SIAM J. OPTIM.
Vol. 29, No. 1, pp. 175–206

VARIANCE-BASED EXTRAGRADIENT METHODS WITH LINE SEARCH FOR STOCHASTIC VARIATIONAL INEQUALITIES*

ALFREDO N. IUSEM†, ALEJANDRO JOFRÉ‡, ROBERTO I. OLIVEIRA†, AND PHILIP THOMPSON†

Abstract. In this paper, we propose dynamic sampled stochastic approximated (DS-SA) extragradient methods for stochastic variational inequalities (SVIs) that are *robust* with respect to an unknown Lipschitz constant L. We propose, to the best of our knowledge, the first provably convergent *robust* SA *method with variance reduction*, either for SVIs or stochastic optimization, assuming just an unbiased stochastic oracle within a large sample regime. This widens the applicability and improves, up to constants, the desired efficient acceleration of previous variance reduction methods, all of which still assume knowledge of L (and, hence, are not robust against its estimate). Precisely, compared to the iteration and oracle complexities of $\mathcal{O}(\epsilon^{-2})$ of previous robust methods with a small stepsize policy, our robust method uses a DS-SA line search scheme obtaining the faster iteration complexity of $\mathcal{O}(\epsilon^{-1})$ with oracle complexity of $(\ln L)\mathcal{O}(d\epsilon^{-2})$ (up to log factors on ϵ^{-1}) for a d-dimensional space. This matches, up to constants, the sample complexity of the sample average approximation estimator which does not assume additional problem information (such as L). Differently from previous robust methods for ill-conditioned problems, we allow an unbounded feasible set and an oracle with *multiplicative noise* (MN) whose variance is not necessarily uniformly bounded. These properties are appreciated in our complexity estimates which depend only on L and *local* variances or fourth moments at solutions. The robustness and variance reduction properties of our DS-SA line search scheme come at the expense of *nonmartingale-like dependencies* (NMDs) due to the needed inner statistical estimation of a lower bound for L. In order to handle an NMD and an MN, our proofs rely on a novel *iterative localization* argument based on empirical process theory.

Key words. stochastic variational inequalities, stochastic approximation, extragradient method, variance reduction, dynamic sampling, line search, empirical process theory

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

DOI. 10.1137/17M1144799

1. Introduction.

In this paper, we are concerned with methods for variational inequality (VI) problems where only a *random perturbation* of the operator is available. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In the following, $\mathbb{E}$ will always denote the expectation with respect to the probability measure $\mathbb{P}$. We shall denote by $\xi : \Omega \to \Xi$ a random variable[1] with distribution $\mathbf{P}$ so that $\mathbf{P}(A) := \mathbb{P}(\xi \in A)$ for any $A \in \mathcal{F}$. Given a closed convex set $X \subset \mathbb{R}^d$ and a measurable random operator $F : \Xi \times X \to \mathbb{R}^d$, we then define the expected operator

$$T(x) = \mathbb{E}\left[F(\xi, x)\right] = \int_{\Xi} F(\xi, x)\, \mathrm{d}\, \mathbf{P}(\xi), \quad (x \in X), \tag{1}$$

*Received by the editors August 25, 2017; accepted for publication (in revised form) November 7, 2018; published electronically January 17, 2019.
http://www.siam.org/journals/siopt/29-1/M114479.html

Funding: The third author's research was supported by a Bolsa de Produtividade em Pesquisa from CNPq, Brazil. His work is part of the activities of the FAPESP Center for Neuromathematics (grant 2013/07699-0, FAPESP - S. Paulo Research Foundation). The fourth author's research was supported by a CNPq Doctoral scholarship while he was a Ph.D. student at IMPA with visit appointments at CMM.

†Instituto de Matemática Pura e Aplicada (IMPA), Rio de Janeiro, RJ, CEP 22460-320, Brazil (iusp@impa.br, rimfo@impa.br, philip@impa.br).

‡Centro de Modelamiento Matemático (CMM & DIM), Santiago, CP 8370448, Chile (ajofre@dim.uchile.cl).

[1]We will sometimes use $\xi \in \Xi$ to denote a point in the sample space if no confusion arises.

assuming it is well defined. Under (1), the *stochastic variational inequality* (SVI) problem, denoted as VI(T, X), is the following problem: find an $x^* \in X$ such that, for all $x \in X$, $\langle T(x^*), x - x^* \rangle \geq 0$. Its solution set will be denoted by X^*. SVIs include *stochastic optimization* (SP) as a special case but also other SVIs for which T is not integrable, such as the *stochastic saddle-point problem*, the *stochastic Nash equilibrium* problem, and the *stochastic system of equations* [8, 15].

The challenge aspect of SVIs, when compared to deterministic VIs, is that the expectation (1) cannot be or is too expensive to be evaluated. However, a practical assumption is that the decision makers have access to the random operator F via samples drawn from the distribution $\mathbf{P}$, a procedure usually named a *stochastic oracle* (SO). Under an SO, a popular methodology to solve SVIs is the *stochastic approximation* (SA) method: samples are accessed in an interior and online fashion along the progress of a chosen algorithm [25]. A different approach is the *sample average approximation* (SAA) method: an external and offline sample is acquired to approximate the SVI, which is then solved by a deterministic algorithm of preferred choice [31].

The SA methodology was first proposed by Robbins and Monro [28] for the problem $\min_{x \in \mathbb{R}^d} \{f(x) := \mathbb{E}[G(\xi, x)]\}$ for a random smooth convex function $G : \Xi \times \mathbb{R}^d \to \mathbb{R}$. Their method takes the form $x^{k+1} := x^k - \alpha_k \nabla G(\xi^k, x^k)$, given an i.i.d. sample sequence $\{\xi^k\}$ and positive stepsize sequence $\{\alpha_k\}$. This is a first instance of the *stochastic gradient* (SG) *method*. This methodology was then extensively explored in numerous works spanning the communities of statistics and stochastic approximation, stochastic optimization, and machine learning (see, e.g., [3, 4] and references therein). More recently, the SA methodology was also analyzed for SVIs. See, e.g., [9, 14, 15, 18, 7, 12, 11].

Given $p \geq 2$ and $(\xi, x) \in \Xi \times X$, consider the maps

$$\epsilon(\xi, x) := F(\xi, x) - T(x), \qquad \sigma_p(x) := \sqrt[p]{\mathbb{E}\left[\|\epsilon(\xi, x)\|^p\right]}. \tag{2}$$

The estimation in SA methods is controlled by the *oracle error* $\epsilon(\xi, x)$. In the analysis of SA methods, assumptions on the *oracle error's p-moment* $\sigma_p(\cdot)$ is as important as assumptions on the smoothness class of T (used to analyze deterministic methods). This is because local surrogate models also need the estimation of T from the SO. In that respect, we will consider Lemma 1.2, which is a consequence of the following assumption.

Assumption 1.1 (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 $\mathbb{E}\left[\|F(\xi, x_*)\|^{\mathsf{a}p}\right] < \infty$ and $\mathbb{E}\left[\mathsf{L}(\xi)^{\mathsf{a}p}\right] < \infty$. We define $L := \mathbb{E}[\mathsf{L}(\xi)]$ and $L_q := \sqrt[q]{\mathbb{E}[\mathsf{L}(\xi)^q]} + L$ for any $q > 0$.

Lemma 1.2 (Hölder continuity of the mean and the standard deviation). *If Assumption* 1.1 *holds, then T is (L, δ)-Hölder continuous on X and $\sigma_q(\cdot)$ is (L_q, δ)-Hölder continuous on X for any $q \in [p, \mathsf{a}p]$.*

In the above, we say T is (L, δ)-Hölder continuous if $\|T(x) - T(y)\| \leq L\|x - y\|^\delta$ for all $x, y \in X$. Lemma 1.2 is a simple consequence of Jensen and Minkowski's inequalities, and hence the proof is omitted. In this paper, we shall only consider the Euclidean norm $\|\cdot\|$. Assumption 1.1 is standard in stochastic optimization [31]. With respect to sampling, our statistical analysis will only require the standard

assumption of an *unbiased oracle* (UO) *with i.i.d. sampling*. 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 $\mathbf{P}$: for all $x \in X$,

$$\widehat{F}(\xi^N, x) := \sum_{j=1}^N \frac{F(\xi_j, x)}{N}, \quad \widehat{\epsilon}(\xi^N, x) := \sum_{j=1}^N \frac{\epsilon(\xi_j, x)}{N}, \quad \widehat{L}(\xi^N) := \sum_{j=1}^N \frac{\mathsf{L}(\xi_j)}{N}. \tag{3}$$

1.1. Related work, proposed methods, and contributions. The performance of first-order methods strongly depends on the *stepsize sequence*. As an example, a classical method to solve a smooth problem $\min_{\mathbb{R}^d} f$ is the gradient method $x^{k+1} := x^k - \alpha_k \nabla f(x^k)$, where $\{\alpha_k\}$ is a positive stepsize sequence. One choice of stepsizes that guarantees its convergence is the *small stepsize policy* (SSP): $\sum_k \alpha_k = \infty$ and $\sum_k \alpha_k^2 < \infty$, a typical choice being $\alpha_k = \mathcal{O}(k^{-1})$. If L is the Lipschitz constant of $\nabla f(\cdot)$, the *constant stepsize policy* (CSP) $\alpha_k = \mathcal{O}(\frac{1}{L})$ has a provable accelerated convergence rate in comparison to the SSP since its stepsize sequence does not vanish. However, the latter has the advantage of *not requiring an estimate of* L, and, in this sense, it is a more *robust* and practical policy since L is rarely known. A significant improvement is the use of *line search schemes* which build endogenous *adaptive* stepsizes bounded away from zero at the expense of a few more gradient evaluations. A standard example is *Armijo's line search* rule [2]. This policy enjoys the accelerated convergence of the CSP and the robustness of the SSP with respect to an unknown L. For VIs, see [17, 13]. As expected, the stepsize policy also affects the performance of SA methods. In [28] and in later developments, it is shown that the SSP is sufficient to progressively *reduce the variance* of the oracle's error trajectory and hence obtain convergence. How to handle such variance efficiently is still an active research subject motivated by problems of machine learning and SP.

The performance of an SA method can be measured by its *iteration complexity* (IC) and *oracle complexity* (OC), given a tolerance $\epsilon > 0$ with respect to a suitable metric. The first is the total number of iterations, a measure for the optimization error, while the second is the total number of samples and oracle calls, a measure for the estimation error. Statistical lower bounds [1] show that the class of smooth convex functions has an optimal OC of $\mathcal{O}(\epsilon^{-2})$ in terms of the optimality gap. A fundamental improvement with respect to estimation error was Polyak and Ruppert's *iterate averaging* (IA) scheme (see, e.g., [27]). This scheme replaces the SSP by *longer stepsizes* $\alpha_k = \mathcal{O}(k^{-\frac{1}{2}})$ with a subsequent final *ergodic average* of the iterates using the stepsizes as weights. *If one oracle call per iteration is postulated*, such a scheme obtains optimal IC and OC of $\mathcal{O}(\epsilon^{-2})$ on the class of smooth convex functions. This is also the size of the final ergodic average, a measure of the additional *averaging effort* implicitly required in IA schemes. Such methods, hence, are efficient in terms of OC. IA was then extensively explored (see, e.g., [25, 15, 4] and references therein). The important work [25] studies the *robustness* of IA in SA methods and shows that such schemes can outperform the SAA approach on relevant convex problems. In [3], IA is shown to be a robust policy on the strongly convex class.

As mentioned above, IA attains optimal IC and OC under the constraint that *the oracle is called once per iteration*. In case the *oracle can be called multiple times*, a question that remains open is the following: Can we obtain an *improved* IC with (approximately) the same *optimal* OC of IA? If yes, one could construct an SA method which is faster but still efficient in terms of sample complexity. In this sense, a rapidly growing line of research proposes SA methods with *variance reduction* (VR) using more than one oracle call per iteration to alleviate the role of the stepsize in reducing

variance. Two representative examples include *gradient aggregation* (GA) *methods* and *dynamic sampling* (DS) *methods*. See section 5 in [4]. These methods can use a CSP and thus obtain an accelerated rate of convergence when compared to the IA scheme. Designed for finitely supported distributions with bounded data, GA methods reduce the variance by combining in a specific manner eventual exact computation of gradients and eventual IA with frequent gradient sampling. See, e.g., [4] and references therein. Designed to solve problems with an arbitrary distribution and online data acquisition (as is the case in many stochastic and simulation optimization problems based on Monte Carlo methods), DS methods reduce variance by estimating the gradient via an *empirical average* associated to a sample whose size (*minibatch*) is increased at every iteration. See, e.g., [6, 11] and references therein. An essential point is that current GA and DS methods achieve, up to constants, the order of the deterministic optimal IC with the *same* (near) optimal OC and averaging effort of IA schemes. *In this sense*, GA and DS methods can be more efficient options than IA.

We now comment on the main purpose of this work. All VR methods mentioned above still use a CSP $\alpha_k = \mathcal{O}(\frac{1}{L})$ *assuming knowledge of the Lipschitz constant*. Hence, although they improve the convergence of SA methods, IA with $\alpha_k = \mathcal{O}(k^{-\frac{1}{2}})$ is still a more robust policy when L is unknown or poorly estimated [25, 3]. In this setting, current VR methods may be impractical. An important question is the following: Can *faster rates of convergence* with (near) *optimal* OC be accomplished by *robust* VR methods? By robust variance reduction we mean the use of adaptive schemes that avoid exogenous estimation of L and produce a stepsize sequence bounded away from zero. Motivated by line search schemes in deterministic methods, our aim is to propose line search schemes for a class of dynamic sampled stochastic approximated (DS-SA) methods. In this paper, we focus on SVIs and pursue an improved complexity analysis of stochastic optimization problems in future research. To the best of our knowledge, line search schemes for SVIs are currently nonexistent. Even for stochastic optimization, considering that Robbins and Monro's seminal work was published in 1951, it seems that only very few existing works treat adaptive stepsize search schemes for SA methods with *stepsizes bounded away from zero* [21, 22, 30, 24, 33, 34, 19, 20]. As explained in the following along our contributions, either these works do not provide a convergence theory, or the one presented is unsatisfactory in the context of SA-line search schemes. For completeness, we mention a different approach taken by some authors [35, 36] using smoothing techniques [32] to deal with the absence of the Lipschitz constant. It should be noted, however, that the authors of [35, 36] do not present VR methods. In fact, they are concerned with the absence of Lipschitz continuity of the operator, a setup which already possesses slower rates in the deterministic case. We, on other hand, assume Lipschitz continuity of the operator with no knowledge of the Lipschitz constant. In this setup, our objective is to obtain faster rates via efficient and robust VR techniques.

Before presenting our contributions, it is important to explain why the analysis of line search schemes in SA methods is intrinsically *different* and *more difficult* than in the deterministic case. This may explain the absence of a satisfying convergence theory of SA methods with line search schemes which (1) do not use knowledge of L, (2) obtain stepsizes bounded away from zero, and (3) only assume a UO. Since [28], it is well known that the analysis of SA methods strongly relies on *martingale processes*. Such a martingale-like property is obtained by (a) *an optimization process*: the iterative algorithm satisfies a fixed-point contraction or Lyapunov principle; (b) *an estimation process*: in standard SA methods, a *fresh i.i.d. sample* is updated at every

iteration; and (c) *exogenous stepsize policies*: for instance, the SSP $\alpha_k = \mathcal{O}(k^{-1})$, longer stepsizes $\alpha_k = \mathcal{O}(k^{-\frac{1}{2}})$ with IA, and the CSP $\alpha_k = \mathcal{O}(\frac{1}{L})$. As an example, consider the SG method with exogenous stepsizes satisfying $0 < \sup_k \alpha_k < \frac{1}{2L}$. Given a solution $x^* \in \operatorname{argmin}_{x\in\mathbb{R}^d}\{f(x) = \mathbb{E}[G(\xi, x)]\}$, it is possible to obtain the recursion

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - [1 - L\mathcal{O}(\alpha_k)]\mathcal{O}(\alpha_k^2)\|\nabla f(x^k)\|^2 + \Delta M_k + \Delta A_k, \tag{4}$$

where $\epsilon^k := \nabla G(\xi^k, x^k) - \nabla f(x^k)$ is the oracle error at the kth iterate, $\Delta M_k := \mathcal{O}(\alpha_k)\langle\epsilon^k, x^* - x^k\rangle$, and $\Delta A_k := \mathcal{O}(\alpha_k^2)\|\epsilon^k\|^2$. The above relation and i.i.d. sampling imply that the *optimization error* sequence $\{\|x^k - x^*\|^2\}$, running over the *iteration time-scale*, defines a "perturbed" *supermartingale* sequence (see Theorem 2.3). By i.i.d. sampling, the gradient *estimation error* sequence $\{\epsilon^k\}$, running over the *estimation time-scale*, defines an *exact martingale difference*, i.e., $\mathbb{E}[\epsilon^k|\mathcal{F}_k] = 0$ for all k. Here $\mathcal{F}_k := \sigma(\xi^0, \ldots, \xi^{k-1})$ is the information collected before iteration k. Similarly, by the tower law of conditional expectations, $\{\Delta M_k\}$ is also an exact martingale difference. To conclude, the quadratic error $\mathbb{E}[\Delta A_k|\mathcal{F}_k]$ can also be handled successfully by using the fact that $\{\epsilon^k\}$ is a martingale difference.

If one considers *adaptive endogenous stepsizes*, a natural choice would be an SA *version of Armijo's rule* (SA-AR): choose α_k as the maximum $\alpha \in \{\theta^\ell\hat{\alpha} : \ell \in \{0\}\cup\mathbb{N}\}$ such that $\widehat{G}\left(\xi^k, x^k(\alpha)\right) - \widehat{G}(\xi^k, x^k) \leq \lambda\langle\nabla\widehat{G}(\xi^k, x^k), x^k(\alpha) - x^k\rangle$, where $\hat{\alpha} \in (0, 1]$, $\theta, \lambda \in (0, 1)$, $\xi^k := \{\xi_j^k\}_{j=1}^{N_k}$ is an i.i.d. sample from $\mathbf{P}$ such that $N_k \to \infty$, and, for all $\alpha > 0$, $x^k(\alpha) := x^k - \alpha\nabla\widehat{G}(\xi^k, x^k)$. In the above, $\widehat{G}(\xi^k, x^k)$ and $\nabla\widehat{G}(\xi^k, x^k)$ denote, respectively, the empirical averages of $G(\cdot, x^k)$ and $\nabla G(\cdot, x^k)$ with respect to the sample ξ^k (note that VR is necessary in order to obtain nonvanishing stepsizes). The challenging aspect of the above scheme is highlighted:

(A): DS-SA *line search schemes intrinsically introduce nonmartingale-like dependencies even when using i.i.d. sampling.*

To see this, first note that the *backtracking* scheme of the SA-AR examines the variation of $\widehat{G}(\xi^k, \cdot)$ along a discrete path $\alpha \mapsto x^k(\alpha)$ so that the chosen stepsize α_k and accepted iterate $x^{k+1} := x^k(\alpha_k)$ are both measurable functions of (ξ^k, x^k). Second, by using the contraction principle produced by the SA-AR, we are forced to estimate the oracle error $\widehat{\epsilon}(\xi^k, x^{k+1}) = \widehat{G}(\xi^k, x^{k+1}) - f(x^{k+1})$. More precisely, it may be shown that an additional quadratic term of the form $\mathcal{O}(\alpha_k^2)\|\widehat{\epsilon}(\xi^k, x^{k+1})\|^2$ needs to be handled in the right-hand side of the recursion (4). The key point is that $\widehat{\epsilon}(\xi^k, x^{k+1})$ *is not a martingale difference*: it is a measurable function of the *coupled* variables ξ^k and x^{k+1} due to backtracking. Even when ξ^k is an i.i.d. sample of $\mathbf{P}$, this coupling is inevitable, and hence the desired convergence

$$\lim_{k\to\infty} \widehat{\epsilon}(\xi^k, x^{k+1}) = \lim_{k\to\infty} \sum_{j=1}^{N_k} \frac{G(\xi_j^k, x^{k+1}) - f(x^{k+1})}{N_k} = 0, \tag{5}$$

either in the almost sure sense or in distribution, *does not* follow from the standard strong law of large numbers or the central limit theorem: the above sum *is not a sum of independent random variables*. The nontrivial aspect here is that an SA method with line search has *two* statistical estimation processes: the gradient estimation of item (b) above *and* the estimation of a lower bound for L replacing (c). In this sense, DS-SA methods with line search schemes are statistically harder than standard SA methods. We finally remark that in all the works [34, 20, 19] the convergence (5) is *postulated*, putting aside the challenging aspect in (A). Thus, their assumptions are far beyond the usual assumption of a UO. Errors of the type $\widehat{\epsilon}(\xi^k, x^{k+1})$ will be

referred to as *correlated errors*. One of the contributions of this paper is to propose the use of empirical process theory in order to handle the nonmartingale sequence $\{\widehat{\epsilon}(\xi^k, x^{k+1})\}$ successfully while exploiting the robustness property of SA-line search schemes.

In this paper, we propose Algorithm 1 to solve SVIs via the SA methodology without requiring knowledge of the Lipschitz constant. In the following, Π denotes the Euclidean projection onto X and $\{N_k\}$ is the sample rate, i.e., the sample size used in the kth iteration to compute the sample mean operator at the current iterates. See (3) and (9)–(10). Our contributions are summarized in the following.

Algorithm 1 DS-SA-extragradient method with a DS-SA line search scheme

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

$$d^k := \begin{cases} 0 & \text{if } \|r^k\| > 0, \\ \delta_k \frac{\hat{x}^k - x^k}{\|\hat{x}^k - x^k\|} \text{ for any } \hat{x}^k \in X \text{ such that } \hat{x}^k \neq x^k & \text{if } \|r^k\| = 0. \end{cases} \tag{6}$$

LINE SEARCH RULE: Define α_k as the maximum $\alpha \in \{\theta^\ell \hat{\alpha} : \ell \in \{0\} \cup \mathbb{N}\}$ such that

$$\alpha \left\| \widehat{F}\left(\xi^k, z^k(\alpha)\right) - \widehat{F}\left(\xi^k, x^k + d^k\right) \right\| \leq \lambda \|z^k(\alpha) - (x^k + d^k)\|, \tag{7}$$

where, for all $\alpha \in (0, \hat{\alpha}]$,

$$z^k(\alpha) := \Pi\left[x^k + d^k - \alpha\left(\widehat{F}(\xi^k, x^k) + \beta d^k\right)\right], \tag{8}$$

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

$$z^k = \Pi\left[x^k - \alpha_k \widehat{F}(\xi^k, x^k)\right], \tag{9}$$

$$x^{k+1} = \Pi\left[x^k - \alpha_k \widehat{F}(\eta^k, z^k)\right]. \tag{10}$$

(i) *Robust VR with efficient oracle complexity and multiplicative noise*: To the best of our knowledge, Algorithm 1 is the first provable *robust* SA method *with VR*, either for SVIs or SPs, with improved IC and near optimal OC. This means that we obtain, up to constants, an optimal IC of $\mathcal{O}(\epsilon^{-1})$ and near optimal OC of $\mathcal{O}(\epsilon^{-2})$ (up to log factors on ϵ and L) in the large sample setting for SVIs without the knowledge of the Lipschitz constant L. Previous *nonrobust* VR methods use the policy $\alpha_k = \mathcal{O}(\frac{1}{L})$ and obtain, up to constants, the same complexities [11] but require an exogenous estimate of L. Such an estimate is often nonexistent in practice. If existent but of a poor quality, it implies a slow convergence. On the other hand, previous robust methods use a *vanishing stepsize policy* with the poorer IC of $\mathcal{O}(\epsilon^{-2})$ for ill-conditioned problems [25]. Concerning line search schemes, they are nonexistent for SVIs. It seems our results are also new for SPs (seen as a particular SVI): all current methods either (1) are not supported by a convergence theory [21, 22], (2) still use the knowledge of L and other parameters or use the SSP [30, 24, 33] (and, hence, have a slower IC and are not robust VR methods), or (3) postulate (5) without giving complexity estimates [34, 20, 19]. Condition (5) puts aside the challenging phenomenon in (A), and it is a much stronger assumption than the standard unbiased oracle, a sufficient assumption

for our analysis. Previous robust methods assume a *compact* X and a global *uniform bound* on the oracle's variance (UBOV), i.e., the existence of some $\sigma > 0$, such that $\sup_{x\in X} \sigma_2(x)^2 \leq \sigma^2$. Our robust method is valid under Assumption 1.1. This includes problems with an unbounded X and an oracle satisfying $\sup_{x\in X} \sigma_2(x)^2 = \infty$, such as unconstrained quadratic SPs and affine SVIs with a *random* matrix. Even if the UBOV holds, it does not lead to the sharpest estimates since typically $\sigma_2(x^*)^2 \ll \sigma^2$ for $x^* \in X^*$ (see Example 3.9 in [11]). Our bounds are local in the sense that they depend on variance at solutions, L, and initial iterates (but neither on a global variance upper bound nor, necessarily, on the diameter of X). Compared to nonrobust VR methods [6, 11], a price to pay in our estimates for not having an estimate of L is that the OC of Algorithm 1 has an additional factor of $\ln(L)\mathcal{O}(d)$. We note, however, that such an upper bound is tight in comparison to the sample complexity of the general SAA estimator which does not assume extra information on the problem (see, e.g., Theorem 5.18 in [31]). We refer the reader to Theorem 3.18 and Corollary 3.19.[2]

(ii) *An iterative local empirical process theory for DS-SA methods*: as mentioned before, DS-SA line search schemes intrinsically introduce nonmartingale-like processes, a fact that has not been properly handled before. Going beyond standard martingale techniques used in SA methods with exogenous stepsize policies, we use a novel *iterative localization* argument based on advanced techniques from *empirical process theory* [5, 26] to analyze correlated errors introduced in SA-line search schemes. Very importantly, we *do not* postulate significantly narrower oracle conditions such as (5) used in [34, 20, 19]; that is, our statistical analysis is solely based on the standard i.i.d. sampling assumption. We refer the reader to section 4 for a detailed description. This is the most sensible part of our work and the cornerstone tool. Our analysis also sets the ground for potential generalizations to other adaptive algorithms based on the SA methodology, such as, e.g., stochastic trust-region methods. In a nutshell, our proposition is to *locally* decouple the dependency in the correlated error up to the control of an empirical process indexed over a suitable ball centered at the current iterate. The intuition here is that the iterate generated after the line search scheme, although highly dependent on the fresh i.i.d. sample, lies at a ball whose radius is dependent on previous information and on a martingale difference error. We refer the reader to section 4 for further details.

In section 2, we give some preliminaries. The convergence theory of Algorithm 1 is presented in section 3. We give particular attention to subsection 3.2, where the main tool of empirical process theory of item (ii) above (Theorem 3.11) is stated and applied to the convergence analysis of Algorithm 1. The proof of Theorem 3.11 is presented in section 4. This paper concludes with a discussion in section 5 and with the proof of a technical lemma in the appendix.

2. Preliminaries and notation. For $x, y \in \mathbb{R}^d$, we denote by $\langle x, y\rangle$ the standard inner product and by $\|x\| = \sqrt{\langle x, x\rangle}$ the correspondent Euclidean norm. Given $C \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$, we use the notations $\mathrm{d}(x, C) := \inf\{\|x - y\| : y \in C\}$, $\mathcal{D}(C) := \sup\{\|x - y\| : x, y \in C\}$, and $\Pi_C(x) := \operatorname{argmin}_{y\in C} \|y - x\|^2$. Given $H : \mathbb{R}^d \to \mathbb{R}^d$, $\mathrm{S}(H, C)$ denotes the solution set of $\mathrm{VI}(H, C)$. The following properties are well known (see, e.g., [8] and [7, Proposition 4.1]).

LEMMA 2.1. *Take a closed and convex set $C \subset \mathbb{R}^d$.*

[2]Our complexities hold for the quadratic natural residual or the D-gap function (see section 2 and [8, Theorems 10.2.3 and 10.3.3 and Proposition 10.3.7]). If X is compact, our method achieves the same complexities, up to constants, in terms of the dual-gap function (see, e.g., [25, 7]).

(i) *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 $x, y \in \mathbb{R}^d$, $\|\Pi_C(x) - \Pi_C(y)\| \leq \|x-y\|$.*
(iii) *Given $H : \mathbb{R}^d \to \mathbb{R}^d$, $\mathrm{S}(H, C) = \{x \in \mathbb{R}^d : x = \Pi_C[x - H(x)]\}$.*

For X as in (1), we use the notation $\Pi := \Pi_X$. Given an operator $H : X \to \mathbb{R}^d$ and $\alpha > 0$, the *natural residual function* is the map

$$r_\alpha(H; x) := \|x - \Pi\left[x - \alpha H(x)\right]\|. \tag{11}$$

For T as in (1), we use the notations $r_\alpha := r_\alpha(T, \cdot)$, $r(H; \cdot) := r_1(H; \cdot)$, and $r := r_1$. We shall need the following lemma (see [8, Proposition 10.3.6]).

LEMMA 2.2. *Given $x \in \mathbb{R}^d$, the function $(0, \infty) \ni \alpha \mapsto \frac{r_\alpha(H,x)}{\alpha}$ is nonincreasing.*

Given sequences $\{x^k\}$ and $\{y^k\}$, we use the notation $x^k = \mathcal{O}(y^k)$ or $\|x^k\| \lesssim \|y^k\|$ to mean that there exists a constant $C > 0$ such that $\|x^k\| \leq C\|y^k\|$ for all k. The notation $\|x^k\| \sim \|y^k\|$ means that $\|x^k\| \lesssim \|y^k\|$ and $\|y^k\| \lesssim \|x^k\|$. Given a σ-algebra $\mathcal{F}$ and a random variable ξ, we denote by $\mathbb{E}[\xi]$, $\mathbb{E}[\xi|\mathcal{F}]$, and $\mathbb{V}[\xi]$ the expectation, conditional expectation, and variance, respectively. Given $p \geq 1$, $|\xi|_p$ is the $\mathcal{L}^p$-norm of ξ and $|\xi\,|\mathcal{F}]_p := \sqrt[p]{\mathbb{E}\left[|\xi|^p\,|\mathcal{F}\right]}$ is the $\mathcal{L}^p$-norm of ξ conditional to $\mathcal{F}$. We denote by $\sigma(\xi_1, \ldots, \xi_k)$ the σ-algebra generated by the random variables $\{\xi_i\}_{i=1}^k$ and $\mathbb{E}[\cdot|\xi_1, \ldots, \xi_k] := \mathbb{E}[\cdot|\sigma(\xi_1, \ldots, \xi_k)]$. Given two σ-algebras $\mathcal{F}$ and $\mathcal{G}$, $\sigma(\mathcal{F} \cup \mathcal{G})$ will denote the σ-algebra generated by $\mathcal{F} \cup \mathcal{G}$. We write a.s. for "almost surely," $\xi \in \mathcal{F}$ for "ξ is $\mathcal{F}$-measurable," $\xi \perp\!\!\!\perp \mathcal{F}$ for "ξ is independent of $\mathcal{F}$," and $\mathbf{1}_A$ for the characteristic function of a set $A \in \mathcal{F}$. Given $x, y \in \mathbb{R}$, $\lceil x \rceil$ denotes the smallest integer greater than x, $x \vee y := \max\{x, y\}$ and $x \wedge y := \min\{x, y\}$. $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, and, for $m \in \mathbb{N}$, we use the notation $[m] = \{1, \ldots, m\}$. $|\mathcal{V}|$ denotes the cardinality of a set $\mathcal{V}$, $\mathbb{B}$ denotes the Euclidean unit ball, and $\mathbb{B}[x, r]$ denotes the Euclidean ball with center x and radius $r > 0$. Recall the following theorem for perturbed supermartingales of Robbins and Siegmund (see, e.g., Theorem 2.2 in [11]).

THEOREM 2.3. *Let $\{y_k\}, \{u_k\}, \{a_k\}, \{b_k\}$ be sequences of nonnegative random variables, adapted to the filtration $\{\mathcal{F}_k\}$, such that a.s. $\sum a_k < \infty$, $\sum b_k < \infty$, and for all $k \in \mathbb{N}$, $\mathbb{E}\left[y_{k+1}\middle|\mathcal{F}_k\right] \leq (1 + a_k)y_k - u_k + b_k$. Then a.s. $\{y_k\}$ converges and $\sum u_k < \infty$.*

3. Convergence analysis of Algorithm 1. We state next additional assumptions needed for the convergence analysis of Algorithm 1. In this section, we always assume that in Assumption 1.1 we have $\delta = 1$. For brevity, we will not mention it any further.

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

Assumption 3.2 (pseudomonotonicity). We assume that $T : X \to \mathbb{R}^d$ as defined in (1) is pseudomonotone: for all $z, x \in X$, $\langle T(x), z - x\rangle \geq 0 \Longrightarrow \langle T(z), z - x\rangle \geq 0$.

Pseudomonotonicity includes monotonicity as a special class [15, 7]. Pseudomonotone SVIs were also considered in [11, 16]. In these works, knowledge of parameters such as the Lipschitz constant are still assumed and no VR schemes are presented in [16]. 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 pseudomonotone gradient. Recall the notation $[N_k] := \{1, \ldots, N_k\}$.

Assumption 3.3 (i.i.d. sampling). In Algorithm 1, the sequences $\{\xi_j^k : k \in \mathbb{N}_0,\, j \in [N_k]\}$ and $\{\eta_j^k : k \in \mathbb{N}_0,\, j \in [N_k]\}$ 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 1, 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 (2), (3), and Algorithm 1, we will define the following oracle errors:

$$\epsilon_1^k := \widehat{\epsilon}(\xi^k, x^k), \qquad \epsilon_2^k := \widehat{\epsilon}(\eta^k, z^k), \qquad \epsilon_3^k := \widehat{\epsilon}(\xi^k, z^k). \tag{12}$$

Assumption 3.3 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 a 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 3.4 (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 1. By the definition of d^k in (6) and convexity of X, we have that, for all $k \in \mathbb{N}$,

$$\|d^k\| \le \delta_k, \quad x^k + d^k \in X. \tag{13}$$

Moreover, it can be shown that if $\beta \in (0, \hat{\alpha}^{-1}]$, then, for all $\alpha \in (0, \hat{\alpha}]$ and $k \in \mathbb{N}$,

$$\|z^k(\alpha) - (x^k + d^k)\| > 0, \tag{14}$$

where $z^k(\alpha)$ is defined in (8) (see the proof of Lemma 3.6 in the next section). In fact, the rule (6) chosen to update d^k could be replaced by any rule satisfying (13)–(14).

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 1), x^k is an exact solution and we can stop the algorithm. In our framework, we use a sample-based line search scheme so that the termination criteria are generally not clear. By choosing β, $\{\delta_k\}$, and d^k as above, the sample-based line search rule (7)–(8) is always clearly specified and terminates in a finite number of iterations. The direction d^k serves merely as a small perturbation to address the case $r^k = 0$. Since $\|d^k\| \le \delta_k$ holds for all k, we can set $\delta_k \to 0$ in any desired rate so as to correct iteratively such small perturbations. In this way, the optimality of the iteration and the oracle complexities of Algorithm 1 are unaltered. We refer the reader to the convergence analysis in the next section for further details.

Remark 3.5 (intuition for the line search scheme). The stochastic approximated line search (7) is motivated by [17]. We make some comments for the case $d^k = 0$ (see Remark 3.4). Using definition (11), inequality (7) can be rewritten as

$$\left\| \widehat{F}\left(\xi^k, z^k(\alpha)\right) - \widehat{F}\left(\xi^k, x^k\right) \right\| \le \lambda \frac{r_\alpha(H_k; x^k)}{\alpha}, \tag{15}$$

where $H_k := \widehat{F}(\xi^k, \cdot)$. Provided that $r_{\hat{\alpha}}(H_k; x^k) \neq 0$, the line search tests (15) for decreasing $\alpha \in (0, \hat{\alpha}]$. The idea is that the right-hand side of (15) does not increase by Lemma 2.2 while the left-hand side tends to 0 by continuity of the operator. Hence, (15) will hold eventually.

3.1. Derivation of an error bound. In this section, we show that Algorithm 1 is well defined and, given some $x^* \in X^*$, we derive a recursive bound for the iteration error sequence $\{\|x^k - x^*\|^2\}$.

LEMMA 3.6 (finite termination of the line search). *Consider Assumption* 1.1. *Then the line search* (7) *in the iteration* k *of Algorithm* 1 *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 (8) and definition (11), 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 (6), 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 (6), 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 2.1(ii) in the last inequality and $0 < \hat{\alpha}\beta \leq 1$ in the last equality.

We now conclude the proof of the lemma. Set $\gamma_\ell := \theta^{-\ell}\hat{\alpha}$. Assuming by contradiction that the line search (7) 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 the definition of $r_\alpha(H_k; \cdot)$ in (11), the fact that $\gamma_\ell \in (0, \hat{\alpha}]$, and Lemma 2.2 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 1.1, 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 (7) either chooses the initial stepsize $\hat{\alpha}$ or it is a 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 3.7 (unbiased lower estimation of the Lipschitz constant). *Consider Assumptions* 1.1 *and* 3.3. *Then* $\alpha_k \geq (\frac{\lambda\theta}{\widehat{L}(\xi^k)}) \wedge \hat{\alpha}$ *a.s. and* $\left|\alpha_k\middle|\mathcal{F}_k\right|_2 \cdot |\mathsf{L}(\xi)|_2 \geq (\lambda\theta)\wedge\hat{\alpha}$.

Proof. If $\hat{\alpha}$ satisfies (7), then $\alpha_k = \hat{\alpha}$. Otherwise, we have

$$\theta^{-1}\alpha_k \left\|\widehat{F}\left(\xi^k, z^k(\theta^{-1}\alpha_k)\right) - \widehat{F}(\xi^k, x^k + d^k)\right\| > \lambda \left\|z^k\left(\theta^{-1}\alpha_k\right) - (x^k + d^k)\right\|. \tag{16}$$

Assumption 1.1 and the definition of $\widehat{F}(\xi^k, \cdot)$ in (3) imply that

$$\left\|\widehat{F}\left(\xi^k, z^k(\theta^{-1}\alpha_k)\right) - \widehat{F}(\xi^k, x^k + d^k)\right\| \leq \widehat{L}(\xi^k) \left\|z^k\left(\theta^{-1}\alpha_k\right) - (x^k + d^k)\right\|. \tag{17}$$

The fact that $z^k\left(\theta^{-1}\alpha_k\right) \neq x^k + d^k$ and (16)–(17) 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

$$\begin{aligned}
(\lambda\theta) \wedge \hat{\alpha} &\leq \mathbb{E}\left[\alpha_k \widehat{L}(\xi^k)\middle|\mathcal{F}_k\right] \\
\text{(by Hölder's inequality)} \quad &\leq \left|\alpha_k\middle|\mathcal{F}_k\right|_2 \cdot \left|\widehat{L}(\xi^k)\middle|\mathcal{F}_k\right|_2 \\
\text{(by convexity of } t \mapsto t^2) \quad &\leq \left|\alpha_k\middle|\mathcal{F}_k\right|_2 \sqrt{\frac{1}{N_k}\sum_{j=1}^{N_k} \mathbb{E}\left[\mathsf{L}(\xi_j^k)^2\middle|\mathcal{F}_k\right]} = \left|\alpha_k\middle|\mathcal{F}_k\right|_2 \cdot |\mathsf{L}(\xi)|_2,
\end{aligned}$$

using Assumption 3.3 in the last equality. ∎

Recall (3) and (12). We define, for $k \in \mathbb{N}_0$ and for $x^* \in X^*$,

$$\Delta A_k := (1 - 8\lambda^2)\hat{\alpha}^2\|\epsilon_1^k\|^2 + 8\hat{\alpha}^2\|\epsilon_2^k\|^2 + 8\hat{\alpha}^2\|\epsilon_3^k\|^2, \tag{18}$$

$$\Delta M_k(x^*) := 2\alpha_k\langle x^* - z^k, \epsilon_2^k\rangle, \tag{19}$$

$$\Delta P_k := 8(2 - \alpha_k\beta)^2[\lambda + \alpha_k\widehat{L}(\xi^k)]^2\delta_k^2. \tag{20}$$

We refer the reader to the definition $r := r_1(T;\cdot)$ in (11).

LEMMA 3.8 (a recursive error bound for Algorithm 1). *Consider Assumptions* 1.1 *and* 3.1–3.2. *The sequence generated by Algorithm* 1 *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 3.8. We divide the proof into two parts. The first uses the extragradient step (9)–(10). The second uses the line search (7)–(8) with some judicious error bounds.

Part 1 (extragradient step). By (9)–(10), we invoke twice Lemma 2.1(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 $x \in X$,

$$2\langle\alpha_k\widehat{F}(\xi^k, x^k), z^k - x\rangle \leq \|x^k - x\|^2 - \|z^k - x\|^2 - \|z^k - x^k\|^2, \tag{21}$$

$$2\langle\alpha_k\widehat{F}(\eta^k, z^k), x^{k+1} - x\rangle \leq \|x^k - x\|^2 - \|x^{k+1} - x\|^2 - \|x^{k+1} - x^k\|^2. \tag{22}$$

We now set $x := x^{k+1}$ in (21) and sum the obtained relation with (22) 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 (2), (3), and (12), 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 $\in X$,

$$\begin{aligned}2\alpha_k\langle T(z^k), z^k - x\rangle &\le 2\alpha_k\langle \widehat{F}(\eta^k, z^k) - \widehat{F}(\xi^k, x^k), z^k - x^{k+1}\rangle + 2\alpha_k\langle \epsilon_2^k, x - z^k\rangle \\ &\quad + \|x^k - x\|^2 - \|x^{k+1} - x\|^2 - \|z^k - x^{k+1}\|^2 - \|z^k - x^k\|^2 \\ &\le 2\alpha_k\|\widehat{F}(\eta^k, z^k) - \widehat{F}(\xi^k, x^k)\|\|z^k - x^{k+1}\| + 2\alpha_k\langle \epsilon_2^k, x - z^k\rangle \\ &\quad + \|x^k - x\|^2 - \|x^{k+1} - x\|^2 - \|z^k - x^{k+1}\|^2 - \|z^k - x^k\|^2 \\ &\le 2\alpha_k^2\|\widehat{F}(\eta^k, z^k) - \widehat{F}(\xi^k, x^k)\|^2 + 2\alpha_k\langle \epsilon_2^k, x - z^k\rangle \\ &\quad + \|x^k - x\|^2 - \|x^{k+1} - x\|^2 - \|z^k - x^k\|^2, \end{aligned} \tag{23}$$

where we used Cauchy–Schwarz in the second inequality and Lemma 2.1(ii) with (9)–(10) 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 (8). We first note that, by (8)–(9), Lemma 2.1(ii), $0 < \alpha_k\beta \le \hat{\alpha}\beta \le 1$, and $\|d^k\| \le \delta_k$,

$$\|\tilde{z}^k - z^k\| \le \|d^k - \alpha_k\beta d^k\| \le (1 - \alpha_k\beta)\delta_k, \quad \|\tilde{x}^k - x^k\| \le \delta_k. \tag{24}$$

Recall that, according to (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 (23), we have, by the triangle inequality,

$$\begin{aligned}\alpha_k\|\widehat{F}(\eta^k, z^k) - \widehat{F}(\xi^k, x^k)\| &\le \alpha_k\|\widehat{F}(\eta^k, z^k) - \widehat{F}(\xi^k, \tilde{z}^k)\| + \alpha_k\|\widehat{F}(\xi^k, \tilde{z}^k) - \widehat{F}(\xi^k, \tilde{x}^k)\| \\ &\quad + \alpha_k\|\widehat{F}(\xi^k, \tilde{x}^k) - \widehat{F}(\xi^k, x^k)\|. \end{aligned} \tag{25}$$

The first term above can be bounded as

$$\begin{aligned}\alpha_k\|\widehat{F}(\eta^k, z^k) - \widehat{F}(\xi^k, \tilde{z}^k)\| &\le \alpha_k\|\widehat{F}(\eta^k, z^k) - T(z^k)\| + \alpha_k\|\widehat{F}(\xi^k, z^k) - T(z^k)\| \\ &\quad + \alpha_k\|\widehat{F}(\xi^k, z^k) - \widehat{F}(\xi^k, \tilde{z}^k)\| \\ &\le \alpha_k\|\epsilon_2^k\| + \alpha_k\|\epsilon_3^k\| + \alpha_k\widehat{L}(\xi^k)\|z^k - \tilde{z}^k\| \\ &\le \alpha_k\|\epsilon_2^k\| + \alpha_k\|\epsilon_3^k\| + \alpha_k\widehat{L}(\xi^k)(1 - \alpha_k\beta)\delta_k, \end{aligned} \tag{26}$$

using the triangle inequality in the first inequality, Assumption 1.1 and the definitions in (2), (3), and (12) in the second inequality, and (24) in the last inequality. Similarly, the third term in (25) satisfies

$$\alpha_k\|\widehat{F}(\xi^k, \tilde{x}^k) - \widehat{F}(\xi^k, x^k)\| \le \alpha_k\widehat{L}(\xi^k)\|\tilde{x}^k - x^k\| \le \alpha_k\widehat{L}(\xi^k)\delta_k. \tag{27}$$

Finally, from the line search (7)–(8) and (24), the second term in (25) satisfies

$$\begin{aligned}\alpha_k\|\widehat{F}(\xi^k, \tilde{z}^k) - \widehat{F}(\xi^k, \tilde{x}^k)\| &\le \lambda\|\tilde{z}^k - \tilde{x}^k\| \\ &\le \lambda\|\tilde{z}^k - z^k\| + \lambda\|z^k - x^k\| + \lambda\|x^k - \tilde{x}^k\| \\ &\le \lambda\|z^k - x^k\| + \lambda(2 - \alpha_k\beta)\delta_k. \end{aligned} \tag{28}$$

Putting together (25)–(28), squaring, using the fact that $(\sum_{i=1}^4 a_i)^2 \le 4\sum_{i=1}^4 a_i^2$, and using definition (20), we obtain

$$2\alpha_k^2\|\widehat{F}(\eta^k, z^k) - \widehat{F}(\xi^k, x^k)\|^2 \le 8\lambda^2\|z^k - x^k\|^2 + 8\hat{\alpha}^2(\|\epsilon_2^k\|^2 + \|\epsilon_3^k\|^2) + \Delta P_k. \tag{29}$$

From $z^k = \Pi[x^k - \alpha_k(T(x^k) + \epsilon_1^k)]$ and Lemma 2.2 with $\alpha_k \in (0, 1]$, we also have

$$\begin{aligned}\alpha_k^2 r^2(x^k) &\le r_{\alpha_k}^2(x^k) \\ &= \|x^k - \Pi[x^k - \alpha_k T(x^k)]\|^2 \\ &\le 2\|x^k - z^k\|^2 + 2\|\Pi[x^k - \alpha_k(T(x^k) + \epsilon_1^k)] - \Pi[x^k - \alpha_k T(x^k)]\|^2 \\ &\le 2\|x^k - z^k\|^2 + 2\hat{\alpha}^2\|\epsilon_1^k\|^2, \end{aligned} \tag{30}$$

where we used Lemma 2.1(ii) in the second inequality. The claim is proved using relations (23) and (29)–(30) with $x := x^*$, for a given $x^* \in X^*$, definitions (18)–(19), and the facts that $0 < 1 - 8\lambda^2 < 1$ (see Algorithm 1) 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 3.2. ∎

3.2. Bound on oracle error. This section is devoted to the control of the oracle errors in (18)–(19). Since $\{\epsilon_1^k\}$, $\{\epsilon_2^k\}$, and $\{\Delta M_k(x^*)\}$ define martingale difference sequences, their control is simpler and uses the following result.

LEMMA 3.9 (local bound for the $\mathcal{L}^q$-norm of the martingale error). *Consider definition* (1)*, and let* $\xi^N := \{\xi_j\}_{j=1}^N$ *be an i.i.d. sample from* $\mathbf{P}$*. Suppose that Assumption* 1.1 *holds, and take* $q \in [p, \mathsf{a}p]$*. Recall the definitions in* (2)–(3)*. Then there is constant* $C_q > 0$ *(depending only on q) such that, for any* $x, x_* \in X$*,*

$$\left|\left|\left|\widehat{\epsilon}(\xi^N, x)\right|\right|\right|_q \leq C_q \frac{\sigma_q(x^*) + L_q \|x - x^*\|^\delta}{\sqrt{N}}.$$

Remark 3.10 (constants of Lemma 3.9). For $q = p = 2$, $C_2 := 1$. Otherwise, C_q in Lemma 3.9 are as in Theorem 4.10 presented in section 4.3.

As mentioned in the introduction, one of the significant challenges in analyzing our sample-based line search scheme is to control the correlated error $\|\epsilon_3^k\|^2$ in (18). 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 the following local iterative result based on empirical process theory. This is the cornerstone tool to analyze the sample-based line search scheme of Algorithm 1.

THEOREM 3.11 (local bound for the $\mathcal{L}^p$-norm of the correlated error). *Under* (1)*, consider the VI*(T, X) *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* 1.1 *holds, recall definitions* (2)–(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* $z\left(\xi^N; \alpha, x\right) := \Pi\big[x - \alpha \widehat{F}\left(\xi^N, x\right)\big]$*. Then the following hold:*

(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* $x^* \in X^*$*,*

$$\left|\left|\left|\widehat{\epsilon}\left(\xi^N, z(\xi^N; \alpha_N, x)\right)\right|\right|\right|_p \leq \frac{C_p \sigma_p(x^*) + L_p^* \mathcal{D}(X)^\delta}{\sqrt{N}},$$

where $L_p^* := \mathsf{d}_2 L_2 + p L_p$*.*

Remark 3.12 (constants of Theorem 3.11). In Theorem 3.11, the constants satisfy

$$\mathsf{c}_1 := 2C_p + C_{2p} C_{\mathsf{L}\hat{\alpha},p}, \qquad \mathsf{c}_3 \lesssim p C_{\mathsf{L}\hat{\alpha},p}^\delta, \qquad \mathsf{c}_4 := C_{2p} C_{\mathsf{L}\hat{\alpha},p},$$

$$\mathsf{c}_2 \lesssim \left[\frac{3^\delta \sqrt{d}}{\sqrt{\delta}\left(\sqrt{2}^\delta - 1\right)} + \sqrt{p}\right] C_{\mathsf{L}\hat{\alpha},p}^\delta, \qquad \mathsf{d}_2 \lesssim \left[\frac{3^\delta \sqrt{d}}{\sqrt{\delta}\left(\sqrt{2}^\delta - 1\right)} + \sqrt{p}\right],$$

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 as in Remark 3.10.

For readability, the proofs of Lemma 3.9 and Theorem 3.11 are presented in section 4.3. We are now ready to obtain the following result.

PROPOSITION 3.13 (bound on oracle error). *Consider Assumptions* 1.1, 3.1, *and* 3.3. *Recall the definitions in* (2), (18), *Lemma* 3.9, *and Theorem* 3.11. *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* $x^* \in X^*$,

$$|\Delta A_k|\mathcal{F}_k|_{\frac{p}{2}} \leq \frac{\mathsf{C}_p\left[\hat{\alpha}\sigma_{\mathsf{a}p}(x^*)\right]^2 + \overline{\mathsf{C}}_p\left(\hat{\alpha}\widetilde{L}_p\right)^2 D_k^2}{N_k}.$$

In the above, for X *compact, we have* $\mathsf{a} = 1$, $\widetilde{L}_p := (C_pL_p) \vee L_p^*$, *and* $D_k :\equiv \mathcal{D}(X)$. *For a general* X, *we have* $\mathsf{a} = 2$, $\widetilde{L}_p := \overline{L}_{2p}$, *and* $D_k := \|x^k - x^*\|$.

Proof of Proposition 3.13. 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 2.1(iii)), $\epsilon_1^k = \widehat{\epsilon}(\xi^k, x^k)$, and $x^k \in \mathcal{F}_k$. From these facts, Lemma 2.1(ii), and the Lipschitz continuity of T, we obtain

$$\left|\|z^k - x^*\||\mathcal{F}_k\right|_p \leq (1 + L\hat{\alpha})\|x^k - x^*\| + \hat{\alpha}\left|\|\epsilon_1^k\||\mathcal{F}_k\right|_p. \tag{31}$$

Lemma 3.9 with $q = p$, (12), and the facts that $x^k \in \mathcal{F}_k$ and $\xi^k \perp\!\!\!\perp \mathcal{F}_k$ imply that

$$\left|\|\epsilon_1^k\||\mathcal{F}_k\right|_p \leq C_p\frac{\sigma_p(x^*) + L_p\|x^k - x^*\|}{\sqrt{N_k}}. \tag{32}$$

Lemma 3.9 with $q = p$, (12), 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

$$\left|\|\epsilon_2^k\||\mathcal{F}_k\right|_p = \left|\left|\|\epsilon_2^k\|\left|\widehat{\mathcal{F}}_k\right|_p\right|\mathcal{F}_k\right|_p \leq C_p\frac{\sigma_p(x^*) + L_p\left|\|z^k - x^*\||\mathcal{F}_k\right|_p}{\sqrt{N_k}}. \tag{33}$$

Finally, Theorem 3.11(i), (12), Assumption 3.3, $0 < \alpha_k \leq \hat{\alpha} \leq 1$, 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

$$\left|\|\epsilon_3^k\||\mathcal{F}_k\right|_p = \left|\left\|\widehat{\epsilon}\left(\xi^k, z(\xi^k; \alpha_k, x^k)\right)\right\|\right|\mathcal{F}_k\Big|_p \leq \frac{\mathsf{c}_1\sigma_{2p}(x^*) + \overline{L}_{2p}\|x^k - x^*\|}{\sqrt{N_k}}. \tag{34}$$

The required claim is proved by putting together relations (18), (31)–(34) and using the facts that $\left|a^2|\mathcal{F}_k\right|_{\frac{p}{2}} = |a|\mathcal{F}_k|_p^2$, $(a+b)^2 \leq 2a^2 + 2b^2$, $\overline{L}_{2p} > L_pC_p$, $\mathsf{c}_1 > C_p$ (as defined in Assumption 1.1, Lemma 3.9, Theorem 3.11, and Remarks 3.10 and 3.12), and $\sigma_{2p}(x^*) \geq \sigma_p(x^*)$. The proof for the case that X is *compact* is analogous but replacing (31) by the facts that $\|x^k - x^*\| \leq \mathcal{D}(X)$ and $\|z^k - x^*\| \leq \mathcal{D}(X)$ and replacing (34) by the bound of Theorem 3.11(ii). ∎

Remark 3.14 (constants of Proposition 3.13). Recall the definitions in Assumption 1.1, Algorithm 1, Lemma 3.9, Theorem 3.11, and Remarks 3.10 and 3.12. Let $\mathsf{G}_p := \sup_k \frac{C_p L_p \hat{\alpha}}{\sqrt{N_k}}$. The constants in Proposition 3.13 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 $\overline{\mathsf{C}}_p := 34-16\lambda^2$.

3.3. Asymptotic convergence, convergence rate, and oracle complexity. In this section, we establish the asymptotic convergence of Algorithm 1 and give bounds on its iteration and oracle complexities. In the following, we set $p = 2$ (see Remark 3.22 for the interest in $p > 2$).

PROPOSITION 3.15 (stochastic quasi-Fejér property). *Consider Assumptions* 1.1 *and* 3.1–3.3 *and the definitions in Proposition* 3.13 *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* 1 *satisfies, for all* $x^* \in X^*$ *and* $k \in \mathbb{N}_0$,

$$\mathbb{E}\left[\|x^{k+1}-x^*\|^2|\mathcal{F}_k\right] \le \|x^k-x^*\|^2 - \nu r^2(x^k) + \frac{\mathsf{C}_2\left[\hat{\alpha}\sigma_{2\mathsf{a}}(x^*)\right]^2 + \overline{\mathsf{C}}_2(\hat{\alpha}\widetilde{L}_2)^2 D_k^2}{N_k} + \mathsf{C}_0\delta_k^2.$$

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 (12). 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 (20), $\alpha_k \le \hat{\alpha}$, and the fact that $(\sum_{i=1}^2 a_i)^2 \le 2\sum_{i=1}^2 a_i^2$, it follows that $\Delta P_k \le 32(2\lambda^2 + 2\hat{\alpha}^2\widehat{L}(\xi^k)^2)\delta_k^2$. This fact and $\mathbb{E}[\widehat{L}(\xi^k)^2|\mathcal{F}_k] = \mathbb{E}[\mathsf{L}(\xi)^2]$ imply that $\mathbb{E}[\Delta P_k|\mathcal{F}_k] \le \mathsf{C}_0\delta_k^2$.

After we take $\mathbb{E}[\cdot|\mathcal{F}_k]$ in Lemma 3.8, the recursion follows from the facts that $\{\Delta M_k(x^*), \mathcal{F}_k\}$ is a martingale difference and $\mathbb{E}[\Delta P_k|\mathcal{F}_k] \le \mathsf{C}_0\delta_k^2$, the facts that $\mathbb{E}\left[\alpha_k^2|\mathcal{F}_k\right] \ge \frac{[(\lambda\theta)\wedge\hat{\alpha}]^2}{|\mathsf{L}(\xi)|_2^2}$ (Lemma 3.7) and $x^k \in \mathcal{F}_k$, and Proposition 3.13 with $p = 2$. ∎

THEOREM 3.16 (asymptotic convergence). *Consider Assumptions* 1.1 *and* 3.1–3.3. *Suppose that* $\sum_k \delta_k^2 < \infty$. *Then Algorithm* 1 *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 3.15 for a general X ($\mathsf{a} := 2$), and the fact that $x^k \in \mathcal{F}_k$, we apply Theorem 2.3 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 1.2) and Π (Lemma 2.1(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 2.1(iii), 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 3.15. ∎

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 IC and OC in terms of local parameters.

PROPOSITION 3.17 ($\mathcal{L}^2$-boundedness of the iterates: unbounded case). *Let Assumptions* 1.1 *and* 3.1–3.3 *hold, and suppose that* $\sum_k \delta_k^2 < \infty$. *Recall the definitions in Algorithm* 1, (2), *Theorem* 3.11, *and Propositions* 3.13 and 3.15 *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*

$$\sum_{i=k_0}^{\infty} \frac{1}{N_i} \leq \frac{\phi}{\overline{\mathsf{C}}_2 \left(\hat{\alpha}\overline{L}_4\right)^2} \qquad \text{and} \qquad \sum_{i=k_0}^{\infty} \delta_i^2 \leq \frac{1}{\mathsf{C}_0}. \tag{35}$$

Then $\sup_{k\geq k_0} \left|\|x^k - x^*\|\right|_2^2 \leq \frac{\left|\|x^{k_0} - x^*\|\right|_2^2 + \frac{\phi \mathsf{C}_2 \sigma_4(x^*)^2}{\overline{\mathsf{C}}_2 \overline{L}_4^2} + 1}{1-\phi}$.

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 (35). Note that such a k_0 always exists since $\sum_k N_k^{-1} < \infty$ by Assumption 3.3. Consider the recursion of Proposition 3.15 for the case that 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

$$|d_k|_2^2 \leq |d_{k_0}|_2^2 + \overline{\mathsf{C}}_2 \left(\hat{\alpha}\overline{L}_4\right)^2 \sum_{i=k_0}^{k-1} \frac{|d_i|_2^2}{N_i} + \mathsf{C}_2 \left[\hat{\alpha}\sigma_4(x^*)\right]^2 \sum_{i=k_0}^{k-1} \frac{1}{N_i} + \mathsf{C}_0 \sum_{i=k_0}^{k-1} \delta_i^2. \tag{36}$$

For any $a > 0$, we define the stopping time $\tau_a := \inf\{k \geq k_0 : |d_k|_2 > a\}$. From (35)–(36) and the 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 the 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 3.18 (rate of convergence). *Consider Assumptions* 1.1 *and* 3.1–3.3. *Take any positive sequence* $\{\delta_k\}$ *such that* $\Delta := \sum_k \delta_k^2 < \infty$. *Recall the definitions in Algorithm* 1, (2), *and Propositions* 3.13 and 3.15 *with* $p = 2$. *Set*

$$N_k := N \left\lceil (k+\mu)(\ln(k+\mu))^{1+b} \right\rceil \tag{37}$$

for any $N \in \mathbb{N}$, $b > 0$, *and* $\mu > 2$. *Then Theorem* 3.16 *holds and the sequence* $\{x^k\}$ *generated by Algorithm* 1 *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* $k \in \mathbb{N}_0$:

$$\min_{i\in\{0,\ldots,k\}} \mathbb{E}[r^2(x^i)] \leq \frac{\nu^{-1}}{k+1} \left\{ \|x^0 - x^*\|^2 + \frac{\mathsf{C}_2[\hat{\alpha}\sigma_{2\mathsf{a}}(x^*)]^2 + \overline{\mathsf{C}}_2 \left(\hat{\alpha}\widetilde{L}_2\right)^2 \mathsf{J}}{Nb[\ln(\mu-1)]^b} + \mathsf{C}_0\Delta \right\}.$$

Proof. Clearly, $\{N_k\}$ satisfies Assumption 3.3 and $\sum_k \delta_k^2 < \infty$. Hence, Theorem 3.16 and Proposition 3.17 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 3.15, 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 $\min_{i\in\{0,\ldots,k\}} \mathbb{E}\left[r^2(x^i)\right] \leq \frac{1}{k+1} \sum_{i=0}^{k} \mathbb{E}\left[r^2(x^i)\right]$. ∎

We end this section with an estimate on the ICs and OCs. Unlike SA methods with endogenous stepsizes, the number of oracle calls in Algorithm 1 is a *random variable*. In order to compute the first operator step (9) of iteration k, the oracle is called $\ell_k N_k$ times using a sample-based line search (which terminates in ℓ_k random iterations).[3] For the second operator step (10) of iteration k, the oracle is called N_k times. We thus present two types of OCs for which $\min_{i\in\{0,\ldots,K\}} \mathbb{E}[r^2(x^i)] \leq \epsilon$ after Algorithm 1 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 OC* $\sum_{i=0}^{K}(1+\mathbb{E}[\ell_i])N_i$. *This will depend on the logarithm of the mean Lipschitz constant* L.

COROLLARY 3.19 (ICs and OCs). *Let the assumptions of Theorem* 3.18 *hold, and set* $N := \mathcal{O}(d)$. *Given* $\epsilon > 0$, *Algorithm* 1 *achieves the tolerance*

$$\min_{i\in\{0,\ldots,K\}} \mathbb{E}[r^2(x^i)] \leq \epsilon \tag{38}$$

after $K = b^{-1}\mathcal{O}(\epsilon^{-1})$ *iterations.*

Additionally, with probability 1, (38) *is ensured with an OC* $\sum_{i=0}^{K}(1+\ell_i)N_i$ *upper bounded by*

$$b^{-2} \cdot \log_{\frac{1}{\theta}} \left(\frac{\hat{\alpha} \max_{i\in\{0,\ldots,K\}} \widehat{L}(\xi^i)}{(\lambda\theta) \wedge \hat{\alpha}} \right) \cdot \left[\ln\left(b^{-1}\epsilon^{-1}\right)\right]^{1+b} \cdot \mathcal{O}(d\epsilon^{-2}),$$

where ℓ_k *is the number of oracle calls used in the line search scheme* (7) *at iteration* k *and* $\widehat{L}(\xi^k) = \frac{1}{N_k}\sum_{j=1}^{N_k} \mathsf{L}(\xi_j^i)$.

Moreover, (38) *is ensured with a mean OC* $\sum_{i=0}^{K}(1+\mathbb{E}[\ell_i])N_i$ *upper bounded by*

$$b^{-2} \cdot \log_{\frac{1}{\theta}} \left(\frac{\hat{\alpha} L}{(\lambda\theta) \wedge \hat{\alpha}} \right) \cdot \left[\ln\left(b^{-1}\epsilon^{-1}\right)\right]^{1+b} \cdot \mathcal{O}(d\epsilon^{-2}).$$

[3] 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 (7). In all such ℓ_k steps, the same sample $\xi^k = \{\xi_j^k\}_{j=1}^{N_k}$ is used.

Proof. We recall the definitions in Assumption 1.1, Lemma 3.9, Theorem 3.11, Remarks 3.10 and 3.12, Propositions 3.13 and 3.15, and Remark 3.14 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 3.18 imply that, up to a constant $B > 0$, for every $k \in \mathbb{N}$, $\min_{i\in\{0,\dots,k\}} \mathbb{E}[r(x^i)^2] \le Bd(Nbk)^{-1}$. Hence, given $\epsilon > 0$, we obtain $\min_{i\in\{0,\dots,K\}} \mathbb{E}[r^2(x^i)] \le \epsilon$ after $K = \mathcal{O}(dN^{-1}b^{-1}\epsilon^{-1})$ iterations.

The total number of oracle calls after K iterations is upper bounded by

$$\sum_{i=0}^{K}(1+\ell_i)N_i \lesssim \left(\max_{i\in\{0,\dots,K\}} \ell_i\right)\sum_{i=1}^{K} Ni(\ln i)^{1+b} \lesssim \left(\max_{i\in\{0,\dots,K\}} \ell_i\right) NK^2(\ln K)^{1+b}$$

$$\lesssim \left(\max_{i\in\{0,\dots,K\}} \ell_i\right) N^{-1}d^2b^{-2}\epsilon^{-2}\left[\ln\left(dN^{-1}b^{-1}\epsilon^{-1}\right)\right]^{1+b}. \tag{39}$$

Moreover, Lemma 3.7 implies that $\ell_k \le \log_{\frac{1}{\theta}}\left(\frac{\hat{\alpha}\widehat{L}(\xi^k)}{(\lambda\theta)\wedge\hat{\alpha}}\right)$. This fact, (39), 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 that

$$\mathbb{E}[\ell_k] \le \mathbb{E}\left[\log_{\frac{1}{\theta}}\left(\frac{\hat{\alpha}\widehat{L}(\xi^k)}{(\lambda\theta)\wedge\hat{\alpha}}\right)\right] \le \log_{\frac{1}{\theta}}\left(\frac{\hat{\alpha}L}{(\lambda\theta)\wedge\hat{\alpha}}\right),$$

where we used that $\mathbb{E}[\widehat{L}(\xi^k)] = L$ by the definitions of $\widehat{L}(\xi^k)$ and L and Assumption 3.3. We take expectation in (39) and 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 3.20 (linear memory budget per operation). The policy in Corollary 3.19 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 1. 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 3.21. Recall the constant definitions in Assumption 1.1, Lemma 3.9, Theorem 3.11, Remarks 3.10 and 3.12, and Remark 3.14 with $p = 2$. Recall also the definition of C_0 in Proposition 3.15. By Proposition 3.17 (X unbounded), the constant J in Theorem 3.18 satisfies

$$\mathsf{J} \le \frac{\max_{k\in\{0,\dots,k_0\}}\left\|\|x^k - x^*\|\right\|_2^2 + \frac{\phi\mathsf{C}_2\sigma_4(x^*)^2}{\overline{\mathsf{C}}_2\overline{L}_4^2} + 1}{1-\phi} \lesssim \max_{k\in\{0,\dots,k_0\}}\left\|\|x^k - x^*\|\right\|_2^2 + \frac{\sigma_4(x^*)^2}{|\mathsf{L}(\xi)|_4^2}. \tag{40}$$

Moreover, if we choose the sequence $\{\delta_k\}$ in Algorithm 1 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 (35) and (37), k_0 in (40) can be estimated by

$$k_0 := \left\lceil \exp\left\{\sqrt[b]{\frac{\overline{\mathsf{C}}_2\left(\hat{\alpha}\overline{L}_4\right)^2}{\phi bN}}\right\} - \mu + 1\right\rceil \bigvee \left\lceil \exp\left\{\sqrt[b]{\frac{\mathsf{C}_0\Delta_0^2}{b}}\right\} - \mu + 1\right\rceil. \tag{41}$$

Remark 3.22 (boundedness in $\mathcal{L}^p$). Adapting the proofs of Propositions 3.13 and 3.17, it is possible to prove, in the case that X is unbounded, that *the sequence* $\{x^k\}$ *is* $\mathcal{L}^p$*-bounded* for any given $p \geq 4$ satisfying Assumption 1.1. This is a significant statistical stability property. The proof exploits that $\Delta M_k(x^*)$ in (19) is still a martingale difference even if ϵ_3^k in (12) is not.

4. An empirical process theory for DS-SA line search schemes. Oracle errors defining *martingale differences*, as found in standard SA methods with exogenous stepsizes, can be controlled in a relatively straightforward way (see Lemma 3.9). The main objective of this section is to prove Theorem 3.11. This is the cornerstone tool to handle *nonmartingale-like* oracle errors in DS-SA line search schemes in the context where L is unknown.

To prove Theorem 3.11, we will crucially require intermediate results which rely on a branch of statistics called *empirical process* (EP) *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 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.

We apply EP theory as a successful way to analyze SA line search schemes. Referring to Algorithm 1 and Theorem 3.11, 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 iteratively construct an EP that *locally decouples* the dependence between ξ^k and z^k in $\widehat{\epsilon}(\xi^k, z^k)$ produced by backtracking. 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 a radius of* $\mathcal{O}(\|x^k - x^*\| + \|\widehat{\epsilon}(\xi^k, x^k)\|)$. Based on this fact and on i.i.d. sampling, our decoupling technique follows these guidelines:

(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*, it is hence easier to estimate.

The developed theory is presented in consecutive sections. The statistical preliminaries used outside the proofs are carefully introduced so as to make the presentation as self-contained as possible. We refer the reader to the excellent book [5], a standard reference in the area. A global outline is as follows. If $Z := \sup_{t\in\mathcal{T}} Z_t$ for a stochastic process $(Z_t)_{t\in\mathcal{T}}$, a possible way to establish an upper bound on $\mathbb{E}[Z]$ is to use *chaining* arguments, assuming that the increments of $(Z_t)_{t\in\mathcal{T}}$ have a *sub-Gaussian* tail (see Definition 4.1 and Lemma 4.5). In section 4.1, we derive instead an upper bound on $|Z|_2 \geq \mathbb{E}[Z]$ in Lemma 4.5 in order to cope with *heavy-tailed* operators (Assumption 1.1). As a consequence, we will need the fact that the square of a sub-Gaussian random variable is a *sub-Gamma* random variable (see Definition 4.1). Lemma 4.5 and the *self-normalized* sub-Gaussian tail inequality of Theorem 4.7 provide an alternative in order to upper bound the *mean* of heavy-tailed Hölder continuous EPs. This technique will be used in the proof of the general lemma (Lemma 4.9) in section 4.2. This lemma provides a *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. Besides Lemma 4.5 and Theorem 4.7, the proof of Lemma 4.9 requires a simple decoupling argument via Hölder's inequality and the control of *variance* terms presented in Theorem 4.6. Finally, the proof of Theorem 3.11 is given in section 4.3.

It relies on Lemma 4.9, the BDG inequality in Hilbert spaces ([23] and Theorem 4.10), and the ideas of items (i)–(iii) above.

4.1. The $\mathcal{L}^2$-norm of suprema of sub-Gaussian processes. In order to bound $\mathbb{E}[Z]$ or $|Z|_2$ of $Z := \sup_{t\in\mathcal{T}} Z_t$ for a stochastic process $(Z_t)_{t\in\mathcal{T}}$, it is important to understand the tail behavior of its increments $(Z_t - Z_{t'})_{(t,t')\in\mathcal{T}\times\mathcal{T}}$. A celebrated sufficient condition is to guarantee that the increments of $(Z_t)_{t\in\mathcal{T}}$ have sub-Gaussian or sub-Gamma tails (see Lemma 4.5). Such random random variables are defined in the following.

DEFINITION 4.1 (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 $s \in \mathbb{R}$, $\ln \mathbb{E}\left[e^{sY}\right] \le \frac{\sigma^2 s^2}{2}$. 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 $0 < s < \frac{1}{c}$, $\ln \mathbb{E}\left[e^{sY}\right] \le \frac{\sigma^2 s^2}{2(1-cs)}$.*

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 [10] in the one-dimensional setting.

THEOREM 4.2 (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 $0 \le s < \frac{1}{2\sigma^2}$, $\ln \mathbb{E}\left[e^{sY^2}\right] \le \sigma^2 s + \frac{\sigma^4 s^2}{1-2\sigma^2 s}$.*

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* [5]. 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 4.3 (metric entropy). *Let $(\mathcal{T}, \mathrm{d})$ 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 $s \ne t$, 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 $\mathrm{d}(s,t) \le \theta$. 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 [5]).

LEMMA 4.4 (metric entropy of Euclidean balls). *Let $\mathbb{B}$ be the Euclidean unit ball of $\mathbb{R}^d$. For all $\theta \in (0, 1]$, $H(\theta, \mathbb{B}) \le d \ln\left(1 + \frac{2}{\theta}\right)$.*

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 discretization precision θ diminishes. This is a key property in order for the chaining method to work. To facilitate the presentation, the proof of Lemma 4.5 is left to the appendix.

LEMMA 4.5 ($\mathcal{L}^2$-norm of suprema of sub-Gaussian processes). *Let $(\mathcal{T}, d)$ be a totally bounded metric space, and let $\theta := \sup_{t\in\mathcal{T}} \mathrm{d}(t, t_0)$ 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 $\lambda > 0$,*

$$\ln \mathbb{E}[\exp\{\lambda(Z_t - Z_{t'})\}] \le a\, \mathrm{d}(t,t')^\delta \lambda + \frac{v\, \mathrm{d}(t,t')^{2\delta}\lambda^2}{2}. \tag{42}$$

Then

$$\left|\sup_{t\in\mathcal{T}} Z_t - Z_{t_0}\right|_2 \le (3\theta)^\delta \sqrt{2(a^2+v)}\left[\frac{1}{2^\delta-1} + \sum_{i=1}^{\infty}\frac{\sqrt[4]{8H\left(\theta 2^{-i},\mathcal{T}\right)} + 2\sqrt{H\left(\theta 2^{-i},\mathcal{T}\right)}}{2^{i\delta}}\right].$$

4.2. Heavy-tailed Hölder continuous operators: Self-normalization and $\mathcal{L}^q$-norms of suprema of EPs. 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 1.1. The main result proved in this section is Lemma 4.9. The proof is based on the following theorem (see Theorem 15.14 in [5]).

THEOREM 4.6 ($\mathcal{L}^q$-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 $q \ge 2$,

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

In order to cope with a heavy-tailed $\mathsf{L}(\xi)$ in Assumption 1.1, we will need Theorem 4.7, a result due to Panchenko (see Theorem 1 in [26] or Theorem 12.3 in [5]). 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. Assumption 1.1 turns out to be sufficient to estimate this quantity.

THEOREM 4.7 (Panchenko's inequality for self-normalized EPs). *Consider a countable family $\mathcal{G}$ of measurable functions $f : \Xi \to \mathbb{R}$ such that $\mathbb{E}[f(\xi)^2] < \infty$. 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*

$$Y := \sup_{f\in\mathcal{G}} \sum_{j=1}^N f(\xi_j) \quad \text{and} \quad V := \mathbb{E}\left\{\sup_{f\in\mathcal{G}} \sum_{j=1}^N [f(\xi_j) - f(\eta_j)]^2 \,\middle|\, \xi_1,\dots,\xi_N\right\}.$$

Then there exists a universal constant $\mathsf{c} > 0$ such that, for all $t > 0$,

$$\mathbb{P}\left\{Y - \mathbb{E}[Y] \ge \mathsf{c}\sqrt{V(1+t)}\right\} \bigvee \mathbb{P}\left\{Y - \mathbb{E}[Y] \le -\mathsf{c}\sqrt{V(1+t)}\right\} \le e^{-t}.$$

Finally, before proving Lemma 4.9, we will need Theorem 4.8, which is a standard tail characterization of sub-Gaussian random variables. Theorem 2.1 in [5] gives a proof for the case $\mathbb{E}[\tilde{Y}] = 0$. The adaptation for the general case is immediate using the integral formula $\mathbb{E}[\tilde{Y}] \le \mathbb{E}[|\tilde{Y}|] = \int_0^\infty \mathbb{P}(|\tilde{Y}| > t)\,\mathrm{d}\,t$ and $\int_0^\infty e^{-\frac{t^2}{2}}\,\mathrm{d}\,t = \sqrt{\frac{\pi}{2}}$.

THEOREM 4.8 (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 $t > 0$,*

$$\mathbb{P}\left\{\tilde{Y} \ge \sqrt{2vt}\right\} \bigvee \mathbb{P}\left\{\tilde{Y} \le -\sqrt{2vt}\right\} \le e^{-t},$$

then, for all $t > 0$, we have $\ln \mathbb{E}\left[e^{t\tilde{Y}}\right] \le e^{\sqrt{\frac{v\pi}{2}}t + 8vt^2}$.

We finish with the proof of Lemma 4.9 using Lemma 4.5 and Theorems 4.6–4.8.

LEMMA 4.9 (local uniform bound for the $\mathcal{L}^p$-norm of empirical error increments). *Consider definition* (1)*, and let* $\xi^N := \{\xi_j\}_{j=1}^N$ *be an i.i.d. sample from* $\mathbf{P}$*. Suppose that Assumption* 1.1 *holds, and recall definitions* (2)–(3)*. Given* $x_* \in X$ *and* $R > 0$*, we define*

$$Z := \sup_{x \in \mathbb{B}[x_*, R] \cap X} \left\| \widehat{\epsilon}(\xi^N, x) - \widehat{\epsilon}(\xi^N, x_*) \right\|. \tag{43}$$

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 4.6. 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

$$\begin{aligned} Z &= \sup_{u \in \mathbb{B}_X} \frac{1}{N} \left\| \sum_{j=1}^N \epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*) \right\| \\ &= \sup_{u \in \mathbb{B}_X} \frac{1}{N} \sup_{y \in \mathbb{B}} \left\langle \sum_{j=1}^N \epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*), y \right\rangle \\ &= \sup_{(u,y) \in \mathbb{B}_X \times \mathbb{B}} \frac{1}{N} \sum_{j=1}^N \langle \epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*), y \rangle, \end{aligned} \tag{44}$$

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

$$X_{j,t} := \frac{1}{N} \langle \epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*), y \rangle, \tag{45}$$

$$\tilde{Z}_t := \sum_{j=1}^N X_{j,t}. \tag{46}$$

From Assumption 1.1, 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

$$\mathrm{d}(t, t') := \|u - u'\| + \|y - y'\|. \tag{47}$$

In particular, $(\tilde{Z}_t)_{t \in \mathcal{T}}$ is also a continuous random process. This fact and the separability of $\mathcal{T}$ imply that $Z = \sup_{t \in \mathcal{T}_0} \tilde{Z}_t = \sup_{t \in \mathcal{T}_0} |\tilde{Z}_t|$ is measurable, where $\mathcal{T}_0$ is a dense countable subset of $\mathcal{T}$. Our next objective is to use Theorem 4.6, 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 4.5 and Theorems 4.7–4.8. At this point, let's fix $t = (u, y) \in \mathcal{T}_0$ and $t' = (u', y') \in \mathcal{T}_0$ 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 $\mathbb{E}[f(\xi)^2] < \infty$ since $\|\|F(\xi,\cdot)\|\|_2 < \infty$ on X (Assumption 1.1). By construction and (45)–(46), 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 $\mathbb{E}\big[\sum_{j=1}^N f(\xi_j)\big] = 0$, using (1), (2), and the fact that $\{\xi_j\}_{j\in[N]}$ is an i.i.d. sample of $\mathbf{P}$.

The previous observations allow us to claim Theorem 4.7 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 4.7 and $\mathbb{E}\big[\sum_{j=1}^N f(\xi_j)\big] = 0$ imply that, for all $\lambda > 0$,

$$\mathbb{P}\left\{\sum_{j=1}^N f(\xi_j) \geq \mathsf{c}\sqrt{V(1+\lambda)}\right\} \bigvee \mathbb{P}\left\{\sum_{j=1}^N f(\xi_j) \leq -\mathsf{c}\sqrt{V(1+\lambda)}\right\} \leq e^{-\lambda} \tag{48}$$

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) and the Hölder continuity of $F(\xi,\cdot)$ and T (Assumption 1.1 and Lemma 1.2) imply that $\epsilon(\xi,\cdot)$ is $(\mathsf{L}(\xi) + L, \delta)$-Hölder continuous on X. This, the 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 $\Delta f_j := N\,|[f(\xi_j) - f(\eta_j)]|$,

$$\begin{aligned}
\Delta f_j &\leq |\langle \epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*) - \epsilon(\eta_j, x_* + Ru) + \epsilon(\eta_j, x_*), y - y'\rangle| \\
&\quad + |\langle \epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_* + Ru') - \epsilon(\eta_j, x_* + Ru) + \epsilon(\eta_j, x_* + Ru'), y'\rangle| \\
&\leq [\mathsf{L}(\xi_j) + \mathsf{L}(\eta_j) + 2L]\, R^\delta \left[\|y - y'\| + \|u - u'\|^\delta\right] \\
&\leq [\mathsf{L}(\xi_j) + \mathsf{L}(\eta_j) + 2L]\, R^\delta 2^{1-\delta} \left[\|y - y'\|^{\frac{1}{\delta}} + \|u - u'\|\right]^\delta \\
&\leq [\mathsf{L}(\xi_j) + \mathsf{L}(\eta_j) + 2L]\, R^\delta 2^{(1-\delta)} \left[\|y - y'\| + \|u - u'\|\right]^\delta,
\end{aligned}$$

where we used the concavity of $\mathbb{R}_+ \ni x \mapsto x^\delta$ in the 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 the last inequality. We take squares in the above inequality and use relation $(\sum_{i=1}^3 a_i)^2 \leq 3\sum_{i=1}^3 a_i^2$ and the definitions of V and (47). We thus obtain

$$\begin{aligned}
V &\leq \frac{3 \cdot 4^{1-\delta} R^{2\delta}\, \mathrm{d}(t,t')^{2\delta}}{N} \left\{\sum_{j=1}^N \frac{\mathsf{L}(\xi_j)^2}{N} + \sum_{j=1}^N \frac{\mathbb{E}\left[\mathsf{L}(\eta_j)^2 | \xi_1, \ldots, \xi_N\right]}{N} + 4L^2\right\} \\
&= \frac{3 \cdot 4^{1-\delta} R^{2\delta}\, \mathrm{d}(t,t')^{2\delta} W_N^2}{N},
\end{aligned} \tag{49}$$

where we have defined

$$W_N := \sqrt{\frac{1}{N}\sum_{j=1}^N \mathsf{L}(\xi_j)^2 + |\mathsf{L}(\xi)|_2^2 + 4L^2} \tag{50}$$

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 (48)–(49) 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 4.8 imply that, for some universal constants $C_1, C_2 > 0$ and for all $\lambda > 0$,

$$\ln \mathbb{E}\left[\exp\left\{\frac{(\tilde{Z}_t - \tilde{Z}_{t'})}{W_N}\lambda\right\}\right] \le \frac{C_1 2^{1-\delta}R^\delta \,\mathrm{d}(t,t')^\delta}{\sqrt{N}}\lambda + \frac{C_2^2 4^{1-\delta}R^{2\delta}\,\mathrm{d}(t,t')^{2\delta}}{2N}\lambda^2. \tag{51}$$

We now observe that (51) holds for any $t, t' \in \mathcal{T}_0$. Inequality (51) and Lemma 4.5 with $(\mathcal{T}_0, \mathrm{d})$ as defined in (47), the continuous process $\mathcal{T}_0 \ni t \mapsto Z_t := \frac{\tilde{Z}_t}{W_N}$, $t_0 := (0,0)$, $\theta := \sup_{t\in\mathcal{T}_0} \mathrm{d}(t,0) \le 2$, $a := \frac{C_1 2^{1-\delta}R^\delta}{\sqrt{N}}$, and $v := \frac{C_2^2 4^{1-\delta}R^{2\delta}}{N}$ imply that

$$\left|\sup_{t\in\mathcal{T}_0} Z_t\right|_2 \le \frac{\sqrt{2}C2^{1-\delta}(6R)^\delta}{\sqrt{N}}\left[\frac{1}{2^\delta - 1} + \sum_{i=1}^\infty \frac{\sqrt[4]{8H\left(2^{-i+1}, \mathcal{T}_0\right)} + 2\sqrt{H\left(2^{-i+1}, \mathcal{T}_0\right)}}{2^{i\delta}}\right], \tag{52}$$

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 4.4 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

$$\begin{aligned}\sum_{i=1}^\infty \frac{\sqrt[4]{8H\left(2^{-i+1}, \mathcal{T}_0\right)} + 2\sqrt{H\left(2^{-i+1}, \mathcal{T}_0\right)}}{2^{i\delta}} &\lesssim \sqrt{d}\sum_{i=1}^\infty \frac{\sqrt{\ln(1+2^{i+1})}}{2^{i\delta}} \\ &\lesssim \sqrt{d}\sum_{i=1}^\infty \frac{\sqrt{i+1}}{2^{i\delta}} \lesssim \frac{\sqrt{d/\delta}}{2^{\frac{\delta}{2}} - 1},\end{aligned} \tag{53}$$

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[4] $\sum_{i=1}^\infty 2^{-\frac{i\delta}{2}} = \frac{1}{2^{\frac{\delta}{2}}-1}$.

Hölder's inequality implies that

$$\mathbb{E}[Z] = \mathbb{E}\left[\sup_{t\in\mathcal{T}_0} |\tilde{Z}_t|\right] = \mathbb{E}\left[\sup_{t\in\mathcal{T}_0} |Z_t| \cdot W_N\right] \le \left|\sup_{t\in\mathcal{T}_0} |Z_t|\right|_2 \cdot |W_N|_2. \tag{54}$$

Since $\{\xi_j\}_{j\in[N]}$ is an i.i.d. sample from $\mathbf{P}$, we also obtain from (50) that $|W_N|_2 \le 2|\mathsf{L}(\xi)|_2 + 2L = 2L_2$. Finally, this, relations (52)–(54), and the facts that $2^{1-\delta}6^\delta = 2\cdot 3^\delta$ and $2^\delta - 1 \ge 2^{\frac{\delta}{2}} - 1$ imply that

$$\mathbb{E}[Z] \lesssim \frac{\sqrt{d}(3R)^\delta L_2}{\left(2^{\frac{\delta}{2}} - 1\right)\sqrt{\delta N}}. \tag{55}$$

Part 2 (an upper bound on M and $\widehat{\sigma}^2$). From the definition of $\widehat{\sigma}^2$ in Theorem 4.6

[4] The previous fact can be derived from the inequality $2^x \ge 1 + (\ln 2)x$.

and (45), we get

$$\begin{aligned}\widehat{\sigma} &= \sqrt{\sup_{(u,y)\in\mathcal{T}_0} \frac{1}{N^2}\sum_{j=1}^{N} \mathbb{E}\left[\langle \epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*), y\rangle^2\right]} \\ &\le \sqrt{\frac{1}{N}\sup_{(u,y)\in\mathcal{T}_0} \mathbb{E}\left[\sum_{j=1}^{N} \frac{(\mathsf{L}(\xi_j)+L)^2}{N} R^{2\delta}\|u\|^{2\delta}\|y\|^2\right]} \\ &\le \frac{R^\delta(|\mathsf{L}(\xi)|_2 + L)}{\sqrt{N}}, \end{aligned} \tag{56}$$

where we used the fact that $\|\epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*)\| \le [\mathsf{L}(\xi_j) + L]R^\delta$ for $u \in \mathbb{B}_X$ (Assumption 1.1 and Lemma 1.2) in the first inequality and the fact that $\{\xi_j\}_{j\in[N]}$ is an i.i.d. sample of $\mathbf{P}$ in the second inequality.

From the definition of M in Theorem 4.6 and (45), we get

$$\begin{aligned}|M|_p^p &= \mathbb{E}\left[\left(\max_{j\in[N]}\sup_{t\in\mathcal{T}_0}|X_{j,t}|\right)^p\right] = \mathbb{E}\left[\max_{j\in[N]}\sup_{t\in\mathcal{T}_0}|X_{j,t}|^p\right] \\ &\le \frac{1}{N^p}\sum_{j=1}^{N}\mathbb{E}\left[\sup_{t\in\mathcal{T}_0}|\langle \epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*), y\rangle|^p\right] \\ &\le \frac{1}{N^{p-1}}\sup_{(u,y)\in\mathcal{T}_0}\mathbb{E}\left[\sum_{j=1}^{N}\frac{(\mathsf{L}(\xi_j)+L)^p}{N}R^{p\delta}\|u\|^{p\delta}\|y\|^p\right] \\ &\le \frac{R^{p\delta}|\mathsf{L}(\xi)+L|_p^p}{N^{p-1}}, \end{aligned}$$

where, again, we used the fact that $\|\epsilon(\xi_j, x_* + Ru) - \epsilon(\xi_j, x_*)\| \le [\mathsf{L}(\xi_j) + L]R^\delta$ for $u \in \mathbb{B}_X$ in the second inequality and the fact that $\{\xi_j\}_{j\in[N]}$ is an i.i.d. sample of $\mathbf{P}$ in the third inequality. We take the pth root in the above inequality and note that for $p \ge 2$ we have $N^{\frac{p-1}{p}} \ge \sqrt{N}$, obtaining

$$|M|_p \le \frac{(|\mathsf{L}(\xi)|_p + L)R^\delta}{\sqrt{N}}. \tag{57}$$

From Theorem 4.6, (55)–(57), and the definitions of L_2 and L_p in Assumption 1.1, we obtain the required claim. ∎

4.3. The proof of Theorem 3.11. With the theory developed in sections 4.1–4.2, we are now ready to prove Theorem 3.11. We shall use Lemma 4.9 and follow the ideas of items (i)–(iii) presented in the introduction of section 4. We will also need Lemma 3.9 to control oracle errors which define martingale differences. The proof of Lemma 3.9 relies on the following theorem [23].

THEOREM 4.10 (Burkholder–Davis–Gundy inequality in $\mathbb{R}^d$). *Let $\|\cdot\|$ be the Euclidean norm in $\mathbb{R}^d$. Then, for all $q \ge 2$, 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*

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

Proof of Lemma 3.9. 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 $\mathcal{G}_t := \sigma(y_0, \dots, y_t)$ 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 an $\mathbb{R}^d$-valued martingale whose increments satisfy

$$|||y_t - y_{t-1}|||_q = \left| \frac{\|\epsilon(\xi,x)\|}{N} \right|_q \le \frac{|||\epsilon(\xi,x_*)|||_q + L_q\|x-y\|^\delta}{N},$$

using that $|||\epsilon(\xi,\cdot)|||_q$ is Hölder continuous with modulus $L_q = |\mathsf{L}(\xi)|_q + L$ and exponent δ (Lemma 1.2) in the inequality. The required claim follows from the above relation, Theorem 4.10, $\widehat{\epsilon}(\xi^N, x) = y_N$, and $|||y_N|||_q \le \left|\sup_{j\le N} \|y_j\|\right|_q$. 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 1.2 imply the claim of the lemma with $C_2 = 1$. ∎

Proof of Theorem 3.11. We fix $x \in X$ and $x^* \in X^*$ as stated in the theorem and set $z^N := z(\xi^N; \alpha_N, x)$. 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 [29] and Assumption 1.1 imply that the map $\Xi \times X \ni (\omega, x) \mapsto \|\widehat{\epsilon}(\xi^N(\omega), x)\|$ is a *normal integrand*, that is,

$$\omega \mapsto \operatorname{epi} \left\|\widehat{\epsilon}(\xi^N(\omega), \cdot)\right\| := \{(x, y) \in X \times \mathbb{R} : \left\|\widehat{\epsilon}(\xi^N(\omega), x)\right\| \le y\}$$

is a set-valued measurable function. This fact and Theorem 14.37 in [29] imply further that, for any measurable function $\epsilon : \Omega \to [0, \infty)$ and $R > 0$,

$$\omega \mapsto \sup_{x' \in \mathbb{B}(\epsilon(\omega)) \cap X} \left\|\widehat{\epsilon}(\xi^N(\omega), x')\right\| \quad \text{and} \quad \omega \mapsto \sup_{x' \in \mathbb{B}[x^*, R] \cap X} \left\|\widehat{\epsilon}(\xi^N(\omega), x')\right\| \tag{58}$$

are measurable functions.

We first prove item (ii) for the easier case when X is compact. We set $R := \mathcal{D}(X)$ and note that $z^N \in \mathbb{B}[x^*, R] \cap X$. This and (58) imply that

$$\begin{aligned}
\left|\|\widehat{\epsilon}(\xi^N, z^N)\|\right|_p &\le \left| \sup_{x' \in \mathbb{B}[x^*,R] \cap X} \|\widehat{\epsilon}(\xi^N, x')\| \right|_p \\
&\le \left| \sup_{x' \in \mathbb{B}[x^*,R] \cap X} \|\widehat{\epsilon}(\xi^N, x') - \widehat{\epsilon}(\xi^N, x^*)\| \right|_p + \left|\|\widehat{\epsilon}(\xi^N, x^*)\|\right|_p \\
&\le c \left[\frac{3^\delta \sqrt{d} L_2}{\sqrt{\delta}\left(\sqrt{2}^\delta - 1\right)} + \sqrt{p} L_2 + p L_p \right] \frac{\mathcal{D}(X)^\delta}{\sqrt{N}} + \frac{C_p |||\epsilon(\xi, x^*)|||_p}{\sqrt{N}}
\end{aligned}$$

for some universal constant $c > 0$, where we used Lemmas 4.9 and 3.9 with $q = p$ in the last inequality. The above inequality and definition (2) prove item (ii).

We now prove item (i) in the case that X may be unbounded. Given $\alpha \in [0, \hat{\alpha}]$, Lemma 2.1(iii) implies that $x^* = \Pi[x^* - \alpha T(x^*)]$. Taking into account this fact, Lemma 2.1(ii), and the definitions of $z(\xi^N; \alpha, x)$, (2), and (3), we get that, for any $\alpha \in [0, \hat{\alpha}]$,

$$\begin{aligned}
\left\|x^* - z(\xi^N; \alpha, x)\right\| &= \left\|\Pi\left[x^* - \alpha T(x^*)\right] - \Pi\left[x - \alpha\left(T(x) + \widehat{\epsilon}(\xi^N, x)\right)\right]\right\| \\
&\le \|x^* - x\| + \alpha\|T(x) - T(x^*)\| + \alpha\left\|\widehat{\epsilon}(\xi^N, x)\right\| \\
&\le (1 + L\hat{\alpha})\left[\|x - x^*\| \vee \|x - x^*\|^\delta\right] + \hat{\alpha}\left\|\widehat{\epsilon}(\xi^N, x)\right\|, \qquad (59)
\end{aligned}$$

where, in the last inequality, we used the Hölder continuity of T (Lemma 1.2).

In what follows, we define the quantities

$$s_* := L_{2p}\Delta(x, x^*) \quad \text{and} \quad \epsilon_N := \left\|\widehat{\epsilon}(\xi^N, x)\right\|. \tag{60}$$

Setting $\alpha := \alpha_N$ in (59), we have that $z^N \in \mathbb{B}(\epsilon_N) \cap X$. We now make the following decomposition:

$$\left|\left\|\widehat{\epsilon}(\xi^N, z^N)\right\|\right|_p = I_1 + I_2, \tag{61}$$

using the definitions

$$I_1 := \left|\left\|\widehat{\epsilon}(\xi^N, z^N)\right\| 1_{\{\epsilon_N \le s_*\}}\right|_p \quad \text{and} \quad I_2 := \left|\left\|\widehat{\epsilon}(\xi^N, z^N)\right\| 1_{\{\epsilon_N > s_*\}}\right|_p.$$

Part 1 (upper bound on I_1). From the fact that $z^N \in \mathbb{B}(\epsilon_N) \cap X$ and (58), we may bound I_1 by

$$\begin{aligned}
I_1 &= \left|\|\widehat{\epsilon}(\xi^N, z^N)\| 1_{\{\epsilon_N \le s_*\}}\right|_p \\
&\le \left| \sup_{x' \in \mathbb{B}(s_*) \cap X} \|\widehat{\epsilon}(\xi^N, x')\| \right|_p \\
&\le \left| \sup_{x' \in \mathbb{B}(s_*) \cap X} \|\widehat{\epsilon}(\xi^N, x') - \widehat{\epsilon}(\xi^N, x^*)\| \right|_p + \left|\|\widehat{\epsilon}(\xi^N, x^*)\|\right|_p \\
&\le c \left[\frac{3^\delta \sqrt{d} L_2}{\sqrt{\delta}\left(\sqrt{2}^\delta - 1\right)} + \sqrt{p} L_2 + p L_p \right] \frac{\mathsf{R}(s_*)^\delta}{\sqrt{N}} + \frac{C_p |||\epsilon(\xi, x^*)|||_p}{\sqrt{N}},
\end{aligned}$$

where we used Lemmas 4.9 and 3.9 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

$$I_1 \le \left[\left(c_\delta \sqrt{d} + c\sqrt{p} \right) L_2 + c p L_p \right] C^\delta_{\mathsf{L}\hat{\alpha}, p} \frac{\Delta(x, x^*)^\delta}{\sqrt{N}} + \frac{C_p |||\epsilon(\xi, x^*)|||_p}{\sqrt{N}}, \tag{62}$$

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

$$\begin{aligned}
\left\|\widehat{\epsilon}(\xi^N, z^N)\right\| &\le \left\|\widehat{\epsilon}(\xi^N, z^N) - \widehat{\epsilon}(\xi^N, x^*)\right\| + \left\|\widehat{\epsilon}(\xi^N, x^*)\right\| \\
&\le \left\| \frac{1}{N} \sum_{j=1}^N \left[F(\xi_j, z^N) - F(\xi_j, x^*) \right] \right\| + \left\| T(z^N) - T(x^*) \right\| + \left\|\widehat{\epsilon}(\xi^N, x^*)\right\| \\
&\le \left(\widehat{L}_N + L \right) \left\| z^N - x^* \right\|^\delta + \left\|\widehat{\epsilon}(\xi^N, x^*)\right\| \\
&\le \left(\widehat{L}_N + L \right) (1 + L\hat{\alpha}) \Delta(x, x^*) + \hat{\alpha} \left(\widehat{L}_N + L \right) \epsilon_N + \epsilon_N^*,
\end{aligned}$$

using Assumption 1.1 and Lemma 1.2 in the third inequality and (59) with $\alpha := \alpha_N$, (60), and the definition $\epsilon_N^* := \left\|\widehat{\epsilon}(\xi^N, x^*)\right\|$ in the fourth inequality. The inequality

above and the definition of I_2 imply that

$$\begin{aligned} I_2 &= \left|\left\|\widehat{\epsilon}(\xi^N, z^N)\right\| 1_{\{\epsilon_N > s_*\}}\right|_p \\ &\le (1+L\hat{\alpha})\Delta(x,x^*)\left|\left(\widehat{L}_N + L\right) 1_{\{\epsilon_N > s_*\}}\right|_p + \hat{\alpha}\left|\left(\widehat{L}_N + L\right)\epsilon_N\right|_p + |\epsilon_N^*|_p \\ &\le (1+L\hat{\alpha})\Delta(x,x^*)\left|\widehat{L}_N + L\right|_{2p}\left|1_{\{\epsilon_N > s_*\}}\right|_{2p} + \hat{\alpha}\left|\widehat{L}_N + L\right|_{2p}|\epsilon_N|_{2p} + |\epsilon_N^*|_p, \end{aligned} \tag{63}$$

where we used Hölder's inequality.

With respect to the last term in the rightmost expression of (63), we have, in view of Lemma 3.9 with $q = p$,

$$|\epsilon_N^*|_p = \left|\|\widehat{\epsilon}(\xi^N, x^*)\|\right|_p \le \frac{C_p |||\epsilon(\xi, x^*)|||_p}{\sqrt{N}}. \tag{64}$$

Concerning the second term in the rightmost expression of (63), Lemma 3.9 with $q = 2p$ implies that

$$|\epsilon_N|_{2p} = \left|\left\|\widehat{\epsilon}(\xi^N, x)\right\|\right|_{2p} \le C_{2p}\frac{|||\epsilon(\xi, x^*)|||_{2p} + L_{2p}\|x - x^*\|^\delta}{\sqrt{N}}. \tag{65}$$

From Markov's inequality and (65) we obtain

$$\begin{aligned} \left|1_{\{\epsilon_N > s_*\}}\right|_{2p} &= \sqrt[2p]{\mathbb{E}\left[1_{\{\epsilon_N > s_*\}}\right]} = \sqrt[2p]{\mathbb{P}\left(\|\widehat{\epsilon}(\xi^N, x)\| > s_*\right)} \\ &\le \sqrt[2p]{\frac{\mathbb{E}\left[\|\widehat{\epsilon}(\xi^N, x)\|^{2p}\right]}{s_*^{2p}}} = \frac{\left|\left\|\widehat{\epsilon}(\xi^N, x)\right\|\right|_{2p}}{s_*} \\ &\le C_{2p}\frac{|||\epsilon(\xi, x^*)|||_{2p} + L_{2p}\|x - x^*\|^\delta}{s_*\sqrt{N}}. \end{aligned} \tag{66}$$

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

$$\left|\widehat{L}_N + L\right|_{2p} \le |\mathsf{L}(\xi)|_{2p} + L = L_{2p}.$$

Using this fact and putting together relations (63)–(66), we get

$$\begin{aligned} I_2 &\le (1+L\hat{\alpha})\frac{\Delta(x,x^*)L_{2p}}{s_*}C_{2p}\frac{|||\epsilon(\xi, x^*)|||_{2p} + L_{2p}\|x - x^*\|^\delta}{\sqrt{N}} \\ &\quad + L_{2p}\hat{\alpha}C_{2p}\frac{|||\epsilon(\xi, x^*)|||_{2p} + L_{2p}\|x - x^*\|^\delta}{\sqrt{N}} + \frac{C_p|||\epsilon(\xi, x^*)|||_p}{\sqrt{N}} \\ &= C_{2p}\left(1 + L\hat{\alpha} + L_{2p}\hat{\alpha}\right)\frac{|||\epsilon(\xi, x^*)|||_{2p}}{\sqrt{N}} + \frac{C_p|||\epsilon(\xi, x^*)|||_p}{\sqrt{N}} \\ &\quad + C_{2p}\left(1 + L\hat{\alpha} + L_{2p}\hat{\alpha}\right)\frac{L_{2p}\|x - x^*\|^\delta}{\sqrt{N}}, \end{aligned} \tag{67}$$

where we used the fact that[5] $s_* = L_{2p}\Delta(x, x^*)$.

Relations (61)–(62) and (67), definition (2), and the facts that $\Delta(x,x^*)^\delta \le \delta_1 \vee \|x - x^*\|^\delta$ and $|||\epsilon(\xi, x^*)|||_p \le |||\epsilon(\xi, x^*)|||_{2p}$ prove item (i). ∎

[5]Note that $[\Delta(x,x^*)\|x - x^*\|^\delta]^2 \lesssim \|x - x^*\|^{4\delta}$ with $4\delta \le 2$. 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 an 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 $\beta_1, \beta_2 \in (0, 1]$.

5. Discussion on the complexity constants for Algorithm 1. For an *exact* oracle, the Lipschitz continuity is only related to the *smoothness class of the operator*. Consequently, the rate estimates using a CSP with a known Lipschitz constant (LC) or a line search scheme are the same, up to universal constants and a factor of $\mathcal{O}(\ln L)$ in the OC. For an SO, the Lipschitz continuity also quantifies the *spread of the oracle's error variance*, either by repeated use of Lemma 1.2 or in chaining and self-normalization arguments (Lemmas 4.5 and 4.9). Consequently, the lack of knowledge of the LC is expected to be 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 3.18 and the constants defined in Assumption 1.1, Lemma 3.9, Theorem 3.11, Remarks 3.10, 3.12, and 3.14, and Propositions 3.13 and 3.15 with $p = 2$.

Suppose first the LC is known. This was already considered in [11] under a more general condition than Assumption 1.1. However, it leads to weaker complexity constants as argued in the following. It is possible to show that if the stronger but fairly general condition of Lemma 1.2 holds and $\hat{\alpha} = \mathcal{O}(\frac{1}{L_2})$, then the rate statement of Theorem 3.18 and the estimates (40)–(41) are valid with $\Delta_0 := 0$ when we replace $\sigma_4(x^*)$ by $\sigma_2(x^*)$, $\overline{L}_4$ by L_2, and 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 [11]: k_0 in (41) 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

$$\min_{i\in\{1,\dots,k\}} \mathbb{E}\left[r^2(x^i)\right] \lesssim \frac{L_2^2\|x^0-x^*\|^2+\sigma_2(x^*)^2}{k}, \tag{68}$$

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 [11], which are of order of $\sigma(x^*)^4\cdot\max_{i\in\{0,\dots,k_0(x^*)\}}\mathbb{E}[\|x^i-x^*\|^2]$ with $k_0(x^*)\in\mathbb{N}_0$ depending on $\sigma(x^*)$ (see Assumption 3.8 and section 3.4.1 in [11]).

Consider now the more challenging regime when the LC is unknown. As expected, the constants in the rate of Theorem 3.18 are less sharp than the ones in (68). First, (68) is not explicitly dependent on the dimension d. In terms of dimension, the rate in Theorem 3.18 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 a 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 3.18 satisfy $\mathsf{C}_2 \lesssim \frac{\mathsf{M}}{N}$, $\overline{\mathsf{C}}_2 \lesssim \mathsf{M}$, $\mathsf{C}_0 \lesssim \mathsf{M}$, and $\frac{(\hat{\alpha}\widetilde{L}_2)^2\mathsf{J}}{N} \lesssim \mathsf{M}^2\max_{0\le k\le k_0}\mathbb{E}[\|x^k-x^*\|^2] + [\hat{\alpha}\sigma_4(x^*)]^2[\hat{\alpha}|\mathsf{L}(\xi)|_4]^2$ for a general[6] X and $\mathsf{C}_2 \lesssim 1$, $\overline{\mathsf{C}}_2 \lesssim 1$, $\mathsf{C}_0 \lesssim \mathsf{M}$, and $\frac{(\hat{\alpha}\widetilde{L}_2)^2\mathsf{J}}{N} \lesssim \mathsf{M}\,\mathcal{D}(X)^2$ for a compact X. Observe that a line search scheme can only estimate a *lower bound* for $|\mathsf{L}(\xi)|_4$. Hence, larger values of $\hat{\alpha}$ lead to larger constants when compared to (68). 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 [25].

Appendix. We now present a proof of Lemma 4.5. We shall need one more preliminary result. It bounds the expectation of the maximum of a *finite* number of

[6] For unbounded X, sharper dependence on M may be obtained via more sophisticated concentration inequalities than used here.

sub-Gamma random variables (see, e.g., Corollary 2.6 of [5]). It is an essential lemma while using chaining arguments.

LEMMA 5.1 (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*

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

Proof of Lemma 4.5. 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 $i \in \mathbb{N}$, $\mathrm{d}(t, \Pi_i(t)) \le \theta_i$. By the triangle inequality, this implies that, for all $t \in \mathcal{T}$ and $i \in \mathbb{N}$,

$$\mathrm{d}(\Pi_i(t), \Pi_{i+1}(t)) \le \theta_i + \theta_{i+1} = 3\theta_{i+1}. \tag{69}$$

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,

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in\mathcal{T}} (Z_t - Z_{t_0})^2\right] &\le \sum_{i=0}^{\infty}\sum_{k=0}^{\infty} \mathbb{E}\left[\sup_{t\in\mathcal{T}} \{\Delta_i(t)\Delta_k(t)\}\right] \\
&\le \sum_{i=0}^{\infty}\sum_{k=0}^{\infty} \left|\sup_{t\in\mathcal{T}} |\Delta_i(t)|\right|_2 \cdot \left|\sup_{t\in\mathcal{T}} |\Delta_k(t)|\right|_2 \\
&= \left[\sum_{i=0}^{\infty} \left|\sup_{t\in\mathcal{T}} |\Delta_i(t)|\right|_2\right]^2, \qquad (70)
\end{aligned}$$

using Hölder's inequality in the second inequality.

Fix $i \in \mathbb{N}$. Since $N(\theta_i, \mathcal{T}) \le N(\theta_{i+1}, \mathcal{T})$, we have that

$$|\{(\Pi_i(t), \Pi_{i+1}(t)) : t \in \mathcal{T}\}| \le N(\theta_{i+1}, \mathcal{T})^2 = e^{2H(\theta_{i+1})}. \tag{71}$$

Relations (42) and (69) imply that, for all $t \in \mathcal{T}$,

$$\ln \mathbb{E}\left[e^{\lambda\Delta_i(t)}\right] \le a\, \mathrm{d}\left(\Pi_i(t), \Pi_{i+1}(t)\right)^{\delta} \lambda + \frac{v\, \mathrm{d}\left(\Pi_i(t), \Pi_{i+1}(t)\right)^{2\delta} \lambda^2}{2} \le a_i\lambda + \frac{v_i\lambda^2}{2},$$

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 4.2, the bound $\Delta_i(t)^2 \le 2[\Delta_i(t) - a_i]^2 + 2a_i^2$, and the change of variables $\lambda \mapsto 2\lambda$ imply that, for all $t \in \mathcal{T}$ and $0 < \lambda < \frac{1}{4v_i}$,

$$\ln \mathbb{E}\left[e^{\lambda\Delta_i(t)^2}\right] \le 2(a_i^2 + v_i)\lambda + \frac{4v_i^2\lambda^2}{(1 - 4v_i\lambda)}, \tag{72}$$

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 (71)–(72) and Lemma 5.1 imply further that

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in\mathcal{T}} \Delta_i(t)^2\right] &\le 2(a_i^2 + v_i) + \sqrt{2 \cdot 8v_i^2 \cdot 2H(\theta_{i+1}, \mathcal{T})} + 4v_i \cdot 2H(\theta_{i+1}, \mathcal{T}) \\ &\le 2 \cdot 9^\delta (a^2 + v) \left[\theta_{i+1}^{2\delta} + \theta_{i+1}^{2\delta}\sqrt{8H(\theta_{i+1}, \mathcal{T})} + 4\theta_{i+1}^{2\delta} H(\theta_{i+1}, \mathcal{T})\right].\end{aligned}$$

Taking the square root in the above relation, we get

(73)
$$\left\|\sup_{t\in\mathcal{T}} |\Delta_i(t)|\right\|_2 \le 3^\delta \sqrt{2(a^2 + v)} \left[\theta_{i+1}^\delta + \theta_{i+1}^\delta \sqrt[4]{8H(\theta_{i+1}, \mathcal{T})} + 2\theta_{i+1}^\delta \sqrt{H(\theta_{i+1}, \mathcal{T})}\right].$$

We now take the square root in (70) and use (73), valid for any $i \in \mathbb{N}$, obtaining

$$\left\|\sup_{t\in\mathcal{T}} Z_t - Z_{t_0}\right\|_2 \le 3^\delta \sqrt{2(a^2 + v)} \left[\sum_{i=1}^\infty \theta_i^\delta + \sum_{i=1}^\infty \theta_i^\delta \sqrt[4]{8H(\theta_i, \mathcal{T})} + 2\sum_{i=1}^\infty \theta_i^\delta \sqrt{H(\theta_i, \mathcal{T})}\right].$$

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. ∎

Acknowledgment. The authors would like to thank the referees for improving the presentation of the paper.

REFERENCES

[1] A. Agarwal, P. Barlett, P. Ravikumar, and M. J. Wainwright, *Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization*, IEEE Trans. Inform. Theory, 58 (2012), pp. 3235–3249.

[2] L. Armijo, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

[3] F. Bach and E. Moulines, *Non-asymptotic analysis of stochastic approximation algorithms for machine learning*, in Advances in Neural Information Processing Systems (NIPS), MIT Press, Cambridge, MA, 2011, pp. 451–459.

[4] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, SIAM Rev., 60 (2018), pp. 223–311, https://doi.org/10.1137/16M1080173.

[5] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, Oxford, UK, 2013.

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

[7] Y. Chen, G. Lan, and Y. Ouyang, *Accelerated schemes for a class of variational inequalities*, Math. Program., 165 (2017), pp. 113–149.

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

[9] J.-B. Hiriart-Urruty, *Algorithmes stochastiques de résolution d'équations et d'inéquations variationnelles*, Z. Wahrsch. Verw. Gebiete, 33 (1975/76), pp. 167–186.

[10] D. Hsu, S. M. Kakade, and T. Zhang, *A tail inequality for quadratic forms of subgaussian random vectors*, Electron. Commun. Probab., 17 (2012), pp. 1–6.

[11] A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson, *Extragradient method with variance reduction for stochastic variational inequalities*, SIAM J. Optim., 27 (2017), pp. 686–724, https://doi.org/10.1137/15M1031953.

[12] A. N. Iusem, A. Jofré, and P. Thompson, *Incremental constraint projection methods for monotone stochastic variational inequalities*, Math. Oper. Res., to appear.

[13] A. N. Iusem and B. F. Svaiter, *A variant of Korpelevich's method for variational inequalities with a new search strategy*, Optimization, 42 (1997), pp. 309–321.

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

[15] A. Juditsky, A. Nemirovski, and C. Tauvel, *Solving variational inequalities with stochastic mirror-prox algorithm*, Stoch. Syst., 1 (2011), pp. 17–58.
[16] A. Kannan and U. V. Shanbhag, *Optimal Stochastic Extragradient Schemes for Pseudomonotone Stochastic Variational Inequality Problems and Their Variants*, preprint, https://arxiv.org/abs/1410.1628, 2017.
[17] E. N. Khobotov, *Modifications of the extragradient method for solving variational inequalities and certain optimization problems*, U.S.S.R. Comput. Math. and Math. Phys., 27 (1987), pp. 120–127.
[18] J. Koshal, A. Nedić, and U. V. Shanbhag, *Regularized iterative stochastic approximation methods for stochastic variational inequality problems*, IEEE Trans. Automat. Control, 58 (2013), pp. 594–609.
[19] N. Krejić, Z. Luzanin, F. Nikolovski, and I. Stojkovska, *A nonmonotone line search method for noisy minimization*, Optim. Lett., 9 (2015), pp. 1371–1391.
[20] N. Krejić, Z. Luzanin, Z. Ovcin, and I. Stojkovska, *Descent direction method with line search for unconstrained optimization in noisy environment*, Optim. Methods Softw., 30 (2015), pp. 1164–1184.
[21] D. Maclaurin, D. Duvenaud, and R. P. Adams, *Gradient-based hyperparameter optimization through reversible learning*, in ICML'15: Proceedings of the 32nd International Conference on Machine Learning, Vol. 37, 2015, pp. 2113–2122.
[22] M. Mahsereci and P. Hennig, *Probabilistic line searches for stochastic optimization*, in Advances in Neural Information Processing Systems (NIPS), MIT Press, Cambridge, MA, 2015, pp. 181–189.
[23] C. Marinelli and M. Röckner, *On the maximal inequalities of Burkholder, Davis and Gundy*, Expo. Math., 34 (2016), pp. 1–26.
[24] P.-Y. Massé and Y. Ollivier, *Speed Learning on the Fly*, preprint, https://arxiv.org/abs/1511.02540, 2015.
[25] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609, https://doi.org/10.1137/070704277.
[26] D. Panchenko, *Symmetrization approach to concentration inequalities for empirical processes*, Ann. Probab., 31 (2003), pp. 2068–2081.
[27] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855, https://doi.org/10.1137/0330046.
[28] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Statistics, 22 (1951), pp. 400–407.
[29] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.
[30] T. Schaul, S. Zhang, and Y. Lecun, *No more pesky learning rates*, in ICML'13: Proceedings of the 30th International Conference on Machine Learning, Vol. 28, 2013, pp. 343–351.
[31] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, SIAM, Philadelphia, 2009, https://doi.org/10.1137/1.9780898718751.
[32] V. A. Steklov, *Sur les expressions asymptotiques de certaines fonctions définies par les équations différentielles du second ordre, et leurs applications au problème du developpement d'une fonction arbitraire en series procédant suivant les-dites fonctions*, Comm. Charkov Math. Soc., 2 (1907), pp. 97–199.
[33] C. Tan, S. Ma, Y.-H. Dai, and Y. Qian, *Barzilai-Borwein Step Size for Stochastic Gradient Descent*, preprint, https://arxiv.org/abs/1605.04131, 2016.
[34] Y. Wardi, *Stochastic algorithms with Armijo stepsizes for minimization of functions*, J. Optim. Theory Appl., 64 (1990), pp. 399–417.
[35] F. Yousefian, A. Nedić, and U. V. Shanbhag, *On stochastic gradient and subgradient methods with adaptive steplength sequences*, Automatica J. IFAC, 48 (2012), pp. 56–67.
[36] F. Yousefian, A. Nedić, and U. V. Shanbhag, *On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems*, Math. Program., 165 (2017), pp. 391–431.