analysis of Algorithm 6 for Hölder continuous operators

With respect to Algorithm 6, we will set $y^k := x^k - \gamma_k \widehat{F}(\xi_k, z^k)$ and study the stochastic process $\{x^k\}$ with respect to the filtration

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

We will replace Assumption 21 by the following one.

Assumption 22 (I.I.D. sampling)**.** *In Algorithm 6, the sequence* $\{\xi_j^k : k \in \mathbb{N}_0, j \in [N_k]\}$ *is an i.i.d. sample drawn from* $\mathbf{P}$ *and* $\sum_{k=0}^{\infty} N_k^{-\frac{1}{2}} < \infty$.

We also define the oracle errors:

$$\bar{\epsilon}_1^k := \widehat{F}(\xi^k, x^k) - T(x^k), \tag{5.76}$$

$$\bar{\epsilon}_2^k := \widehat{F}(\xi^k, z^k) - T(z^k), \tag{5.77}$$

$$\bar{\epsilon}_3^k := \widehat{F}(\xi^k, \hat{z}^k) - T(z^k), \tag{5.78}$$

where $\hat{z}^k := \bar{z}^k(\theta^{-1}\alpha_k)$ (see line search (5.12) for the definition of $\bar{z}^k(\alpha)$). We remark that $\overline{\epsilon}_2^k$ and $\overline{\epsilon}_3^k$ are correlated errors in the sense that z^k and $\hat{z}^k$ are dependent on ξ^k. In the setting of Theorem 12, this means that $z^k = \overline{z}_{\beta_k}(\xi^k; \alpha_k, x^k)$ and $\hat{z}^k = \overline{z}_{\beta_k}(\xi^k; \theta^{-1}\alpha_k, x^k)$. We start by showing the line search (5.12) in Algorithm 6 is well defined.

Lemma 19 (Good definition of the line search)**.** *Consider Assumption 1. Then*

i) The line search (5.12) *in* Algorithm 6 *terminates after a finite number of iterations.*

ii) If Algorithm 6 *does not stop at iteration* $k+1$*, then* $\left\langle \widehat{F}(\xi^k, z^k), x^k - z^k \right\rangle > 0$. *In particular,* $\gamma_k > 0$ *in* (5.14).

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

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

We let $\ell \to \infty$ above and by continuity of $\widehat{F}(\xi^k, \cdot)$, resulting from Assumption 1, 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 20. *Consider Assumptions 2-20 and* (5.77). *Suppose that* Algorithm 6 *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 $\gamma_k \langle \bar{\epsilon}_2^k, x - z^k\rangle$. This term is not a martingale difference, since z^k depends on ξ^k. We shall need the following lemma.

Lemma 21. *Suppose that* Algorithm 6 *does not stop at iteration* $k+1$. *Then*

$$0 < \gamma_k < \frac{\alpha_k \beta_k}{\lambda} \leq \frac{\widehat{\alpha}\beta_k}{\lambda}. \tag{5.79}$$

Proof. We only need to prove the second inequality. The line search (5.12) and the fact that $x^k - z^k = \alpha_k(x^k - \Pi(g^k))$ imply that

$$\langle \widehat{F}(\xi^k, z^k), x^k - z^k\rangle \geq \frac{\lambda}{\alpha_k \beta_k} \|x^k - z^k\|^2. \tag{5.80}$$

From (5.80) and the definition of γ_k we get

$$\gamma_k = \frac{\langle \widehat{F}(\xi^k, z^k), x^k - z^k\rangle}{\|\widehat{F}(\xi^k, z^k)\|^2} > \frac{\lambda}{\alpha_k \beta_k} \frac{\|x^k - z^k\|^2}{\|\widehat{F}(\xi^k, z^k)\|^2}, \tag{5.81}$$

while the definition of γ_k gives

$$\gamma_k = \frac{\langle \widehat{F}(\xi^k, z^k), x^k - z^k\rangle}{\|\widehat{F}(\xi^k, z^k)\|^2} \leq \frac{\|\widehat{F}(\xi^k, z^k)\| \|x^k - z^k\|}{\|\widehat{F}(\xi^k, z^k)\|^2} = \frac{\|x^k - z^k\|}{\|\widehat{F}(\xi^k, z^k)\|}, \tag{5.82}$$

using the Cauchy-Schwartz inequality. Inequalities (5.81)-(5.82) imply the claim. □

Lemma 22 (Error decay)**.** *Consider Assumptions 1, 2 and 22 and* (5.77)*. Suppose that* Algorithm 6 *does not stop at iteration* $k+1$*. Then, for all* $x^* \in X^*$*,*

$$\left|\gamma_k\langle\bar{\epsilon}_2^k, x^* - z^k\rangle\right|\mathcal{F}_k\Big|_{\frac{p}{2}} \lesssim \frac{1+\|x^k - x^*\|^2}{\sqrt{N_k}}.$$

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

$$x^* - z^k = \alpha_k(x^* - \tilde{z}^k) + (1-\alpha_k)(x^* - x^k), \tag{5.83}$$

using the fact that $x^* = \alpha_k x^* + (1-\alpha_k)x^*$. In view of (5.83), we have

$$\begin{aligned} \gamma_k\langle\bar{\epsilon}_2^k, x^* - z^k\rangle &= \gamma_k\alpha_k\langle\bar{\epsilon}_2^k, x^* - \tilde{z}^k\rangle + \gamma_k(1-\alpha_k)\langle\bar{\epsilon}_2^k, x^* - x^k\rangle \\ &\leq \frac{\tilde{\beta}}{\lambda}\|\bar{\epsilon}_2^k\|\left(\|x^* - \tilde{z}^k\| + \|x^* - x^k\|\right), \end{aligned} \tag{5.84}$$

using the Cauchy-Schwarz inequality, Lemma 21, 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

$$\begin{aligned} \|\tilde{z}^k - x^*\| &= \|\Pi[x^k - \beta_k(T(x^k) + \bar{\epsilon}_1^k)] - \Pi[x^* - \beta_k T(x^*)]\| \\ &\leq \|x^k - x^* + \beta_k(T(x^*) - T(x^k)) - \beta_k\bar{\epsilon}_1^k\| \\ &\leq \|x^k - x^*\| + \tilde{\beta}L\|x^k - x^*\|^\delta + \tilde{\beta}\|\bar{\epsilon}_1^k\|, \end{aligned} \tag{5.85}$$

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

Using (5.84)-(5.85) and the fact that $\|x^k - x^*\|^\delta \leq 1 + \|x^k - x^*\|$, we take $|\cdot|\mathcal{F}_k|_{\frac{p}{2}}$ and get

$$\left|\gamma_k\langle\bar{\epsilon}_2^k, x^* - z^k\rangle\right|\mathcal{F}_k\Big|_{\frac{p}{2}} \leq \left[\tilde{\beta}L + (2+\tilde{\beta}L)\|x^k - x^*\|\right]\frac{\tilde{\beta}}{\lambda}\left|\|\bar{\epsilon}_2^k\|\right|\mathcal{F}_k\Big|_{\frac{p}{2}} + \frac{\tilde{\beta}^2}{\lambda}\left|\|\bar{\epsilon}_1^k\|\|\bar{\epsilon}_2^k\|\right|\mathcal{F}_k\Big|_{\frac{p}{2}}, \tag{5.86}$$

using the fact that $x^k \in \mathcal{F}_k$. By Lemma 15 with $q = p$ and the facts that $x^k \in \mathcal{F}_k$ and $\xi^k \perp\!\!\!\perp \mathcal{F}_k$, we get

$$\left|\|\bar{\epsilon}_1^k\|\right|\mathcal{F}_k\Big|_p \leq C_p\frac{\sigma_p(x^*) + L_p + L_p\|x^k - x^*\|}{\sqrt{N_k}}, \tag{5.87}$$

where we used the fact that $\|x^k - x^*\|^\delta \leq 1 + \|x^k - x^*\|$. By Theorem 12, (5.77) and the facts that $z^k = \overline{z}_{\beta_k}(\xi^k; \alpha_k, x^k)$, $x^k \in \mathcal{F}_k$ and $\alpha_k \in (0, 1]$, we get

$$\left\|\|\bar{\epsilon}_2^k\|\middle|\mathcal{F}_k\right|_{\frac{p}{2}} \leq \left\|\|\bar{\epsilon}_2^k\|\middle|\mathcal{F}_k\right|_p \lesssim \frac{\sigma_{2p}(x^*) + \|x^k - x^*\|}{\sqrt{N_k}}, \tag{5.88}$$

where we used the fact that $\delta \vee \|x^k - x^*\|^\delta \leq 1 + \|x^k - x^*\|$. Invoking Hölder's inequality, we also get

$$\left\|\|\bar{\epsilon}_1^k\|\|\bar{\epsilon}_2^k\|\middle|\mathcal{F}_k\right|_{\frac{p}{2}} \leq \left\|\|\epsilon_1^k\|\middle|\mathcal{F}_k\right|_p \cdot \left\|\|\epsilon_2^k\|\middle|\mathcal{F}_k\right|_p. \tag{5.89}$$

Relations (5.86)-(5.89) prove the claim. □

Proposition 14 (Stochastic quasi-Fejér property)**.** *Consider Assumptions 1, 2-20 and 22. Assume that* Algorithm 6 *generates an infinite sequence* $\{x^k\}$*. Then*

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

$$\mathbb{E}\Big[\|x^{k+1} - x^*\|^2\Big|\mathcal{F}_k\Big] \leq \|x^k - x^*\|^2 - \mathbb{E}\Big[\|y^k - x^k\|^2\Big|\mathcal{F}_k\Big] + c(x^*)\frac{1 + \|x^k - x^*\|^2}{\sqrt{N_k}}.$$

(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 20, 22 and the fact that $x^k \in \mathcal{F}_k$, after taking $\mathbb{E}[\cdot|\mathcal{F}_k]$ in Lemma 20.

ii) Set $\mathsf{c}_k(x^*) := \frac{c(x^*)}{\sqrt{N_k}}$. From (i), for all $k \in \mathbb{N}_0$,

$$\mathbb{E}\Big[\|x^{k+1} - x^*\|^2\Big|\mathcal{F}_k\Big] \leq [1 + \mathsf{c}_k(x^*)]\,\|x^k - x^*\|^2 + \mathsf{c}_k(x^*). \tag{5.90}$$

By Assumption 22, we have $\sum_k \mathsf{c}_k(x^*) < \infty$. Hence, from (5.90) 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 (5.90) and the fact that $x^k \in \mathcal{F}_k$ imply

$$\mathbb{E}\Big[\,\mathrm{d}(x^{k+1}, X^*)^2\Big|\mathcal{F}_k\Big] \leq \Big[1 + \mathsf{c}_k(\bar{x}^k)\Big]\,\mathrm{d}(x^k, X^*)^2 + \mathsf{c}_k(\bar{x}^k). \tag{5.91}$$

The boundedness of $\{\bar{x}^k\}$ and Assumption 22 imply that a.s. $\sum_k \mathsf{c}_k(\bar{x}^k) < \infty$. Hence, Theorem 1 and (5.91) imply that $\{\mathrm{d}(x^k, X^*)\}$ a.s.-converges.

iii) Suppose that a.s. there exists $\bar{x} \in X^*$ and a subsequence $\{k_\ell\}$ such that $\lim_{\ell\to\infty} \|x^{k_\ell} - \bar{x}\| = 0$. Clearly, $\mathrm{d}(x^{k_\ell}, X^*) \leq \|x^{k_\ell} - \bar{x}\|$ a.s., 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 6.

Theorem 19 (Asymptotic convergence)**.** *Under Assumptions 1, 2-20 and 22, either* Algorithm 6 *stops at iteration $k+1$, in which case x^k is a solution of* $\mathrm{VI}(T, X)$*, or it generates an infinite sequence $\{x^k\}$ 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 6 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 have

$$\langle \widehat{F}(\xi^k, x^k), x - x^k \rangle \geq 0, \qquad \forall x \in X. \tag{5.92}$$

From Assumption 22, (1.2) and the facts that $x^k \in \mathcal{F}_k$ and $\xi^k \perp\!\!\!\perp \mathcal{F}_k$, 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 take $\mathbb{E}[\cdot|\mathcal{F}_k]$ in (5.92) and obtain $\langle T(x^k), x - x^k \rangle \geq 0$, for all $x \in X$. Hence $x^k \in X^*$.

We now suppose that the sequence $\{x^k\}$ is infinite. By Proposition 14(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 14, set $\mathsf{c}_k(x^*) := \frac{c(x^*)}{\sqrt{N_k}}$. Using the property that $\mathbb{E}[\mathbb{E}[\cdot|\mathcal{F}_k]] = \mathbb{E}[\cdot]$, we take the expectation in Proposition 14(i), and get, for all $k \in \mathbb{N}_0$,

$$\mathbb{E}\Big[\|x^{k+1} - x^*\|^2 \Big| \mathcal{F}_k\Big] \leq [1 + \mathsf{c}_k(x^*)]\, \mathbb{E}\left[\|x^k - x^*\|^2\right] - \mathbb{E}\left[\|y^k - x^k\|^2\right] + \mathsf{c}_k(x^*). \tag{5.93}$$

From the fact that $\sum_k \mathsf{c}_k(x^*) < \infty$ (Assumption 22), (5.93) and Theorem 1 we conclude that

$$\sum_{k=0}^{\infty} \mathbb{E}\left[\|y^k - x^k\|^2\right] < \infty, \tag{5.94}$$

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 6, 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 (5.94) we get

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{\langle T(z^k) + \bar{\epsilon}_2^k, x^k - z^k\rangle^2}{\|T(z^k) + \bar{\epsilon}_2^k\|^2}\right] = 0. \tag{5.95}$$

From the definitions of $\{\bar{\epsilon}_1^k, \bar{\epsilon}_2^k, \bar{\epsilon}_3^k\}$ in (5.76)-(5.78), Lemma 15 with $q = p = 2$, Theorem 12(i) and the facts that $z^k = \overline{z}_{\beta_k}(\xi^k; \alpha_k, x^k)$ and $\widehat{z}^k = \overline{z}_{\beta_k}(\xi^k; \theta^{-1}\alpha_k, x^k)$, 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

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

for $s \in \{1, 2, 3\}$ and all $k \in \mathbb{N}_0$. Since $\lim_{k\to\infty} N_k^{-1} = 0$ (Assumption 22), we have in particular that, for $s \in \{1, 2, 3\}$,

$$\lim_{k\to\infty} \mathbb{E}[\|\bar{\epsilon}_s^k\|^2] = 0. \tag{5.96}$$

Since $\mathcal{L}^2$-convergence implies a.s.-convergence along a subsequence, from (5.95)-(5.96), we may take a (deterministic) subsequence $\{k_\ell\}_{\ell=1}^\infty$ such that a.s. for $s \in \{1, 2, 3\}$,

$$\lim_{\ell\to\infty} \frac{\alpha_{k_\ell}\langle T(z^{k_\ell}) + \bar{\epsilon}_2^{k_\ell}, x^{k_\ell} - \Pi(g^{k_\ell})\rangle}{\|T(z^{k_\ell}) + \bar{\epsilon}_2^{k_\ell}\|} = 0, \tag{5.97}$$

$$\lim_{\ell\to\infty} \bar{\epsilon}_s^{k_\ell} = 0, \tag{5.98}$$

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 $\beta > 0$,

$$\lim_{\ell\to\infty} \beta_{k_\ell} = \beta. \tag{5.99}$$

From Proposition 14(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

$$\lim_{k\in\mathfrak{N}} x^k = x^*, \tag{5.100}$$

for some (random) $x^* \in \mathbb{R}^d$. Using the fact that $g^k = x^k - \beta_k[T(x^k) + \bar{\epsilon}_1^k]$, (5.98)-(5.100) and the continuity of T and Π, for the event Ω_1, we have

$$g^* := \lim_{k\in\mathfrak{N}} g^k = x^* - \beta T(x^*). \tag{5.101}$$

Also, for the event Ω_1, from the definition of z^k in (5.13), the fact that $\alpha_k \in (0,1]$, (5.98) and (5.100)-(5.101), we get that $\{T(z^k) + \bar{\epsilon}_2^k\}_{k\in\mathfrak{N}}$ is bounded so that, since (5.97), we obtain

$$\lim_{k\in\mathfrak{N}} \alpha_k \langle T(z^k) + \bar{\epsilon}_2^k, x^k - \Pi(g^k)\rangle = 0. \tag{5.102}$$

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 (5.102) that on Ω_1,

$$\lim_{k\in\mathfrak{N}} \langle T(z^k) + \bar{\epsilon}_2^k, x^k - \Pi(g^k)\rangle = 0. \tag{5.103}$$

From (5.12)-(5.13), we get

$$\langle T(z^k) + \bar{\epsilon}_2^k, x^k - \Pi(g^k)\rangle \geq \frac{\lambda}{\beta_k}\|x^k - \Pi(g^k)\|^2 \geq \frac{\lambda}{\bar{\beta}}\|x^k - \Pi(g^k)\|^2 \tag{5.104}$$

for all k. Relations (5.103)-(5.104) imply that, on Ω_1,

$$0 = \lim_{k\in\mathfrak{N}} \|x^k - \Pi(g^k)\|. \tag{5.105}$$

From (5.100)-(5.101), we take limits in (5.105) and obtain, by continuity of Π,

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

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

$$\lim_{k\in\mathfrak{N}} \theta^{-1}\alpha_k = 0. \tag{5.106}$$

Since $\hat{z}^k := \theta^{-1}\alpha_k \Pi(g^k) + (1 - \theta^{-1}\alpha_k)x^k$ and $\{g^k\}_{k\in\mathfrak{N}}$ is bounded, we get from (5.100) and (5.106) that

$$\lim_{k\in\mathfrak{N}} \hat{z}^k = x^*. \tag{5.107}$$

Observe that, by the definition of the line search rule (5.12) and (5.78), we have

$$\langle T(\hat{z}^k) + \bar{\epsilon}_3^k, x^k - \Pi(g^k)\rangle < \frac{\lambda}{\beta_k}\|x^k - \Pi(g^k)\|^2, \tag{5.108}$$

for all $k \in \mathbb{N}_0$. We take limit in (5.108) along $\mathfrak{N}$, and we get, using the continuity of T and Π and relations (5.98)-(5.101) and (5.107) that

$$\langle T(x^*), x^* - \Pi(g^*)\rangle \leq \frac{\lambda}{\beta}\|x^* - \Pi(g^*)\|^2. \tag{5.109}$$

Since the sequence $\{x^k\}$ is feasible and X is closed, the limit point x^* belongs to X. Thus, from (5.109) and Lemma 1(v), we get that, on Ω_1,

(5.110)

$$\lambda\|x^* - \Pi(g^*)\|^2 \geq \beta\langle T(x^*), x^* - \Pi(g^*)\rangle = \langle x^* - g^*, x^* - \Pi(g^*)\rangle \geq \|x^* - \Pi(g^*)\|^2.$$

Since $\lambda \in (0,1)$, (5.110) implies that $\|x^* - \Pi(g^*)\| = 0$. Hence, in view of (5.101), we have $x^* = \Pi(x^* - \beta T(x^*))$. By Lemma 1(iv), we conclude that $x^* \in X^*$.

We have proved that on 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 14(iii). □

5.4 Discussion on the complexity constants of Algorithm 5

Suppose the oracle is *exact*. In that case, Algorithm 5 would have essentially the same rate estimates, up to universal constants and a factor of $\mathcal{O}(\ln L)$ in the oracle complexity, either if a line search scheme is used or a CSP is used with a known Lipschitz constant (LC). The reason is that the Lipschitz continuity is only related to the *smoothness class* of the operator. The situation is different when the oracle is *stochastic*: the Lipschitz continuity also quantifies the *spread of the oracle's error variance*.[7] Consequently, the lack of knowledge of the LC is much more demanding in the stochastic case. It is instructive to compare the complexity constants when the LC is known or not. In the following, we recall the rate of convergence of Theorem 18 and the constants defined in Assumption 1, Theorem 12, Lemma 15, Remarks 9 and 12 and Proposition 11 with $p = 2$.

[7]This is true either for the martingale difference errors $\{\epsilon_i^k\}_{i=1,2}$ or the correlated error ϵ_3^k in (5.46). The Lipschitz continuity in the analysis of ϵ_3^k is crucial in our chaining and self-normalization arguments of Lemmas 13 and 14.

Suppose first the LC is known. This was already considered in Chapter 4 under a more general condition than Assumption 1. However, it leads to weaker complexity constants as argued in the following.[8] It is possible to show that if the stronger but fairly general condition of Lemma 1 holds and $\hat{\alpha} = \mathcal{O}(\frac{1}{L_2})$, then the rate statement of Theorem 18 and the estimates (5.74)-(5.75) are valid when we replace $\sigma_4(x^*)$ by $\sigma_2(x^*)$, $\overline{L}_4$ by L_2 and[9] the coefficient $(1-6\lambda^2)[(\lambda\theta)\wedge\hat{\alpha}]$ by a term of order $1-\mathcal{O}(1)(\hat{\alpha}L_2)^2$. Since $\hat{\alpha}L_2 \lesssim 1$ we also have $\mathsf{C}_2 \lesssim 1$ and $\overline{\mathsf{C}}_2 \lesssim 1$. Assuming L_2 is known, we obtain a property not satisfied by the estimates in Chapter 4: k_0 in (5.75) is *independent of the oracle's error variances* $\{\sigma_2(x)^2\}_{x\in X}$ *over* X and there exist b, N and μ and policy $\hat{\alpha} = \mathcal{O}(\frac{1}{L_2})$ such that $k_0 := 0$. It is then possible to obtain the rate

$$\min_{i=0,\dots,k} \mathbb{E}\left[r(x^i)^2\right] \lesssim \frac{L_2^2\|x^0 - x^*\|^2 + \sigma_2(x^*)^2}{k}, \tag{5.111}$$

which depends only on the *local* variance $\sigma_2(x^*)^2$ and the *initial* iterate x^0. This can be seen as a *variance localization property*. We note that the above rate is sharper than those obtained in Chapter 4.[10]

Consider now the more challenging regime when the LC is unknown. As expected, the constants in the rate of Theorem 18 are less sharp then the ones in (5.111). First, (5.111) is not explicitly dependent on the dimension d. In terms of dimension, the rate in Theorem 18 is of $\mathcal{O}(\frac{d}{N})$ and, thus, it is valid in the large sample regime $N := \mathcal{O}(d)$. This is a manifestation of our need to treat correlated errors when using a line search scheme. Such scheme is an inner statistical estimator for the LC. Second, if we set $\mathsf{M} := (\hat{\alpha}|\mathsf{L}(\xi)|_4)^2$, then the constants in the rate of Theorem 18 satisfy $\mathsf{C}_2 \lesssim \frac{\mathsf{M}}{N}$, $\overline{\mathsf{C}}_2 \lesssim \mathsf{M}$ and $\frac{(\hat{\alpha}\widetilde{L}_2)^2\mathsf{J}}{N} \lesssim \mathsf{M}^2\mathsf{J}$, for a general[11] X and $\mathsf{C}_2 \lesssim 1$, $\overline{\mathsf{C}}_2 \lesssim 1$ and $\frac{(\hat{\alpha}\widetilde{L}_2)^2\mathsf{J}}{N} \lesssim \mathsf{M}\,\mathrm{diam}(X)^2$, for a compact X. Observe that a line search scheme can only estimate a *lower bound* for $|\mathsf{L}(\xi)|_4$. For a large $\hat{\alpha}$,

[8] Differently than Lemma 1, it allows the multiplicative noise to depend on the reference point $x^* \in X^*$.

[9] Up to universal constants, C_2 and $\overline{\mathsf{C}}_2$ are unchanged.

[10] In Chapter 4, given $x^* \in X^*$, the rate is of the order of $\sigma(x^*)^4 \cdot \max_{0\le i\le k_0(x^*)} \mathbb{E}[\|x^i - x^*\|^2]$, where $k_0(x^*) \in \mathbb{N}_0$ depends on $\sigma(x^*)$.

[11] The given order of dependence on M for an unbounded X is an artifact of our proof techniques. We believe a sharper dependence can be obtained via more sophisticated concentration inequalities (instead of moment inequalities).

the lack of an *upper bound* leads to a rate with larger constants when compared to (5.111). This is a manifestation of our absence of information of the LC. Note that robust methods are expected to have nonoptimal constants since the endogenous parameters are unknown [60]. Third, note that (5.111) only depends on the *initial* iterate x^0. This is possible since k_0 in (5.75) can be calibrated using the knowledge of the LC. For an unknown LC and for an unbounded X, k_0 depends on M but it still *independent of the oracle's error moments* $\{\sigma_4(x)\}_{x\in X}$ *over* X. Differently than (5.111), for a large $\hat{\alpha}$ (implying a larger value for M), the rate in Theorem 18 will depend on $D_{k_0}^2(x^*) := \max_{k=0,\dots k_0} \mathbb{E}[\|x^k - x^*\|^2]$ for a possibly large k_0. Although not as sharp as (5.111), the resulted rate estimate for a large k_0 is not a limiting issue. It is still in accordance to, and in fact generalize, previous estimates which rely on compactness of X (see e.g. [60]): for a compact X, we have $\max_{k=0,\dots k_0} \mathbb{E}[\|x^k - x^*\|^2] \leq \mathrm{diam}(X)^2$.

Appendix

Proof of Lemma 1. By Jensen's inequality and Assumption 1 we get

$$\|T(x) - T(x_*)\| \leq \mathbb{E}\left[\|F(\xi, x) - F(\xi, x_*)\|\right] \leq \mathbb{E}[\mathsf{L}(\xi)]\|x - y\|^\delta.$$

Using this fact and definition (5.1), we get

$$\begin{aligned}
\left|\left|\left|\epsilon(\xi, x)\right|\right|\right|_q &\leq \left|\left|\left|F(\xi, x) - F(\xi, x_*)\right|\right|\right|_q + \left|\left|\left|F(\xi, x_*) - T(x_*)\right|\right|\right|_q + \left|\left|\left|T(x) - T(x_*)\right|\right|\right|_q \\
&\leq \left|\mathsf{L}(\xi)\|x - x_*\|^\delta\right|_q + \left|\left|\left|\epsilon(\xi, x_*)\right|\right|\right|_q + L\|x - x_*\|^\delta \\
&= \left|\left|\left|\epsilon(\xi, x_*)\right|\right|\right|_q + \left(\left|\mathsf{L}(\xi)\right|_q + L\right)\|x - y\|^\delta,
\end{aligned}$$

where we used the triangle inequality for $\|\cdot\|$ and Minkowski's inequality for $|\cdot|_q$. The claim is proved from the above fact, (5.2) and $L_q = |\mathsf{L}(\xi)|_q + L$. □

Proof of Lemma 20. By Lemma 19(ii), we have that $\gamma_k > 0$. Thus

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \|\Pi(y^k) - x^*\|^2 \\
&\leq \|y^k - x^*\|^2 - \|y^k - \Pi(y^k)\|^2 \\
&\leq \|y^k - x^*\|^2 \\
&= \|(x^k - x^*) - \gamma_k(T(z^k) + \bar{\epsilon}_2^k)\|^2 \\
&= \|x^k - x^*\|^2 + \gamma_k^2\|T(z^k) + \bar{\epsilon}_2^k\|^2 - 2\gamma_k\langle T(z^k) + \bar{\epsilon}_2^k, x^k - x^*\rangle, \qquad (5.112)
\end{aligned}$$

using Lemma 1(ii) in the first inequality. Concerning the last term in the rightmost expression of (5.112), we have

$$\begin{aligned} -2\gamma_k\langle T(z^k)+\bar{\epsilon}_2^k, x^k-x^*\rangle &= -2\gamma_k\langle T(z^k)+\bar{\epsilon}_2^k, x^k-z^k\rangle + \\ &\quad 2\gamma_k\langle T(z^k), x^*-z^k\rangle + 2\gamma_k\langle \bar{\epsilon}_2^k, x^*-z^k\rangle \\ &= -2\gamma_k(\gamma_k\|T(z^k)+\bar{\epsilon}_2^k\|^2) \\ &\quad +2\gamma_k\langle T(z^k), x^*-z^k\rangle + 2\gamma_k\langle \bar{\epsilon}_2^k, x^*-z^k\rangle \\ &\leq -2\gamma_k^2\|T(z^k)+\bar{\epsilon}_2^k\|^2 + 2\gamma_k\langle \bar{\epsilon}_2^k, x^*-z^k\rangle, \end{aligned} \tag{5.113}$$

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 (5.112)-(5.113) we get

$$\begin{aligned} \|x^{k+1}-x^*\|^2 &\leq \|x^k-x^*\|^2 + \gamma_k^2\|T(z^k)+\bar{\epsilon}_2^k\|^2 - 2\gamma_k^2\|T(z^k)+\bar{\epsilon}_2^k\|^2 + 2\gamma_k\langle \bar{\epsilon}_2^k, x^*-z^k\rangle \\ &= \|x^k-x^*\|^2 - \gamma_k^2\|T(z^k)+\bar{\epsilon}_2^k\|^2 + 2\gamma_k\langle \bar{\epsilon}_2^k, x^*-z^k\rangle \\ &= \|x^k-x^*\|^2 - \|y^k-x^k\|^2 + 2\gamma_k\langle \bar{\epsilon}_2^k, x^*-z^k\rangle, \end{aligned} \tag{5.114}$$

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. □

Chapter 6

Conclusions and open questions

This thesis proposes stochastic approximation methods for the solution of stochastic variational inequalities, paying attention to asymptotic convergence (stability), convergence rate, oracle complexity, knowledge of problem parameters, data availability and distributed solution. See Section 1.5 for the precise statements of our contributions. We make some comments regarding possible interesting open questions.

In Chapter 3, we have proposed an incremental projection method for monotone SVIs using regularization. Motivated by the results of Chapter 4 (which avoids regularization by means of an extragradient method with a variance reduction procedure), we would like to device, if possible, an incremental projection method for plain monotone SVIs without regularization. By avoiding regularization, we may be able to prove optimal convergence rates, which are not reported in Chapter 3.

Regarding Chapters 4 and 5, we would like to prove convergence rates and oracle complexity with exponentially high probability, refine the analysis for important classes of VIs and support the results with a computational study of relevant large-scale problems. Importantly, we would like to maintain the assumptions in Chapters 4 and 5 of unboundedness of the feasible set and non-uniform variance of the oracle (which were major improvements compared to previous works).

The variance reduction scheme in Chapter 4 accelerates the rate without compromising the oracle complexity. Interestingly, our variance reduction scheme is robust. We would like to study a stochastic extragradient method for SVIs which

combines our robust variance reduction scheme with robust stepsizes in the sense of [60]. These features are very relevant in practice. Another important question for large-scale problems (such as equilibrium problems over large networks) is the devise of methods with optimal dependence on data *dimension* or *diameter* of the feasible set. Thus, we would also like to include dimension-reduction techniques in the mentioned proposals. Another relevant question in practice is the use of *inexact* projections. We also would like to include this feature in our method. In case of distributed solution of Cartesian variational inequalities, another interesting question would be to explore the topology of the network in order to require minimal coordination between agents as possible.

Bibliography

[1] Auslender, A. and Teboulle, M. (2005) Interior projection-like methods for monotone variational inequalities, *Mathematical Programming*, Vol. 104, pp. 39-68.

[2] Bach, F. and Moulines, E. (2011) Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning, *Advances in Neural Information Processing Systems (NIPS)*, conference paper.

[3] Bauschke, H.H. (2001) Projection algorithms: results and open problems. In: Butnariu, D., Censor, Y., Reich, Y. (eds.) *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, Elsevier, Amsterdam, pp. 11-22.

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

[5] Bauschke, H.H., Combettes, H.H. and Luke, D.R. (2003) Hybrid projection-reflection method for phase retrieval, *Journal of the Optical Society of America A*, Vol. 20, pp. 1025-1034.

[6] Bello Cruz, J.Y. and Iusem, A.N. (2010) Convergence of direct methods for paramonotone variational inequalities, *Computational Optimization and Applications*, Vol. 46, pp. 247–263.

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

[8] Bello Cruz, J.Y. and Iusem A.N. (2015) Full convergence of an approximate projections method for nonsmooth variational inequalities, *Mathematics and Computers in Simulation*, Vol. 114, pp. 2-13.

[9] Bertsekas, D.P. (2011) Incremental proximal methods for large scale convex optimization, *Mathematical Programming*, Vol. 129, pp. 163-195.

[10] Billingsley, P. (1968) *Convergence of Probability Measures*, John Wiley, New York.

[11] Boucheron, S., Lugosi G. and Massart, P. (2013) *Concentration inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, Oxford.

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

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

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

[15] Byrd, R.H., Chin, G.M., Nocedal, J. and Wu, Y. (2012) Sample size selection in optimization methods for machine learning, *Mathematical Programming*, Vol. 134, pp. 127-155.

[16] Cegielski, A. and Suchocka, A. (2008) Relaxed alternating projection methods, *SIAM Journal on Optimization*, Vol. 19, pp. 1093-1106.

[17] Censor, Y. (1981) Row-action methods for huge and sparse systems and its applications, *SIAM Review*, Vol. 23, pp. 444-464.

[18] Censor, Y. and Gibali, A. (2008) Projections onto super-half-spaces for monotone variational inequality problems in finite-dimensional spaces, *Journal of Nonlinear and Convex Analysis*, Vol. 9, pp. 461-474.

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

[20] Chen, Y., Lan, G. and Ouyang, Y. Accelerated schemes for a class of variational inequalities, *http://arxiv.org/abs/1403.4164*, preprint.

[21] Dang, C.D. and Lan, G. (2015) On the convergence properties of non-Euclidean extragradient methods for variational inequalities with generalized monotone operators, *Computational Optimization and Applications*, Vol. 60, pp. 277-310.

[22] Deng, G. and Ferris, M.C. (2009) Variable-number sample-path optimization, *Mathematical Programming*, Vol. 117, pp. 81-109.

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

[24] Duchi, J.C., Bartlett, P.L. and Wainwright, M.J. (2012) Randomized smoothing for stochastic optimization, *SIAM Journal on Optimization*, Vol. 22, pp. 674-701.

[25] Dudley, R.M. (1967) The sizes of compact subsets of Hilbert space and continuity of Gaussian processes, *Journal of Functional Analysis*, 1, pp. 290-330.

[26] Durret, R. (2010) *Probability: Theory and Examples*, Cambridge University Press, Cambridge.

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

[28] Ferris, M.C. and Pang, J.S. (1997) Engineering and economic applications of complementarity problems, *SIAM Review*, Vol. 39, No. 4, pp. 669-713.

[29] Friedlander, M.P. and Goh, G. Tail bounds for stochastic approximation, *http://arxiv.org/abs/1304.5586*, pre-print.

[30] Fukushima, M. (1986) A relaxed projection method for variational inequalities, *Mathematical Programming*, Vol. 35, pp. 58-70.

[31] Gürkan, G., Özge, A.Y. and Robinson, S.M. (1999) Sample-path solution of stochastic variational inequalities, *Mathematical Programmming*, Vol. 84, pp. 313-333.

[32] Homem-de-Mello, T. (2003) Variable-Sample Methods for Stochastic Optimization, *ACM Transactions on Modeling and Computer Simulation*, Vol. 13, pp. 108-133.

[33] Hsu, D., Kakade, S.M. and Zhang, T. (2012) A tail inequality for quadratic forms of subgaussian random vectors, *Electronic Communications in Probability*, Vol. 17, pp. 1-6.

[34] Huang, J., Subramanian, V.G., Agrawal, R. and Berry, R. (2009) Joint scheduling and resource allocation in uplink OFDM systems for broadband wireless access networks. *IEEE Journal on Selected Areas in Communications*, Vol. 27, pp. 226-234.

[35] Iusem, A.N. (1998) On some properties of paramonotone operators, *Journal of Convex Analysis*, Vol. 5, pp. 269-278.

[36] Iusem, A.N., Jofré A. and Thompson, P. (2017) Incremental constraint projection methods for monotone stochastic variational inequalities, *Mathematics of Operations Research*, under review.

[37] Iusem, A.N., Jofré A., Oliveira, R. and Thompson, P. (2016) Extragradient method with variance reduction for stochastic variational inequalities, *SIAM Journal on Optimization*, to appear.

[38] Iusem, A.N., Jofré A., Oliveira, R. and Thompson, P. (2017) Variane-based extragradient methods with line search for stochastic variational inequalities, *SIAM Journal on Optimization*, under review.

[39] Iusem, A.N. and Svaiter, B.F. (1997) A variant of Korpelevich's method for variational inequalities with a new search strategy, *Optimization*, Vol. 42, pp. 309-321.

[40] Jiang, H. and Xu, H. (2008) Stochastic approximation approaches to the stochastic variational inequality problem, *IEEE Transactions on Automatic Control*, Vol. 53, pp. 1462-1475.

[41] Jofré, A., Wets, Rockafellar, R. Terry and Wets, Roger J.-B. (2007) Variational inequalities and economic equilibrium, *Mathematics of Operations Research*, Vol. 32, No.1, pp. 32-50.

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

[43] Kannan, A. and Shanbhag, U.V. (2012) Distributed computation of equilibria in monotone Nash games via iterative regularization techniques, *SIAM Journal on Optimization*, Vol. 22, pp. 1177-1205.

[44] Kannan, A. and Shanbhag, U.V. (2014) The pseudomonotone stochastic variational inequality problem: Analytical statements and stochastic extragradient schemes, *American Control Conference (ACC)*, Portland, USA, pp. 2930-2935.

[45] 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.

[46] Kibardin, V.M. (1980) Decomposition into functions in the minimization problem, *Automation and Remote Control*, Vol. 40, pp. 1311-1323.

[47] Khobotov, E.N. (1987) Modifications of the extragradient method for solving variational inequalities and certain optimization problems, *USSR Computational Mathematics and Mathematical Physics*, Vol. 27, pp. 120-127.

[48] Konnov, I.V. (2007) *Equilibrium Models and Variational Inequalities*, Elsevier, Amsterdam.

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

[50] Koshal, J., Nedić, A. and Shanbhag U.V. (2013) Regularized Iterative Stochastic Approximation Methods for Stochastic Variational Inequality Problems, *IEEE Transactions on Automatic Control*, Vol. 58, pp. 594-609.

[51] Kushner, H.J. and Yin, G.G. (2003) *Stochastic approximation and recursive algorithms and applications*, Springer, New York.

[52] Luo, Z.-Q. and Tseng, P. (1994) Analysis of an approximate gradient projection method with applications to the backpropagation algorithm, *Optimization Methods and Software*, Vol. 4, pp. 85-101.

[53] Marcotte, P. and Zhu, D. (1998) Weak sharp solutions of variational inequalities, *SIAM Journal on Optimization*, Vol. 9, pp. 179-189.

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

[55] Monteiro, R.D. and Svaiter, B.F. (2010) On the complexity of the hybrid proximal extra-gradient method for the iterates and the ergodic mean, *SIAM Journal on Optimization*, Vol. 20, pp. 275-287.

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

[57] Nedić, A. (2011) Random algorithms for convex minimization problems, *Mathematical Programming Ser. B*, vol. 129, pp. 225-253.

[58] Nedić, A. and Bertsekas, D.P. (2001) Incremental subgradient method for nondifferentiable optimization, *SIAM Journal on Optimization*, Vol. 12, 109-138.

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

[60] Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A. (2009) Robust stochastic approximation approach to stochastic programming, *SIAM Journal on Optimization*, Vol. 19, pp. 1574-1609.

[61] Nesterov, Y. (2009) Primal-dual subgradient methods for convex problems, *Mathematical Programming*, Vol. 120, pp. 261-283.

[62] Panchenko, D. (2003) Symmetrization approach to concentration inequalities for empirical processes, *The Annals of Probability*, Vol. 1, pp. 2068-2081.

[63] Polyak, B.T. (1969) Minimization of unsmooth functionals, *U.S.S.R. Computational Mathematics and Mathermatical Physics*, Vol. 9, pp. 14-29.

[64] Polyak, B.T. (1987) *Introduction to Optimization*, Optimization Software, New York.

[65] Polyak, B.T. (2001) Random algorithms for solving convex inequalities, In: Butnariu, D., Censor, Y., Reich, S. (eds.) *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, Elsevier, Amsterdam, pp. 409-422.

[66] Robbins, H. and Monro, S. (1951) A Stochastic Approximation Method, *The Annals of Mathematical Statistics*, Vol. 22, pp. 400-407.

[67] Robbins, H. and Siegmund, D.O. (1971) A convergence theorem for non negative almost supermartingales and some applications, *Optimizing methods in statistics*, Academic Press, New York, pp. 233-257.

[68] Rockafellar, R.T. and Wets, R.J-B. (1998) *Variational Analysis*, Springer, Berlin.

[69] Rockafellar, R.T. and Wets, R.J-B., Stochastic Variational Inequalities: single-stage to multistage, pre-print.

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

[71] Solodov, M.V. and Svaiter, B.F. (1999) A new projection method for monotone variational inequality problems, *SIAM Journal on Control and Optimization*, Vol. 37, pp. 765-776.

[72] Wang, M. and Bertsekas, D.P. (2013) Incremental Constraint Projection-Proximal Methods for Nonsmooth Convex Optimization, *Lab. for Information and Decision Systems Report LIDS-P-2907*, MIT.

[73] Wang, M. and Bertsekas, D.P. (2015) Incremental Constraint Projection Methods for Variational Inequalities, *Mathematical Programming Ser. A*, pp. 150, pp. 321-363.

[74] Wang, Y.J., Xiu, N.H. and Wang, C.Y. (2001) Unified Framework of Extragradient-Type Methods for Pseudomonotone Variational Inequalities, *Journal of Optimization Theory and Applications*, Vol. 111, pp. 641-656.

[75] 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.

[76] 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.

[77] 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.