Computational Optimization and Applications
https://doi.org/10.1007/s10589-022-00399-3

Stochastic relaxed inertial forward-backward-forward splitting for monotone inclusions in Hilbert spaces

Shisheng Cui[1] · Uday Shanbhag[1] · Mathias Staudigl[2] · Phan Vuong[3]

Received: 2 November 2021 / Accepted: 30 June 2022

Abstract

We consider monotone inclusions defined on a Hilbert space where the operator is given by the sum of a maximal monotone operator T and a single-valued monotone, Lipschitz continuous, and expectation-valued operator V. We draw motivation from the seminal work by Attouch and Cabot (Attouch in AMO 80:547–598, 2019, Attouch in MP 184: 243–287) on relaxed inertial methods for monotone inclusions and present a stochastic extension of the relaxed inertial forward–backward-forward method. Facilitated by an online variance reduction strategy via a mini-batch approach, we show that our method produces a sequence that weakly converges to the solution set. Moreover, it is possible to estimate the rate at which the discrete velocity of the stochastic process vanishes. Under strong monotonicity, we demonstrate strong convergence, and give a detailed assessment of the iteration and oracle complexity of the scheme. When the mini-batch is raised at a geometric (polynomial) rate, the rate statement can be strengthened to a linear (suitable polynomial) rate while the oracle complexity of computing an ϵ-solution improves to $\mathcal{O}(1/\epsilon)$. Importantly, the latter claim allows for possibly biased oracles, a key theoretical advancement allowing for far broader applicability. By defining a restricted gap function based on the Fitzpatrick function, we prove that the expected gap of an averaged sequence diminishes at a sublinear rate of $\mathcal{O}(1/k)$ while the oracle complexity of computing a suitably defined ϵ-solution is $\mathcal{O}(1/\epsilon^{1+a})$ where $a > 1$. Numerical results on two-stage games and an overlapping group Lasso problem illustrate the advantages of our method compared to competitors.

Keywords Monotone operator splitting · Stochastic approximation · Complexity · Variance reduction · Dynamic sampling

✉ Mathias Staudigl
m.staudigl@maastrichtuniversity.nl

Extended author information available on the last page of the article

Published online: 30 July 2022

1 Introduction

1.1 Problem formulation and motivation

A wide range of problems in areas such as optimization, variational inequalities, game theory, signal processing, or traffic theory, can be reduced to solving inclusions involving set-valued operators in a Hilbert space H, i.e. to find a point $x \in \mathsf{H}$ such that $0 \in F(x)$, where $F : \mathsf{H} \to 2^{\mathsf{H}}$ is a set-valued operator. In many applications such inclusion problems display specific structure revealing that the operator F can be additively decomposed. This leads us to the main problem we consider in this paper.

Problem 1 Let H be a real separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and associated norm $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. Let $T : \mathsf{H} \to 2^{\mathsf{H}}$ and $V : \mathsf{H} \to \mathsf{H}$ be maximally monotone operators, such that V is L-Lipschitz continuous. The problem is to

$$\text{find } x \in \mathsf{H} \text{ such that } 0 \in F(x) \triangleq V(x) + T(x), \tag{MI}$$

We assume that Problem 1 is well-posed:

Assumption 1 $\mathsf{S} \triangleq \mathsf{Zer}(F) \neq \emptyset$.

We are interested in the case where (MI) is solved by an iterative algorithm based on a *stochastic oracle (SO) representation* of the operator V. Specifically, when solving the problem, the algorithm calls to the SO. At each call, the SO receives as input a search point $x \in \mathsf{H}$ generated by the algorithm on the basis of past information so far, and returns the output $\hat{V}(x, \xi)$, where ξ is a random variable defined on some given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, taking values in a measurable set Ξ with law $\mathsf{P} = \mathbb{P} \circ \xi^{-1}$. In most parts of this paper, and the vast majority of contributions on stochastic variational problems in general, it is assumed that the output of the SO is unbiased,

$$V(x) = \mathbb{E}_{\xi}[\hat{V}(x, \xi)] = \int_{\Xi} \hat{V}(x, z)\, d\mathsf{P}(z) \qquad \forall x \in \mathsf{H}. \tag{1}$$

Such stochastic inclusion problems arise in numerous problems of fundamental importance in mathematical optimization and equilibrium problems, either directly or through an appropriate reformulation. An excellent survey on the existing techniques for solving problem (MI) can be found in [3] (in general Hilbert spaces) and [4] (in the finite-dimensional case).

1.2 Motivating examples

In what follows, we provide some motivating examples.

Example 1 (Stochastic Convex Optimization) Let $\mathsf{H}_1, \mathsf{H}_2$ be separable Hilbert spaces. A large class of stochastic optimization problems, with wide range of applications in signal processing, machine learning and control, is given by

$$\min_{u \in \mathsf{H}_1} \{f(u) + h(u) + g(Lu)\} \tag{2}$$

where $h : \mathsf{H}_1 \to \mathbb{R}$ is a convex differentiable function with a Lipschitz continuous gradient ∇h, represented as $h(u) = \mathbb{E}_\xi[\hat{h}(u, \xi)]$. $f : \mathsf{H}_1 \to (-\infty, \infty]$ and $g : \mathsf{H}_2 \to (-\infty, \infty]$ are proper, convex lower semi-continuous functions, and $L : \mathsf{H}_1 \to \mathsf{H}_2$ is a bounded linear operator. Problem (2) gains particular relevance in machine learning, where usually $h(u)$ is a convex data fidelity term (e.g. a *population risk* functional), and $g(Lu)$ and $f(u)$ embody penalty or regularization terms; see e.g. total variation [5], hierarchical variable selection [6, 7], and graph regularization [8, 9]. Applications in control and engineering are given in [10, 11]. We refer to (2) as the *primal problem*. Using Fenchel-Rockafellar duality [3, ch.19], the dual problem of (2) is given by

$$\min_{v \in \mathsf{H}_2} \{(f + h)^*(-L^* v) + g^*(v)\}, \tag{3}$$

where g^* is the Fenchel conjugate of g and $(f + h)^*(w) = f^* \Box h^*(w) = \inf_{u \in \mathsf{H}_1} \{f^*(u) + h^*(w - u)\}$ represents the infimal convolution of the functions f and h. Combining the primal problem (2) with its dual (3), we obtain the saddle-point problem

$$\inf_{u \in \mathsf{H}_1} \sup_{v \in \mathsf{H}_2} \{f(u) + h(u) - g^*(v) + \langle Lu, v \rangle\}. \tag{4}$$

Following classical Karush-Kuhn-Tucker theory [12], the primal-dual optimality conditions associated with (4) are concisely represented by the following monotone inclusion: Find $\bar{x} = (\bar{u}, \bar{v}) \in \mathsf{H}_1 \times \mathsf{H}_2 \equiv \mathsf{H}$ such that

$$-L^* \bar{v} \in \partial f(\bar{u}) + \nabla h(\bar{u}), \text{ and } L\bar{u} \in \partial g^*(\bar{v}). \tag{5}$$

We may compactly summarize these conditions in terms of the zero-finding problem (MI) using the operators V and T, defined as

$$V(u, v) \triangleq (\nabla h(u) + L^* v, -Lu) \text{ and } T(u, v) \triangleq \partial f(u) \times \partial g^*(v).$$

Note that the operator $V : \mathsf{H} \to \mathsf{H}$ is the sum of a maximally monotone and a skew-symmetric operator. Hence, in general, it is not cocoercive. Conditions on the data guaranteeing Assumption 1 are stated in [13].

Since $h(u)$ is represented as an expected value, we need to appeal to simulation based methods to evaluate its gradient. Also, significant computational speedups can be made if we are able to sample the skew-symmetric linear operator $(u, v) \mapsto (L^* u, -Lu)$ in an efficient way. Hence, we assume that there exists a SO that can provide unbiased estimator to the gradient operators $\nabla h(u)$ and $(L^* v, -Lu)$. More specifically, given the current position $x = (u, v) \in \mathsf{H}_1 \times \mathsf{H}_2$, the oracle will output the random estimators $\hat{H}(u, \xi), \hat{L}_u(u, \xi), \hat{L}_v(v, \xi)$ such that

$$\mathbb{E}_{\xi}[\hat{H}(u,\xi)] = \nabla h(u),\ \mathbb{E}_{\xi}[\hat{L}_u(u,\xi)] = Lu,\ \text{and}\ \mathbb{E}_{\xi}[\hat{L}_v(v,\xi)] = L^*v.$$

This oracle feedback generates the *random operator* $\hat{V}(x,\xi) = (\hat{H}(u,\xi) + \hat{L}_v(v,\xi), -\hat{L}_u(u,\xi))$, which allows us to approach the saddle-point problem (4) via simulation-based techniques.

Example 2 (Stochastic variational inequality problems) There are a multitude of examples of monotone inclusion problems (MI) where the single-valued map V is not the gradient of a convex function. An important model class where this is the case is the *stochastic variational inequality (SVI)* problem. Due to their huge number of applications, SVI's received enormous interest over the last several years from various communities [14–17]. This problem emerges when $V(x)$ is represented as an expected value as in (1) and $T(x) = \partial g(x)$ for some proper lower semi-continuous function $g : \mathsf{H} \to (-\infty, \infty]$. In this case, the resulting structured monotone inclusion problem can be equivalently stated as

$$\text{find } \bar{x} \in \mathsf{H} \text{ s.t. } \langle V(\bar{x}), x - \bar{x}\rangle + g(x) - g(\bar{x}) \geq 0 \quad \forall x \in \mathsf{H}. \tag{6}$$

An important and frequently studied special case of (6) arises if g is the indicator function of a given closed and convex subset $\mathsf{C} \subset \mathsf{H}$. In this cases the set-valued operator T becomes the normal cone map

$$T(x) = \mathsf{N}_{\mathsf{C}}(x) \triangleq \begin{cases} \{p \in \mathsf{H} | \sup_{y \in \mathsf{C}} \langle y - x, p\rangle \leq 0\} & \text{if } x \in \mathsf{C}, \\ \emptyset & \text{else.} \end{cases} \tag{7}$$

This formulation includes many fundamental problems including fixed point problems, Nash equilibrium problems and complementarity problems [4]. Consequently, the equilibrium condition (6) reduces to

$$\text{find } \bar{x} \in \mathsf{C} \text{ s.t. } \langle V(\bar{x}), x - \bar{x}\rangle \geq 0 \quad \forall x \in \mathsf{C}.$$

1.3 Contributions

Despite the advances in stochastic optimization and variational inequalities, the algorithmic treatment of general monotone inclusion problems under stochastic uncertainty is a largely unexplored field. This is rather surprising given the vast amount of applications of maximally monotone inclusions in control and engineering, encompassing distributed computation of generalized Nash equilibria [18–20], traffic systems [21–23], and PDE-constrained optimization [24]. The first major aim of this manuscript is to introduce and investigate a relaxed inertial stochastic forward-backward-forward (RISFBF) method, building on an operator splitting scheme originally due to Paul Tseng [25]. RISFBF produces three sequences $\{(X_k, Y_k, Z_k); k \in \mathbb{N}\}$, defined as

$$\begin{aligned} Z_k &= X_k + \alpha_k(X_k - X_{k-1}), \\ Y_k &= J_{\lambda_k T}(Z_k - \lambda_k A_k(Z_k)), \\ X_{k+1} &= (1-\rho_k)Z_k + \rho_k[Y_k + \lambda_k(A_k(Z_k) - B_k(Y_k))]. \end{aligned} \tag{RISFBF}$$

The data involved in this scheme are explained as follows:

- $A_k(Z_k)$ and $B_k(Y_k)$ are random estimators of V obtained by consulting the SO at search points Z_k and Y_k, respectively;
- $(\alpha_k)_{k\in\mathbb{N}}$ is a sequence of non-negative numbers regulating the memory, or *inertia* of the method;
- $(\lambda_k)_{k\in\mathbb{N}}$ is a positive sequence of step-sizes;
- $(\rho_k)_{k\in\mathbb{N}}$ is a non-negative *relaxation* sequence.

If $\alpha_k = 0$ and $\rho_k = 1$ the above scheme reduces to the stochastic forward-backward-forward method developed in [26, 27], with important applications in Gaussian communication networks [16] and dynamic user equilibrium problems [28]. However, even more connections to existing methods can be made.

Stochastic Extragradient If $T = \{0\}$, we obtain the inertial extragradient method

$$\begin{aligned} Z_k &= X_k + \alpha_k(X_k - X_{k-1}), \\ Y_k &= Z_k - \lambda_k A_k(Z_k), \\ X_{k+1} &= Z_k - \rho_k \lambda_k B_k(Y_k). \end{aligned}$$

If $\alpha_k = 0$, this reduces to a generalized extragradient method

$$\begin{aligned} Y_k &= X_k - \lambda_k A_k(X_k), \\ X_{k+1} &= X_k - \lambda_k \rho_k B_k(Y_k), \end{aligned}$$

recently introduced in [29].

Proximal Point Method If $V = 0$, the method reduces to the well-known deterministic proximal point algorithm [2], overlaid by inertial and relaxation effects. The scheme reads explicitly as

$$\begin{aligned} Z_k &= X_k + \alpha_k(X_k - X_{k-1}), \\ X_{k+1} &= (1-\rho_k)Z_k + \rho_k J_{\lambda_k T}(Z_k). \end{aligned}$$

The list of our contributions reads as follows:

(i) *Wide Applicability* A key argument in favor of Tseng's operator splitting method is that it is provably convergent when solving structured monotone inclusions of the type (MI), without imposing *cocoercivity* of the single-valued part V. This is a remarkable advantage relative to the perhaps more familiar and direct forward-backward splitting methods (aka projected (stochastic) gradient descent in the potential case). In particular, our scheme is applicable to the primal-dual splitting described in Example 1.

(ii) *Asymptotic guarantees* We show that under suitable assumptions on the relaxation sequence $(\rho_k)_{k\in\mathbb{N}}$, the non-decreasing inertial sequence $(\alpha_k)_{k\in\mathbb{N}}$, and step-length sequence $(\lambda_k)_{k\in\mathbb{N}}$, the generated stochastic process $(X_k)_{k\in\mathbb{N}}$ weakly almost surely converges to a random variable with values in S. Assuming demiregularity of the operators yields strong convergence in the real (possibly infinite-dimensional) Hilbert space.

(iii) *Non-asymptotic linear rate under strong monotonicity of V* When V is strongly monotone, strong convergence of the last iterate is shown and the sequence admits a non-asymptotic linear rate of convergence *without a conditional unbiasedness of the SO*. In particular, we show that the iteration and oracle complexity of computing an ϵ-solution is no worse than $\mathcal{O}(\log(\frac{1}{\epsilon}))$ and $\mathcal{O}(\frac{1}{\epsilon})$, respectively.

(iv) *Non-asymptotic sublinear rate under monotonicity of V* When V is monotone, by leveraging the *Fitzpatrick function* [3, 30, 31] associated with the structured operator $F = T + V$, we propose a restricted gap function. We then prove that the expected gap of an averaged sequence diminishes at the rate of $\mathcal{O}(\frac{1}{k})$. This allows us to derive an $\mathcal{O}(\frac{1}{\epsilon})$ upper bound on the iteration complexity, and an $\mathcal{O}(\frac{1}{\epsilon^{2+\delta}})$ upper bound (for $\delta > 0$) on the oracle complexity for computing an ϵ-solution.

The above listed contributions shed new light on a set of open questions, which we summarize below:

(i) *Absence of rigorous asymptotics* So far no aymptotic convergence guarantees have been available when considering relaxed inertial FBF schemes when T is maximally monotone and V is a single-valued monotone expectation-valued map.

(ii) *Unavailability of rate statements* We are not aware of any known non-asymptotic rate guarantees for algorithms solving (MI) under stochastic uncertainty. A key barrier in monotone and stochastic regimes in developing such statements has been in the availability of a residual function. Some recent progress in the special stochastic variational inequality case has been made by [26, 32, 33], but the general Hilbert-space setting involving set-valued operators seems to be largely unexplored (we will say more in Sect. 1.4).

(iii) *Bias requirements* A standard assumption in stochastic optimization is that the SO generates signals which are unbiased estimators of the deterministic operator $V(x)$. Of course, the requirement that the noise process is unbiased may often fail to hold in practice. In the present Hilbert space setting this is in some sense even expected to be the rule rather than the exception, since most operators are derived from complicated dynamical systems or the optimization method is applied to discretized formulations of the original problem. See the recent work [34, 35] for an interesting illustration in the context of PDE-constrained optimization. Some of our results go beyond the standard unbiasedness assumption.

1.4 Related research

Understanding the role of inertial and relaxation effects in numerical schemes is a line of research which received enormous interest over the last two decades. Below, we try to give a brief overview about related algorithms.

Inertial, Relaxation, and Proximal schemes

In the context of convex optimization, Polyak [36] introduced the *Heavy-ball method*. This is a two-step method for minimizing a smooth convex function *f*. The algorithm reads as

$$\begin{cases} Z_k = X_k + \alpha_k(X_k - X_{k-1}), \\ X_{k+1} = Z_k - \lambda_k \nabla f(X_k) \end{cases} \tag{HB}$$

The difference from the gradient method is that the base point of the gradient descent step is taken to be the extrapolated point Z_k, instead of X_k. This small difference has the surprising consequence that (HB) attains optimal complexity guarantees for strongly convex functions with Lipschitz continuous gradients. Hence, (HB) resembles an optimal method [37]. The acceleration effects can be explained by writing the process entirely in terms of a single updating equation as

$$X_{k+1} - 2X_k - X_{k-1} + (1-\alpha_k)(X_k - X_{k-1}) + \lambda_k \nabla f(X_k) = 0.$$

Choosing $\alpha_k = 1 - a_k\delta_k$ and $\lambda_k = \gamma_k\delta_k^2$ for δ_k a small parameter, we arrive at

$$\frac{1}{\delta_k^2}(X_{k+1} - 2X_k - X_{k-1}) + \frac{a_k}{\delta_k}(X_k - X_{k-1}) + \gamma_k \nabla f(X_k) = 0.$$

This can be seen as a discrete-time approximation of the second-order dynamical system

$$\ddot{x}(t) + \frac{a}{t}\dot{x}(t) + \gamma(t)\nabla f(x(t)) = 0,$$

introduced by [38]. Since then, it has received significant attention in the potential, as well as in the non-potential case (see e.g [39–41] for an appetizer). As pointed out in [42], if $\gamma(t) = 1$, the above system reduces to a continuous version of Nesterov's fast gradient method [43]. Recently, [44] defined a stochastic version of the Heavy-ball method.

Motivated by the development of such fast methods for convex optimization, Attouch and Cabot [1] studied a *relaxed-inertial forward-backward algorithm*, reading as

$$\begin{cases} Z_k = X_k + \alpha_k(X_k - X_{k-1}), \\ Y_k = J_{\lambda_k T}(Z_k - \lambda_k V(Z_k)) \\ X_{k+1} = (1-\rho_k)Z_k + \rho_k Y_k. \end{cases} \tag{RIFB}$$

If $V = 0$, this reduces to a relaxed inertial proximal point method analyzed by Attouch and Cabot [2]. If $\rho_k = 1$, an inertial forward-backward splitting method is recovered, first studied by Lorenz and Pock [45].

Convergence guarantees for the forward-backward splitting rely on the cocoercivity (inverse strong monotonicity) of the single-valued operator V. Example 1, in which V is given by a monotone plus a skew-symmetric linear operator, illustrates an important instance for which this assumption is not satisfied (see [46] for further examples). A general-purpose operator splitting framework, relaxing the cocoercivity property, is the forward-backward-forward (FBF) method due to Tseng [25]. Inertial [47] and relaxed-inertial [48] versions of FBF have been developed. An all-encompassing numerical scheme can be compactly described as

$$\begin{cases} Z_k = X_k + \alpha_k(X_k - X_{k-1}), \\ Y_k = J_{\lambda_k T}(Z_k - \lambda_k V(Z_k)), \\ X_{k+1} = (1-\rho_k)Z_k + \rho_k[Y_k - \lambda_k(V(Y_k) - V(Z_k))]. \end{cases} \tag{RIFBF}$$

Weak and strong convergence under appropriate conditions on the involved operators and parameter sequences are established in [48], but no rate statements are given.

Related work on stochastic approximation Efforts in extending stochastic approximation methods to variational inequality problems have considered standard projection schemes [14] for Lipschitz and strongly monotone operators. Extragradient and (more generally) mirror-prox algorithms [49, 50] can contend with merely monotone operators, while iterative smoothing [51] schemes can cope with with the lack of Lipschitz continuity. It is worth noting that extragradient schemes have recently assumed relevance in the training of generative adversarial networks (GANS) [52, 53]. Rate analysis for stochastic extragradient (SEG) have led to optimal rates for Lipschitz and monotone operators [50], as well as extensions to non-Lipschitzian [51] and pseudomonotone settings [32, 54]. To alleviate the computational complexity single-projection schemes, such as the stochastic forward-backward-forward (SFBF) method [26, 27], as well as subgradient-extragradient and projected reflected algorithms [55] have been studied as well.

SFBF has been shown to be nearly optimal in terms of iteration and oracle complexity, displaying significant empirical improvements compared to SEG. While the role of inertia in optimization is well documented, in stochastic splitting problems, the only contribution we are aware of is the work by Rosasco et al. [56]. In that paper asymptotic guarantees for an inertial stochastic forward-backward (SFB) algorithm are presented under the hypothesis that the operators V and T are maximally monotone and the single-valued operator V is cocoercive.

Variance reduction approaches Variance-reduction schemes address the deterioration in convergence rate and the resulting poorer practical behavior via two commonly adopted avenues:

(i) If the single-valued part V appears as a finite-sum (see e.g. [52, 57]), variance-reduction ideas from machine learning [58] can be used.

(ii) Mini-batch schemes that employ an increasing batch-size of gradients [59] lead to deterministic rates of convergence for stochastic strongly convex [60], convex [61], and nonconvex optimization [62], as well as for pseudo-monotone SVIs via extragradient [32], and splitting schemes [26].

In terms of run-time, improvements in iteration complexities achieved by mini-batch approaches are significant; e.g. in strongly monotone regimes, the iteration complexity improves from $\mathcal{O}(\frac{1}{\epsilon})$ to $\mathcal{O}(\ln(\frac{1}{\epsilon}))$ [27, 55]. Beyond run-time advantages, such avenues provide asymptotic and rate guarantees under possibly weaker assumptions on the problem as well as the oracle; in particular, mini-batch schemes allow for possibly biased oracles and state-dependency of the noise [55]. Concerns about the sampling burdens are, in our opinion, often overstated since such schemes are meant to provide ϵ-solutions; e.g. if $\epsilon = 10^{-3}$ and the obtained rate is $\mathcal{O}(1/k)$, then the batch-size $m_k = \lfloor k^a \rfloor$ where $a > 1$, implying that the batch-sizes are $\mathcal{O}(10^{3a})$, a relatively modest requirement, given the advances in computing.

Outline The remainder of the paper is organized in five sections. After dispensing with the preliminaries in Sect. 2, we present the (RISFBF) scheme in Sect. 3. Asymptotic and rate statements are developed in Sect. 4 and preliminary numerics are presented in Sect. 5. We conclude with some brief remarks in Sect. 6. Technical results are collected in Appendix 1.

2 Preliminaries

Throughout, H is a real separable Hilbert space with scalar product $\langle\cdot,\cdot\rangle$, norm $\|\cdot\|$, and Borel σ-algebra $\mathcal{B}$. The symbols $\to$ and $\rightharpoonup$ denote strong and weak convergence, respectively. $\mathrm{Id} : \mathsf{H} \to \mathsf{H}$ denotes the identity operator on H. Stochastic uncertainty is modeled on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, endowed with a filtration $\mathbb{F} = (\mathcal{F}_k)_{k\in\mathbb{N}_0}$. By means of the Kolmogorov extension theorem, we assume that $(\Omega, \mathcal{F}, \mathbb{P})$ is large enough so that all random variables we work with are defined on this space. A H-valued random variable is a measurable function $X : (\Omega, \mathcal{F}) \to (\mathsf{H}, \mathcal{B})$. Let $\mathcal{G} \subset \mathcal{F}$ be a given sub-sigma algebra. The conditional expecation of the random variable X is denoted by $\mathbb{E}(X|\mathcal{G})$. If $\mathcal{A} \subset \mathcal{G} \subset \mathcal{F}$, the tower-property says that

$$\mathbb{E}[\mathbb{E}(X|\mathcal{G})|\mathcal{A}] = \mathbb{E}[\mathbb{E}(X|\mathcal{A})|\mathcal{G}] = \mathbb{E}(X|\mathcal{A}).$$

We denote by $\ell^0(\mathbb{F})$ the set of sequences of real-valued random variables $(\xi_k)_{k\in\mathbb{N}}$ such that, for every $k \in \mathbb{N}$, ξ_k is $\mathcal{F}_k$-measurable. For $p \in [1, \infty]$, we set

$$\ell^p(\mathbb{F}) \triangleq \left\{ (\xi_k)_{k\in\mathbb{N}} \in \ell^0(\mathbb{F}) | \sum_{k\geq 1} |\xi_k|^p < \infty \quad \mathbb{P}\text{-a.s.} \right\}.$$

We denote the set of summable non-negative sequences by $\ell^1_+(\mathbb{N})$.

We now collect some concepts from monotone operator theory. For more details, we refer the reader to [3]. Let $F : \mathsf{H} \to 2^{\mathsf{H}}$ be a set-valued operator. Its domain and graph are defined as $\operatorname{dom} F \triangleq \{x \in \mathsf{H} | F(x) \neq \emptyset\}$, and $\operatorname{gr}(F) \triangleq \{(x,u) \in \mathsf{H} \times \mathsf{H} | u \in F(x)\}$, respectively. A single-valued operator $C : \mathsf{H} \to \mathsf{H}$ is *cocoercive* if there exists $\beta > 0$ such that $\langle C(x) - C(y), x - y\rangle \geq \beta\|C(x) - C(y)\|^2$. A set-valued operator $F : \mathsf{H} \to 2^{\mathsf{H}}$ is called *monotone* if

$$\langle v - w, x - y\rangle \geq 0 \qquad \forall (x,v),(y,w) \in \operatorname{gr}(F). \tag{8}$$

The set of zeros of F, denoted by $\mathsf{Zer}(T)$, defined as $\mathsf{Zer}(F) \triangleq \{x \in \mathsf{H} | 0 \in T(x)\}$. The inverse of F is $F^{-1} : \mathsf{H} \to 2^{\mathsf{H}}, u \mapsto F^{-1}(u) = \{x \in \mathsf{H} | u \in F(x)\}$. The resolvent of F is $J_F \triangleq (\operatorname{Id} + F)^{-1}$. If F is maximally monotone, then J_F is a single-valued map. We also need the classical notion of *demiregularity* of an operator.

Definition 1 An operator $F : \mathsf{H} \to 2^{\mathsf{H}}$ is demiregular at $x \in \operatorname{dom}(F)$ if for every sequence $\{(y_n, u_n)\}_{n\in\mathbb{N}} \subset \operatorname{gr}(F)$ and every $u \in F(y)$, we have

$$[y_n \rightharpoonup y, v_n \to v] \Rightarrow y_n \to y.$$

The notion of demiregularity captures various properties typically used to establish strong convergence of dynamical systems. [10] exhibits a large class of possibly set-valued operators F which are demiregular. In particular, demiregularity holds if F is uniformly or strongly monotone, or when F is the subdifferential of a uniformly convex lower semi-continuous function f. We often use the Young inequality

$$ab \leq \frac{a^2}{2\varepsilon} + \frac{\varepsilon b^2}{2} \quad (a, b \in \mathbb{R}). \tag{9}$$

3 Algorithm

Our aim is to solve the monotone inclusion problem (MI) under the following assumption:

Assumption 2 Consider Problem 1. The set-valued operator $T : \mathsf{H} \to 2^{\mathsf{H}}$ is maximally monotone with an efficiently computable resolvent. The single-valued operator $V : \mathsf{H} \to \mathsf{H}$ is maximally monotone and L-Lipschitz continuous ($L > 0$) with full domain $\operatorname{dom} V = \mathsf{H}$.

Assumption 2 guarantees that the operator $F = T + V$ is maximally monotone [3, Corolllary 24.4].

For numerical tractability, we make a *finite-dimensional noise* assumption, common to stochastic optimization problems in (possibly infinite-dimensional) Hilbert spaces [63].[1]

[1] Our analysis does not rely on this assumption. It is made here only for concreteness and because it is the most prevalent one in applications.

Assumption 3 (Finite-dimensional noise) All randomness can be described via a finite dimensional random variable $\xi : (\Omega, \mathcal{F}) \to (\Xi, \mathcal{E})$, where $\Xi \subseteq \mathbb{R}^d$ is a measurable set with Borel sigma algebra $\mathcal{E}$. The law of the random variable ξ is denoted by P, i.e. $\mathsf{P}(\Gamma) \triangleq \mathbb{P}(\{\omega \in \Omega | \xi(\omega) \in \Gamma\})$ for all $\Gamma \in \mathcal{E}$.

To access new information about the values of the operator $V(x)$, we adopt a *stochastic approximation* (SA) approach where samples are accessed iteratively and online: At each iteration, we assume to have access to a stochastic oracle (SO) which generates some estimate on the value of the deterministic operator $V(x)$ when the current position is x. This information is obtained by drawing an iid sample form the law P. These fresh samples are then used in the numerical algorithm after an initial extrapolation step delivering the point $Z_k = X_k + \alpha_k(X_k - X_{k-1})$, for some extrapolation coefficient $\alpha_k \in [0, 1]$. Departing from Z_k, we call the SO to retrieve the mini-batch estimator with *sample rate* $m_k \in \mathbb{N}$:

$$A_k(Z_k, \omega) \triangleq \frac{1}{m_k} \sum_{t=1}^{m_k} \hat{V}(Z_k, \xi_k^{(t)}(\omega)). \tag{10}$$

$\xi_k \triangleq (\xi_k^{(1)}, \dots, \xi_k^{(m_k)})$ is the data sample employed by the SO to return the estimator $A_k(Z_k)$. Subsequently we perform a forward-backward update with step size $\lambda_k > 0$:

$$Y_k = J_{\lambda_k T}\big(Z_k - \lambda_k A_k(Z_k)\big). \tag{11}$$

In the final updates, a second independent call of the SO is made, using the data set $\eta_k = (\eta_k^{(1)}, \dots, \eta_k^{(m_k)})$, yielding the estimator

$$B_k(Y_k, \omega) \triangleq \frac{1}{m_k} \sum_{t=1}^{m_k} \hat{V}(Y_k, \eta_k^{(t)}(\omega)), \tag{12}$$

and the new state

$$X_{k+1} = (1 - \rho_k)Z_k + \rho_k \big[Y_k + \lambda_k(A_k(Z_k) - B_k(Y_k))\big] \tag{13}$$

This iterative procedure generates a stochastic process $\{(Z_k, Y_k, X_k)\}_{k\in\mathbb{N}}$, defining the *relaxed inertial stochastic forward-backward-forward* (RISFBF) scheme. A pseudocode is given as Algorithm 1 below.

Algorithm 1 RISFBF

```
**Input:** $X_0, X_1 \in \operatorname{dom} T$, Terminal time $K \in \mathbb{N}$, nonnegative sequences $(\alpha_k)_{k=1}^K, (\lambda_k)_{k=1}^K, (\rho_k)_{k=1}^K, (m_k)_{k=1}^K$. $k = 1, \dots, K$
Compute $Z_k = X_k + \alpha_k(X_k - X_{k-1})$
Draw an iid sample $\{\xi_k^{(1)}, \dots, \xi_k^{(m_k)}\}$ from the law $\mathsf{P}$
Compute $A_k(Z_k)$ and $Y_k$ as in (10) and (11)
Draw an iid sample $\{\eta_k^{(1)}, \dots, \eta_k^{(m_k)}\}$ from the law $\mathsf{P}$
Compute $B_k(Y_k)$ and $X_{k+1}$ as in (12) and (13).
Report
    $\bar{X}_K = \sum_{k=1}^K \frac{\rho_k Y_k}{\sum_{k=1}^K \rho_k}.$    (14)
```

$$\bar{X}_k = \sum_{k=1}^K \frac{\rho_k Y_k}{\sum_{k=1}^K \rho k} \tag{14}$$

Note that RISFBF is still conceptual since we have not explained how the sequences $(\alpha_k)_{k\in\mathbb{N}}, (\lambda_k)_{k\in\mathbb{N}}$ and $(\rho_k)_{k\in\mathbb{N}}$ should be chosen. We will make this precise in our complexity analysis, starting in Sect. 4.

3.1 Equivalent form of RISFBF

We can collect the sequential updates of RISFBF as the fixed-point iteration

$$\begin{cases} Z_k = X_k + \alpha_k(X_k - X_{k-1}), \\ X_{k+1} = Z_k - \rho_k \Phi_{k,\lambda_k}(Z_k) \end{cases} \tag{15}$$

where $\Phi_{k,\lambda} : \mathsf{H} \times \Omega \to \mathsf{H}$ is the time-varying map given by

$$\Phi_{k,\lambda}(x, \omega) \triangleq x - \lambda A_k(x, \omega) - (\operatorname{Id}_{\mathsf{H}} - \lambda B_k(\cdot, \omega)) \circ J_{\lambda T} \circ (\operatorname{Id}_{\mathsf{H}} - \lambda A_k(\cdot, \omega))(x).$$

Formulating the algorithm in this specific way establishes the connection between RISFBF and the heavy-ball system. Indeed, combining the iterations in (15) in one, we get a second-order difference equations, closely resembling the structure present in (HB):

$$\frac{1}{\rho_k}(X_{k+1} - 2X_k - X_{k-1}) + \frac{(1-\alpha_k)}{\rho_k}(X_k - X_{k-1}) + \Phi_{k,\lambda_k}(X_k + \alpha_k(X_k - X_{k-1})) = 0.$$

Also, it reveals the Markovian nature of the process $(X_k)_{k\in\mathbb{N}}$; It is clear from the formulation (15) that X_k is Markov with respect to the sigma-algebra $\sigma(\{X_0, \dots, X_{k-1}\})$.

3.2 Assumptions on the stochastic oracle

In order to tame the stochastic uncertainty in RISFBF, we need to impose some assumptions on the distributional properties of the random fields $(A_k(x))_{k\in\mathbb{N}}$ and $(B_k(x))_{k\in\mathbb{N}}$. One crucial statistic we need to control is the SO variance. Define the *oracle error* at a point $x \in \mathsf{H}$ as

$$\varepsilon(x,\xi) \triangleq \hat{V}(x,\xi) - V(x). \tag{16}$$

Assumption 4 (Oracle Noise) We say that the SO

(i) is conditionally unbiased if $\mathbb{E}_{\xi}[\varepsilon(x,\xi)|x] = 0$ for all $x \in \mathsf{H}$;
(ii) enjoys a uniform variance bound: $\mathbb{E}_{\xi}[\|\varepsilon(x,\xi)\|^2|x] \leq \sigma^2$ for some $\sigma > 0$ and all $x \in \mathsf{H}$.

Define

$$U_k(\omega) \triangleq \frac{1}{m_k}\sum_{t=1}^{m_k} \varepsilon(Z_k(\omega), \xi_k^{(t)}(\omega)), \text{ and } W_k(\omega) \triangleq \frac{1}{m_k}\sum_{t=1}^{m_k} \varepsilon(Y_k(\omega), \eta_k^{(t)}(\omega)).$$

The introduction of these two processes allows us to decompose the random estimator into a mean component and a residual, so that

$$A_k(Z_k) = V(Z_k) + U_k, \text{ and } B_k(Y_k) = V(Y_k) + W_k$$

If Assumption 4(i) holds true then $\mathbb{E}[W_k|\hat{\mathcal{F}}_k] = 0 = \mathbb{E}[U_k|\mathcal{F}_k] = 0$. Hence, under conditional unbiasedness, the processes $\{(U_k,\mathcal{F}_k);k\in\mathbb{N}\}$ and $\{(W_k,\hat{\mathcal{F}}_k);k\in\mathbb{N}\}$ are martingale difference sequences, where the filtrations are defined as $\mathcal{F}_0 \triangleq \hat{\mathcal{F}}_0 \triangleq \mathcal{F}_1 \triangleq \sigma(X_0, X_1)$, and iteratively, for $k \geq 1$,

$$\hat{\mathcal{F}}_k \triangleq \sigma(X_0, X_1, \xi_1, \eta_1, \dots, \eta_{k-1}, \xi_k),\ \mathcal{F}_{k+1} \triangleq \sigma(X_0, X_1, \xi_1, \eta_1, \dots, \xi_k, \eta_k).$$

Observe that $\mathcal{F}_k \subseteq \hat{\mathcal{F}}_k \subseteq \mathcal{F}_{k+1}$ for all $k \geq 1$. The uniform variance bound, Assumption 4(ii), ensures that the processes $\{(U_k,\mathcal{F}_k);k\in\mathbb{N}\}$, $\{(W_k,\hat{\mathcal{F}}_k);k\in\mathbb{N}\}$ have finite second moment.

Remark 1 For deriving the stochastic estimates in the analysis to come, it is important to emphasize that X_k is $\mathcal{F}_k$-measurable for all $k \geq 0$, and Y_k is $\hat{\mathcal{F}}_k$-measurable.

The mini-batch sampling technology implies an online variance reduction effect, summarized in the next lemma, whose simple proof we omit.

Lemma 1 (Variance of the SO) *Suppose Assumption 4 holds. Then for* $k \geq 1$,

$$\mathbb{E}[\|W_k\|^2|\mathcal{F}_k] \leq \frac{\sigma^2}{m_k} \text{ and } \mathbb{E}[\|U_k\|^2|\mathcal{F}_k] \leq \frac{\sigma^2}{m_k}, \qquad \mathbb{P}-\text{a.s.} \tag{17}$$

We see that larger sampling rates lead to more precise point estimates of the single-valued operator. This comes at the cost of more evaluations of the stochastic operator. Hence, any mini-batch approach faces a trade-off between the *oracle complexity* and the iteration complexity. We want to use mini-batch estimators to achieve an online variance reduction scheme, motivating the next assumption.

Assumption 5 (Batch Size) The batch size sequence $(m_k)_{k\in\mathbb{N}}$ is non-decreasing and satisfies $\sum_{k=1}^{\infty}\frac{1}{m_k}<\infty$.

4 Analysis

This section is organized into three subsections. The first subsection derives asymptotic convergence guarantees, while the second and third subsections provides linear and sublinear rate statements in strongly monotone and monotone regimes, respectively.

4.1 Asymptotic convergence

Given $\lambda>0$, we define the residual function for the monotone inclusion (MI) as

$$\mathsf{res}_\lambda(x)\triangleq\left\|x-J_{\lambda T}(x-\lambda V(x))\right\|. \tag{18}$$

Clearly, for every $\lambda>0$, $x\in\mathsf{S}\Leftrightarrow\mathsf{res}_\lambda(x)=0$. Hence, $\mathsf{res}_\lambda(\cdot)$ is a *merit function* for the monotone inclusion problem. To put this merit function into context, let us consider the special case where T is the subdifferential of a lower semi-continuous convex function $g:\mathsf{H}\to(-\infty,\infty]$, i.e. $T=\partial g$. In this case, the resolvent $J_{\lambda T}$ reduces to the well-known *proximal-operator*

$$\operatorname{prox}_{\lambda g}(x)\triangleq\operatorname*{argmin}_{u\in\mathsf{H}}\{\lambda g(u)+\frac{1}{2}\|u-x\|^2\}.$$

In the potential case, where $V(x)=\nabla f(x)$ for some smooth convex function $f:\mathsf{H}\to\mathbb{R}$, the residual function is thus seen to be a constant multiple of the norm of the so-called *gradient mapping* $\left\|x-\operatorname{prox}_{\lambda g}(x-\lambda V(x))\right\|$, which is a standard merit function in convex [64] and stochastic [65, 66] optimization. We use this function to quantify the per-iteration progress of RISFBF. The main result of this subsection is the following.

Theorem 2 (Asymptotic Convergence) *Let $\bar{\alpha},\bar{\varepsilon}\in(0,1)$ be fixed parameters. Suppose that Assumption 1-5 hold true. Let $(\alpha_k)_{k\in\mathbb{N}}$ be a non-decreasing sequence such that $\lim_{k\to\infty}\alpha_k=\bar{\alpha}$. Let $(\lambda_k)_{k\in\mathbb{N}}$ be a converging sequence in $(0,\frac{1}{4L})$ such that $\lim_{k\to\infty}\lambda_k=\lambda\in(0,\frac{1}{4L})$. If $\rho_k=\frac{5(1-\bar{\varepsilon})(1-\bar{\alpha})^2}{4(2\alpha_k^2-\alpha_k+1)(1+L\lambda_k)}$ for all $k\geq 1$, then*

(i) $\lim_{k\to\infty}\mathsf{res}_{\lambda_k}(Z_k)=0$ *in* $L^2(\mathbb{P})$;

(ii) *the stochastic process* $(X_k)_{k\in\mathbb{N}}$ *generated by algorithm RISFBF weakly converges to a* S*-valued limiting random variable* X;

(iii) $\sum_{k=1}^{\infty}\left[(1-\alpha_k)\left(\frac{5(1-\alpha_k}{4\rho_k(1+L\lambda_k)}-1\right)-2\alpha_k^2\right]\|X_k-X_{k-1}\|^2<\infty \quad \mathbb{P}\text{-}a.s.$

We prove this Theorem via a sequence of technical Lemmas.

Lemma 3 *For all* $k\geq 1$, *we have*

$$-\|Z_k-Y_k\|^2\leq \lambda_k^2\|U_k\|^2-\frac{1}{2}\mathsf{res}^2_{\lambda_k}(Z_k). \tag{19}$$

Proof By definition,

$$\begin{aligned}\frac{1}{2}\mathsf{res}^2_{\lambda_k}(Z_k)&=\frac{1}{2}\left\|Z_k-J_{\lambda_k T}(Z_k-\lambda_k V(Z_k))\right\|^2\\&=\frac{1}{2}\left\|Z_k-Y_k+J_{\lambda_k T}(Z_k-\lambda_k A_k(Z_k))-J_{\lambda_k T}(Z_k-\lambda_k V(Z_k))\right\|^2\\&\leq\|Z_k-Y_k\|^2+\left\|J_{\lambda_k T}(Z_k-\lambda_k A_k(Z_k))-J_{\lambda_k T}(Z_k-\lambda_k V(Z_k))\right\|^2\\&\leq\|Z_k-Y_k\|^2+\lambda_k^2\|U_k\|^2,\end{aligned}$$

where the last inequality uses the non-expansivity property of the resolvent operator. Rearranging terms gives the claimed result. □

Next, for a given pair $(p,p^*)\in\operatorname{gr}(F)$, we define the stochastic processes $(\Delta M_k)_{k\in\mathbb{N}}$, $(\Delta N_k(p,p^*))_{k\in\mathbb{N}}$, and $(\mathsf{e}_k)_{k\in\mathbb{N}}$ as

$$\Delta M_k\triangleq\frac{5\rho_k\lambda_k^2}{2(1+L\lambda_k)}\|\mathsf{e}_k\|^2+\frac{\rho_k\lambda_k^2}{2}\|U_k\|^2, \tag{20}$$

$$\Delta N_k(p,p^*)\triangleq 2\rho_k\lambda_k\langle W_k+p^*,p-Y_k\rangle,\ \text{and} \tag{21}$$

$$\mathsf{e}_k\triangleq W_k-U_k. \tag{22}$$

Key to our analysis is the following energy bound on the evolution of the anchor sequence $\left(\|X_k-p\|^2\right)_{k\in\mathbb{N}}$.

Lemma 4 (Fundamental Recursion) *Let* $(X_k)_{k\in\mathbb{N}}$ *be the stochastic process generated by RISFBF with* $\alpha_k\in(0,1)$, $0\leq\rho_k<\frac{5}{4(1+L\lambda_k)}$, *and* $\lambda_k\in(0,1/4L)$. *For all* $k\geq 1$ *and* $(p,p^*)\in\operatorname{gr}(F)$, *we have*

$$\begin{aligned}\|X_{k+1}-p\|^2 \leq\,& (1+\alpha_k)\|X_k-p\|^2-\alpha_k\|X_{k-1}-p\|^2-\frac{\rho_k}{4}\mathsf{res}^2_{\lambda_k}(Z_k)\\&+\Delta M_k+\Delta N_k(p,p^*)-2\rho_k\lambda_k\langle V(Y_k)-V(p),Y_k-p\rangle\\&+\alpha_k\|X_k-X_{k-1}\|^2\left(2\alpha_k+\frac{5(1-\alpha_k)}{4\rho_k(1+L\lambda_k)}\right)\\&-(1-\alpha_k)\left(\frac{5}{4\rho_k(1+L\lambda_k)}-1\right)\|X_{k+1}-X_k\|^2.\end{aligned}$$

Proof To simplify the notation, let us call $A_k \equiv A_k(Z_k)$ and $B_k \equiv B_k(Y_k)$. We also introduce the intermediate update $R_k \triangleq Y_k+\lambda_k(A_k-B_k)$. For all $k \geq 0$, it holds true that

$$\begin{aligned}\|Z_k-p\|^2 =&\|Z_k-Y_k+Y_k-R_k+R_k-p\|^2\\=&\|Z_k-Y_k\|^2+\|Y_k-R_k\|^2+\|R_k-p\|^2+2\langle Z_k-Y_k,Y_k-p\rangle\\&+2\langle Y_k-R_k,R_k-p\rangle\\=&\|Z_k-Y_k\|^2+\|Y_k-R_k\|^2+\|R_k-p\|^2+2\langle Z_k-Y_k,Y_k-p\rangle\\&+2\langle Y_k-R_k,R_k-p\rangle\\=&\|Z_k-Y_k\|^2+\|Y_k-R_k\|^2+\|R_k-p\|^2+2\langle Z_k-Y_k,Y_k-p\rangle\\&+2\langle Y_k-R_k,Y_k-p\rangle+2\langle Y_k-R_k,R_k-Y_k\rangle\\=&\|Z_k-Y_k\|^2+\|Y_k-R_k\|^2+\|R_k-p\|^2+2\langle Z_k-R_k,Y_k-p\rangle\\&+2\langle Y_k-R_k,R_k-Y_k\rangle\\=&\|Z_k-Y_k\|^2-\|Y_k-R_k\|^2+\|R_k-p\|^2+2\langle Z_k-R_k,Y_k-p\rangle.\end{aligned}$$

Since

$$\begin{aligned}\|Y_k-R_k\|^2 &= \lambda_k^2\|B_k(Y_k)-Y_k(Z_k)\|^2\\&\leq \lambda_k^2\|V(Y_k)-V(Z_k)+W_{k+1}-U_{k+1}\|^2\\&\leq \lambda_k^2\|V(Y_k)-V(Z_k)\|^2+\lambda_k^2\|W_k-U_k\|^2+2\lambda_k^2\langle V(Y_k)-V(Z_k),W_k-U_k\rangle\\&\leq L^2\lambda_k^2\|Y_k-Z_k\|^2+\lambda_k^2\|W_k-U_k\|^2+2\lambda_k^2\langle V(Y_k)-V(Z_k),W_k-U_k\rangle\\&\leq 2L^2\lambda_k^2\|Y_k-Z_k\|^2+2\lambda_k^2\|W_k-U_k\|^2.\end{aligned}$$

Introducing the process $(\mathsf{e}_k)_{k\in\mathbb{N}}$ from eq. (22), the aforementioned set of inequalities reduces to

$$\|Y_k-R_k\|^2 \leq 2L^2\lambda_k^2\|Y_k-Z_k\|^2+2\lambda_k^2\|\mathsf{e}_k\|^2.$$

Hence,

$$\|Z_k-p\|^2 \geq (1-2L^2\lambda_k^2)\|Z_k-Y_k\|^2-2\lambda_k^2\|\mathsf{e}_k\|^2+\|R_k-p\|^2+2\langle Z_k-R_k,Y_k-p\rangle.$$

But $Y_k+\lambda_k T(Y_k) \ni Z_k-\lambda_k A_k$, implying that

$$\frac{1}{\lambda_k}(Z_k - Y_k - \lambda_k A_k) \in T(Y_k).$$

Pick $(p, p^*) \in \operatorname{gr}(F)$, so that $p^* - V(p) \in T(p)$. Then, the monotonicity of T yields the estimate

$$\left\langle \frac{1}{\lambda_k}(Z_k - Y_k - \lambda_k A_k) - p^* + V(p), Y_k - p \right\rangle \geq 0.$$

This is equivalent to

$$\begin{aligned} &\left\langle \frac{1}{\lambda_k}(Z_k - R_k - \lambda_k B_k) - p^* + V(p), Y_k - p \right\rangle \geq 0, \\ &\text{or } \langle Z_k - R_k, Y_k - p\rangle \geq \lambda_k \langle W_k + p^*, Y_k - p\rangle + \lambda_k \langle V(Y_k) - V(p), Y_k - p\rangle. \end{aligned} \tag{23}$$

This implies that

$$\langle Z_k - R_k, Y_k - x^* \rangle \geq \lambda_k \langle W_k, Y_k - x^* \rangle.$$

Hence, we obtain the following,

$$\begin{aligned} \|Z_k - p\|^2 \geq (1 - 2L^2\lambda_k^2)\|Y_k - Z_k\|^2 + \|R_k - p\|^2 - 2\lambda_k^2\|\mathsf{e}_k\|^2 \\ + 2\lambda_k\langle W_k + p^*, Y_k - p\rangle + 2\lambda_k\langle V(Y_k) - V(p), Y_k - p\rangle. \end{aligned}$$

Rearranging terms, we arrive at the following bound on $\|R_k - p\|^2$:

$$\begin{aligned} \|R_k - p\|^2 \leq &\|Z_k - p\|^2 - (1 - 2L^2\lambda_k^2)\|Y_k - Z_k\|^2 + 2\lambda_k^2\|\mathsf{e}_k\|^2 + 2\lambda_k\langle W_k + p^*, p - Y_k\rangle \\ &+ 2\lambda_k\langle V(Y_k) - V(p), p - Y_k\rangle \end{aligned} \tag{24}$$

Next, we observe that $\|X_{k+1} - p\|^2$ may be bounded as follows.

$$\begin{aligned} \|X_{k+1} - p\|^2 &= \|(1 - \rho_k)Z_k + \rho_k R_k - p\|^2 \\ &= \|(1 - \rho_k)(Z_k - p) - \rho_k(R_k - p)\|^2 \\ &= (1 - \rho_k)^2\|Z_k - p\|^2 + \rho_k^2\|R_k - p\|^2 - 2\rho_k(1 - \rho_k)\langle Z_k - p, R_k - p\rangle \\ &= (1 - \rho_k)\|Z_k - p\|^2 - \rho_k(1 - \rho_k)\|Z_k - p\|^2 \\ &\quad + \rho_k\|R_k - p\|^2 - \rho_k(1 - \rho_k)\|R_k - p\|^2 \\ &\quad + 2\rho_k(1 - \rho_k)\langle Z_k - p, R_k - p\rangle \\ &= (1 - \rho_k)\|Z_k - p\|^2 + \rho_k\|R_k - p\|^2 - \rho_k(1 - \rho_k)\|R_k - Z_k\|^2 \\ &= (1 - \rho_k)\|Z_k - p\|^2 + \rho_k\|R_k - p\|^2 - \frac{1 - \rho_k}{\rho_k}\|X_{k+1} - Z_k\|^2. \end{aligned} \tag{25}$$

We may then derive a bound on the expression in (25),

$$\begin{aligned}
&(1-\rho_k)\|Z_k-p\|^2+\rho_k\|R_k-p\|^2-\frac{1-\rho_k}{\rho_k}\|X_{k+1}-Z_k\|^2\\
&\quad\le \|Z_k-p\|^2-\frac{1-\rho_k}{\rho_k}\|X_{k+1}-Z_k\|^2-\rho_k(1-2L^2\lambda_k^2)\|Z_k-Y_k\|^2\\
&\qquad+2\lambda_k^2\rho_k\|\mathsf{e}_k\|^2-2\rho_k\lambda_k\langle W_k+p^*,Y_k-p\rangle+2\rho_k\lambda_k\langle V(Y_k)-V(p),p-Y_k\rangle
\end{aligned}\tag{26}$$

$$\begin{aligned}
&\quad=\|Z_k-p\|^2-\frac{1-\rho_k}{\rho_k}\|X_{k+1}-Z_k\|^2-\rho_k(1/2-2L^2\lambda_k^2)\|Z_k-Y_k\|^2\\
&\qquad+2\lambda_k^2\rho_k\|\mathsf{e}_k\|^2-2\rho_k\lambda_k\langle W_k+p^*,Y_k-p\rangle\\
&\qquad-\frac{\rho_k}{2}\|Y_k-Z_k\|^2+2\rho_k\lambda_k\langle V(Y_k)-V(p),p-Y_k\rangle.
\end{aligned}\tag{27}$$

By invoking (19), we arrive at the estimate

$$\begin{aligned}
\|X_{k+1}-p\|^2\le \|Z_k-p\|^2&-\frac{1-\rho_k}{\rho_k}\|X_{k+1}-Z_k\|^2+2\lambda^2\rho_k\|\mathsf{e}_k\|^2\\
&-2\rho_k\lambda_k\langle W_k+p^*,Y_k-p\rangle\\
&-\rho_k(1/2-2L^2\lambda_k^2)\|Y_k-Z_k\|^2-\frac{\rho_k}{4}\mathsf{res}^2_{\lambda_k}(Z_k)\\
&+\frac{\rho_k\lambda_k^2}{2}\|U_k\|^2+2\rho_k\lambda_k\langle V(Y_k)-V(p),p-Y_k\rangle.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\frac{1}{\rho_k}\|X_{k+1}-Z_k\|=\|R_k-Z_k\|&\le\|R_k-Y_k\|+\|Y_k-Z_k\|\\
&\le\lambda_k\|B_k-A_k\|+\|Y_k-Z_k\|\\
&\le(1+L\lambda_k)\|Y_k-Z_k\|+\lambda_k\|\mathsf{e}_k\|,
\end{aligned}$$

which implies that

$$\frac{1}{2\rho_k^2}\|X_{k+1}-Z_k\|^2\le(1+L\lambda_k)^2\|Y_k-Z_k\|^2+\lambda_k^2\|\mathsf{e}_k\|^2.\tag{28}$$

Multiplying both sides by $\frac{\rho_k(1/2-2L\lambda_k)}{1+L\lambda_k}$, a positive scalar since $\lambda_k\in(0,\frac{1}{4L})$, we obtain

$$\begin{aligned}
\frac{1/2-2L\lambda_k}{2\rho_k(1+L\lambda_k)}\|X_{k+1}-Z_k\|^2\le&\,\rho_k(1/2-2L\lambda_k)(1+L\lambda_k)\|Y_k-Z_k\|^2\\
&+\frac{\lambda_k^2\rho_k(1/2-2L\lambda_k)}{1+L\lambda_k}\|\mathsf{e}_k\|^2.
\end{aligned}\tag{29}$$

Rearranging terms, and noting that $(1/2-2L\lambda_k)(1+L\lambda_k)\le 1/2-2L^2\lambda_k^2$, the above estimate becomes

$$-\rho_k(1/2-2L^2\lambda_k^2)\|Y_k-Z_k\|^2 \leq -\frac{1/2-2L\lambda_k}{2\rho_k(1+L\lambda_k)}\|X_{k+1}-Z_k\|^2 + \frac{\rho_k\lambda_k^2(1/2-2L\lambda_k)}{1+L\lambda_k}\|\mathsf{e}_k\|^2. \tag{30}$$

Substituting this bound into the first majorization of the anchor process $\|X_{k+1}-p\|^2$, we see

$$\begin{aligned}\|X_{k+1}-p\|^2 \leq \|Z_k-p\|^2 &- \left(\frac{1-\rho_k}{\rho_k}+\frac{1/2-2L\lambda_k}{2\rho_k(1+L\lambda_k)}\right)\|X_{k+1}-Z_k\|^2\\ &+\rho_k\lambda_k^2\|\mathsf{e}_k\|^2\left(2+\frac{1/2-2L\lambda_k}{1+L\lambda_k}\right)+2\rho_k\lambda_k\langle V(Y_k)-V(p),p-Y_k\rangle\\ &-2\rho_k\lambda_k\langle W_k+p^*,Y_k-p\rangle-\frac{\rho_k}{4}\mathsf{res}^2_{\lambda_k}(Z_k)+\frac{\rho_k\lambda_k^2}{2}\|U_k\|^2\\ =\ &\|Z_k-p\|^2-\frac{\rho_k}{4}\mathsf{res}^2_{\lambda_k}(Z_k)+\frac{\rho_k\lambda_k^2}{2}\|U_k\|^2-2\rho_k\lambda_k\langle W_k+p^*,Y_k-p\rangle\\ &-\frac{5/2-2\rho_k(1+L\lambda_k)}{2\rho_k(1+L\lambda_k)}\|X_{k+1}-Z_k\|^2\\ &+\frac{5\rho_k\lambda_k^2}{2(1+L\lambda_k)}\|\mathsf{e}_k\|^2+2\rho_k\lambda_k\langle V(Y_k)-V(p),p-Y_k\rangle.\end{aligned}$$

Observe that

$$\begin{aligned}\|X_{k+1}-Z_k\|^2&=\|(X_{k+1}-X_k)-\alpha_k(X_k-X_{k-1})\|^2\\ &\geq(1-\alpha_k)\|X_{k+1}-X_k\|^2+(\alpha_k^2-\alpha_k)\|X_k-X_{k-1}\|^2,\end{aligned} \tag{31}$$

and Lemma 16 gives

$$\|Z_k-p\|^2=(1+\alpha_k)\|X_k-p\|^2-\alpha_k\|X_{k-1}-p\|^2+\alpha_k(1+\alpha_k)\|X_k-X_{k-1}\|^2. \tag{32}$$

By hypothesis, $\alpha_k,\rho_k,\lambda_k$ are defined such that $\frac{5/2-2\rho_k(1+L\lambda_k)}{2\rho_k(1+L\lambda_k)}>0$. Then, using both of these relations in the last estimate for $\|X_{k+1}-p\|^2$, we arrive at

$$\begin{aligned}\|X_{k+1}-p\|^2\leq(1+\alpha_k)\|X_k-p\|^2&-\alpha_k\|X_{k-1}-p\|^2+\alpha_k(1+\alpha_k)\|X_k-X_{k-1}\|^2\\ &-2\rho_k\lambda_k\left\langle W_{k+1}+p^*,Y_k-p\right\rangle\\ &-\frac{\rho_k}{4}\mathsf{res}^2_{\lambda_k}(Z_k)+\frac{5\rho_k\lambda_k^2}{2(1+L\lambda_k)}\|\mathsf{e}_k\|^2\\ &+\frac{\rho_k\lambda_k^2}{2}\|U_k\|^2+2\rho_k\lambda_k\langle V(Y_k)-V(p),p-Y_k\rangle\\ &-\left(\frac{5}{4\rho_k(1+L\lambda_k)}-1\right)\Big[(1-\alpha_k)\|X_{k+1}-X_k\|^2+(\alpha_k^2-\alpha_k)\|X_k-X_{k-1}\|^2\Big].\end{aligned}$$

Using the respective definitions of the stochastic increments ΔM_k, $\Delta N_k(p,p^*)$ in (20) and (21), we arrive at

$$\begin{aligned}\|X_{k+1}-p\|^2 \leq (1+\alpha_k)\|X_k-p\|^2 - \alpha_k\|X_{k-1}-p\|^2 - \frac{\rho_k}{4}\mathsf{res}^2_{\lambda_k}(Z_k)\\ + \Delta M_k + \Delta N_k(p,p^*) - 2\rho_k\lambda_k\langle V(Y_k)-V(p), Y_k-p\rangle\\ + \alpha_k\|X_k-X_{k-1}\|^2\left(2\alpha_k+\frac{5(1-\alpha_k)}{4\rho_k(1+L\lambda_k)}\right)\\ -(1-\alpha_k)\left(\frac{5}{4\rho_k(1+L\lambda_k)}-1\right)\|X_{k+1}-X_k\|^2.\end{aligned} \tag{33}$$

□

Recall that Y_k is $\hat{\mathcal{F}}_k$-measurable. By the law of iterated expectations, we therefore see

$$\mathbb{E}[\Delta N_k(p,p^*)|\mathcal{F}_k] = \mathbb{E}\left\{\mathbb{E}[\Delta N_k(p,p^*)|\hat{\mathcal{F}}_k]|\mathcal{F}_k\right\} = 2\rho_k\lambda_k\mathbb{E}[\langle p^*, p-Y_k\rangle|\mathcal{F}_k],$$

for all $(p,p^*) \in \operatorname{gr}(F)$. Observe that if we choose $(p,0) \in \operatorname{gr}(F)$, meaning that $p \in \mathsf{S}$, then $\Delta N_k(p,0) \equiv \Delta N_k(p)$ is a martingale difference sequence. Furthermore, for all $k \geq 1$,

$$\mathbb{E}[\Delta M_k|\mathcal{F}_k] \leq \frac{5\rho_k\lambda_k^2}{1+L\lambda_k}\mathbb{E}[\|W_k\|^2|\mathcal{F}_k] + \lambda_k^2\left(\frac{5\rho_k}{1+L\lambda_k}+\frac{\rho_k}{2}\right)\mathbb{E}[\|U_k\|^2|\mathcal{F}_k] \leq \frac{\mathsf{a}_k\sigma^2}{m_k}, \tag{34}$$

where $\mathsf{a}_k \triangleq \lambda_k^2\left(\frac{10\rho_k}{1+L\lambda_k}+\frac{\rho_k}{2}\right)$.

To prove the a.s. convergence of the stochastic process $(X_k)_{k\in\mathbb{N}}$, we rely on the following preparations. Motivated by the analysis of deterministic inertial schemes, we are interested in a regime under which α_k is non-decreasing.

For a fixed reference point $p \in \mathsf{H}$, define the anchor sequences $\phi_k(p) \triangleq \frac{1}{2}\|X_k-p\|^2$, and the energy sequence $\Delta_k \triangleq \frac{1}{2}\|X_k-X_{k-1}\|^2$. In terms of these sequences, we can rearrange the fundamental recursion from Lemma 4 to obtain

$$\begin{aligned}\phi_{k+1}(p)-\alpha_k\phi_k(p) - (1-\alpha_k)\left(\frac{5}{4\rho_k(1+L\lambda_k)}-1\right)\Delta_{k+1} \leq \phi_k(p)-\alpha_k\phi_{k-1}(p)\\ -(1-\alpha_k)\left(\frac{5}{4\rho_k(1+L\lambda_k)}-1\right)\Delta_k + \frac{1}{2}\Delta M_k + \frac{1}{2}\Delta N_k(p,p^*)\\ -\rho_k\lambda_k\langle V(Y_k)-V(p), Y_k-p\rangle + \Delta_k\left[2\alpha_k^2+(1-\alpha_k)\left(1-\frac{5(1-\alpha_k)}{4\rho_k(1+L\lambda_k)}\right)\right]\\ -\frac{\rho_k}{8}\mathsf{res}^2_{\lambda_k}(Z_k).\end{aligned}$$

For a given pair $(p,p^*) \in \operatorname{gr}(F)$, define

$$Q_k(p) \triangleq \phi_k(p) - \alpha_k\phi_{k-1}(p) + (1-\alpha_k)\left(\frac{5}{4\rho_k(1+L\lambda_k)} - 1\right)\Delta_k. \tag{35}$$

Then, in terms of the sequence

$$\beta_{k+1} \triangleq (1-\alpha_k)\left(\frac{5}{4\rho_k(1+L\lambda_k)} - 1\right) - (1-\alpha_{k+1})\left(\frac{5}{4\rho_{k+1}(1+L\lambda_{k+1})} - 1\right), \tag{36}$$

and using the monotonicity of V, guaranteeing that $\langle V(Y_k) - V(p), Y_k - p\rangle \geq 0$, we get

$$\begin{aligned} Q_{k+1}(p) \leq Q_k(p) - \frac{\rho_k}{8}\mathsf{res}^2_{\lambda_k}(Z_k) + \frac{1}{2}\Delta M_k + \frac{1}{2}\Delta N_k(p,p^*) + (\alpha_k - \alpha_{k+1})\phi_k(p) \\ + \left[2\alpha_k^2 + (1-\alpha_k)\left(1 - \frac{5(1-\alpha_k)}{4\rho_k(1+L\lambda_k)}\right)\right]\Delta_k - \beta_{k+1}\Delta_{k+1}. \end{aligned}$$

Defining

$$\theta_k \triangleq \frac{\rho_k}{8}\mathsf{res}^2_{\lambda_k}(Z_k) - \left[2\alpha_k^2 + (1-\alpha_k)\left(1 - \frac{5(1-\alpha_k)}{4\rho_k(1+L\lambda_k)}\right)\right]\Delta_k,$$

we arrive at

$$Q_{k+1}(p) \leq Q_k(p) - \theta_k + \frac{1}{2}\Delta M_k + \frac{1}{2}\Delta N_k(p,p^*) + (\alpha_k - \alpha_{k+1})\phi_k(p) - \beta_{k+1}\Delta_{k+1}. \tag{37}$$

Our aim is to use $Q_k(p)$ as a suitable energy function for RISFBF. For that to work, we need to identify a specific parameter sequence pair (ρ_k, α_k) so that $\beta_k \geq 0$ and $\theta_k \geq 0$, taking the following design criteria into account:

1. $\alpha_k \in (0, \bar{\alpha}] \subset (0,1)$ for all $k \geq 1$;
2. α_k is non-decreasing with

$$\sup_{k\geq 1}\alpha_k = \bar{\alpha}, \text{ and} \inf_{k\geq 1}\alpha_k > 0. \tag{38}$$

Incorporating these two restrictions on the inertia parameter α_k, we are left with the following constraints:

$$\beta_k \geq 0 \text{ and } 2\alpha_k^2 + (1-\alpha_k)\left(1 - \frac{5(1-\alpha_k)}{4\rho_k(1+L\lambda_k)}\right) \leq 0. \tag{39}$$

To identify a constellation of parameters (α_k, ρ_k) satisfying these two conditions, define

$$h_k(x,y) \triangleq (1-x)\left(\frac{5}{4y(1+L\lambda_k)} - 1\right). \tag{40}$$

Then,

$$
\begin{aligned}
0 &\geq 2\alpha_k^2 - (1-\alpha_k)\big(h_k(\alpha_k,\rho_k) + (1-\alpha_k) - 1\big) \\
&= 2\alpha_k^2 + \alpha_k(1-\alpha_k) - (1-\alpha_k)h_k(\alpha_k,\rho_k) \\
&= \alpha_k(1+\alpha_k) - (1-\alpha_k)h_k(\alpha_k,\rho_k),
\end{aligned}
$$

which gives

$$
h_k(\alpha_k,\rho_k) \geq \frac{\alpha_k(1+\alpha_k)}{1-\alpha_k}. \tag{41}
$$

Solving this condition for ρ_k reveals that $\frac{1}{\rho_k} \geq \frac{4(2\alpha_k^2-\alpha_k+1)(1+L\lambda_k)}{5(1-\alpha_k)^2}$. Using the design condition $\alpha_k \leq \bar{\alpha} < 1$, we need to choose the relaxation parameter ρ_k so that $\rho_k \leq \frac{5(1-\alpha_k)^2}{4(1+L\lambda_k)(2\alpha_k^2-\alpha_k+1)}$. This suggests to use the relaxation sequence $\rho_k = \rho_k(\alpha_k,\lambda_k) \triangleq \frac{5(1-\bar{\varepsilon})(1-\bar{\alpha})^2}{4(1+L\lambda_k)(2\alpha_k^2-\alpha_k+1)}$. It remains to verify that with this choice we can guarantee $\beta_k \geq 0$. This can be deduced as follows: Recalling (40), we get

$$
h_k(\alpha_k,\rho_k) = (1-\alpha_k)\left(\frac{5}{4\rho_k(1+L\lambda)} - 1\right) = \frac{(1-\alpha_k)(2\alpha_k^2-\alpha_k+1)}{(1-\bar{\varepsilon})(1-\bar{\alpha})^2} + \alpha_k - 1.
$$

In particular, we note that if $f(\alpha) \triangleq \frac{(1-\alpha)(2\alpha^2-\alpha+1)}{(1-\bar{\varepsilon})(1-\bar{\alpha})^2} + \alpha - 1$, then

$$
f'(\alpha) = \frac{(1-\alpha)(4\alpha-1)-(2\alpha^2-\alpha+1)}{(1-\bar{\varepsilon})(1-\bar{\alpha})^2} + 1 = \frac{-6\alpha^2+6\alpha-2+(1-\bar{\varepsilon})(1-\bar{\alpha})^2}{(1-\bar{\varepsilon})(1-\bar{\alpha})^2} = \frac{-6(\alpha-\frac{1}{2})^2-\frac{1}{2}+(1-\bar{\varepsilon})(1-\bar{\alpha})^2}{(1-\bar{\varepsilon})(1-\bar{\alpha})^2}
$$

We consider two cases:

Case 1: $\bar{\alpha} \leq 1/2$. In this case

$$
f'(\alpha) \leq \frac{-6(\bar{\alpha}-\frac{1}{2})^2-\frac{1}{2}+(1-\bar{\varepsilon})(1-\bar{\alpha})^2}{(1-\bar{\varepsilon})(1-\bar{\alpha})^2} \leq \frac{-5\bar{\alpha}^2+4\bar{\alpha}-1}{(1-\bar{\varepsilon})(1-\bar{\alpha})^2} < 0.
$$

Case 2: $1/2 < \bar{\alpha} < 1$. In this case

$$
f'(\alpha) \leq \frac{-6(1/2-1/2)^2 - 1/2 + (1-\bar{\varepsilon})(1-\bar{\alpha})^2}{(1-\bar{\varepsilon})(1-\bar{\alpha})^2} \leq \frac{-1/2 + (1-\bar{\varepsilon})(1-1/2)^2}{(1-\bar{\varepsilon})(1-\bar{\alpha})^2} < 0.
$$

Thus, $f(\alpha)$ is decreasing in $\alpha \in (0,\bar{\alpha}]$, where $0 < \bar{\alpha} < 1$.

Using these relations, we see that (37) reduces to

$$
Q_{k+1}(p) \leq Q_k(p) - \theta_k + \frac{1}{2}\Delta M_k + \frac{1}{2}\Delta N_k(p,p^*), \tag{42}
$$

where $\theta_k \geq 0$. This is the basis for our proof of Theorem 2.

Proof of Theorem 2 We start with (i). Consider (42), with the special choice $p^* = 0$, so that $p \in \mathsf{S}$. Taking conditional expectations on both sides of this inequality, we arrive at

$$
\mathbb{E}[Q_{k+1}(p)|\mathcal{F}_k] \leq Q_k(p) - \theta_k + \psi_k,
$$

where $\psi_k \triangleq \frac{\mathsf{a}_k\sigma^2}{2m_k}$. By design of the relaxation sequence ρ_k, we see that

$$\mathsf{a}_k = \lambda_k^2\left(\frac{10\rho_k}{1+L\lambda_k}+\frac{\rho_k}{2}\right) = \lambda_k^2\left(\frac{10}{1+L\lambda_k}+\frac{1}{2}\right)\frac{5(1-\bar{\varepsilon})}{4(2\alpha_k^2-\alpha_k+1)(1+L\lambda_k)}.$$

Since $\lim_{k\to\infty}\lambda_k = \lambda \in (0,1/4L)$, and $\lim_{k\to\infty}\alpha_k = \bar{\alpha} \in (0,1)$, we conclude that the sequence $(\mathsf{a}_k)_{k\in\mathbb{N}}$ is bounded. Consequently, thanks to Assumption 5, the sequence $(\psi_k)_{k\in\mathbb{N}}$ is in $\ell_+^1(\mathbb{N})$. We next claim that $Q_k(p) \geq 0$. To verify this, note that

$$\begin{aligned}
Q_k(p) &= \frac{1}{2}\|X_k-p\|^2 - \frac{\alpha_k}{2}\|X_{k-1}-p\|^2 + \frac{(1-\alpha_k)}{2}\left(\frac{5}{4\rho_k(1+L\lambda_k)}-1\right)\|X_k-X_{k-1}\|^2\\
&= \frac{1}{2}\|X_k-p\|^2 + \left(\frac{(1-\alpha_k)(2\alpha_k^2+1-\alpha_k)}{(1-\bar{\varepsilon})(1-\bar{\alpha})^2}-1+\alpha_k\right)\frac{1}{2}\|X_k-X_{k-1}\|^2\\
&\quad - \frac{\alpha_k}{2}\|X_{k-1}-p\|^2\\
&\geq \frac{1}{2}\|X_k-p\|^2 + \left(\frac{(1-\alpha_k)(\alpha_k^2+1-\alpha_k)}{(1-\bar{\varepsilon})(1-\bar{\alpha})^2}-1+\alpha_k\right)\frac{1}{2}\|X_k-X_{k-1}\|^2\\
&\quad - \frac{\alpha_k}{2}\|X_{k-1}-p\|^2\\
&\geq \frac{1}{2}\|X_k-p\|^2 + \left(\frac{(1-\alpha_k)(\alpha_k^2+1-\alpha_k)}{(1-\alpha_k)^2}-1+\alpha_k\right)\frac{1}{2}\|X_k-X_{k-1}\|^2\\
&\quad - \frac{\alpha_k}{2}\|X_{k-1}-p\|^2\\
&= (\alpha_k+(1-\alpha_k))\|X_k-p\|^2 + \left(\alpha_k+\frac{\alpha_k^2}{1-\alpha_k}\right)\|X_k-X_{k-1}\|^2\\
&\quad - \alpha_k\|X_{k-1}-p\|^2\\
&\geq \frac{\alpha_k}{2}\left(\|X_k-p\|^2+\|X_k-X_{k-1}\|^2\right)\\
&\quad - \frac{\alpha_k}{2}\|X_{k-1}-p\|^2 + \alpha_k\|X_k-p\|\cdot\|X_k-X_{k-1}\|\\
&\geq \frac{\alpha_k}{2}\left(\|X_k-p\|+\|X_k-X_{k-1}\|\right)^2 - \frac{\alpha_k}{2}\|X_{k-1}-p\|^2 \geq 0.
\end{aligned}$$

where the first and second inequality uses $\bar{\varepsilon} < 1$ and $\alpha_k \leq \bar{\alpha} \in (0,1)$, the third inequality makes use of the Young inequality: $\frac{1-a}{2a}\|X_k-p\|^2 + \frac{a}{2(1-a)}\|X_k-X_{k-1}\|^2 \geq \|X_k-p\|\cdot\|X_k-X_{k-1}\|$. Finally, the fourth inequality uses the triangle inequality $\|X_{k-1}-p\| \leq \|X_k-X_{k-1}\|+\|X_k-p\|$. Lemma 17 readily yields the existence of an a.s. finite limiting random variable $Q_\infty(p)$ such that $Q_k(p)\to Q_\infty(p)$, $\mathbb{P}$-a.s., and $(\theta_k)_{k\in\mathbb{N}} \in \ell_+^1(\mathbb{F})$. Since $\lambda_k \to \lambda$, we get $\lim_{k\to\infty}\rho_k = \frac{5(1-\bar{\varepsilon})(1-\bar{\alpha})^2}{4(1+L\lambda)(2\bar{\alpha}^2+1-\bar{\alpha})}$. Hence,

$$\lim_{k\to\infty}\left(2\alpha_k^2-(1-\alpha_k)\left(1-\tfrac{5(1-\alpha_k)}{4\rho_k(1+L\lambda_k)}\right)\right)\left\|X_k(\omega)-X_{k-1}(\omega)\right\|^2=0,\ \text{ and}$$
$$\lim_{k\to\infty}\frac{\rho_k}{4}\mathsf{res}^2_{\lambda_k}(Z_k(\omega))=0.$$

$\mathbb{P}$-a.s. We conclude that $\lim_{k\to\infty}\mathsf{res}^2_{\lambda_k}(Z_k)=0$, $\mathbb{P}$-a.s..

To prove (ii) observe that, since $\bar{\varepsilon}\in(0,1)$ and $\lim_{k\to\infty}\alpha_k=\bar{\alpha}$, it follows

$$\left[2\alpha_k^2+(1-\alpha_k)\left(1-\frac{5(1-\alpha_k)}{4\rho_k(1+L\lambda_k)}\right)\right]\leq\frac{-\bar{\varepsilon}}{1-\bar{\varepsilon}}(2\bar{\alpha}^2+1-\bar{\alpha})<0.$$

Consequently, $\lim_{k\to\infty}\left\|X_k-X_{k-1}\right\|^2=0$, $\mathbb{P}$-a.s., and $\left(\phi_k(p)-\alpha_k\phi_{k-1}(p)\right)_{k\in\mathbb{N}}$ is almost surely bounded. Hence, for each $\omega\in\Omega$, there exists a bounded random variable $C_1(\omega)\in[0,\infty)$ such that

$$\phi_k(p,\omega)\leq C_1(\omega)+\alpha_k\phi_{k-1}(p,\omega)\leq C_1(\omega)+\bar{\alpha}\phi_{k-1}(p,\omega)\qquad\forall k\geq 1.$$

Iterating this relation, using the fact that $\bar{\alpha}\in[0,1)$, we easily derive

$$\phi_k(p,\omega)\leq\frac{C_1(\omega)}{1-\bar{\alpha}}+\bar{\alpha}^k\phi_1(p,\omega).$$

Hence, $(\phi_k(p))_{k\in\mathbb{N}}$ is a.s. bounded, which implies that $(X_k)_{k\in\mathbb{N}}$ is bounded $\mathbb{P}$-a.s. We next claim that $(\left\|X_k-p\right\|)_{k\in\mathbb{N}}$ converges to a $[0,\infty)$-valued random variable $\mathbb{P}$-a.s. Indeed, take $\omega\in\Omega$ such that $\phi_k(p,\omega)\equiv\phi_k(\omega)$ is bounded. Suppose there exists $\mathtt{t}_1(\omega)\in[0,\infty),\mathtt{t}_2(\omega)\in[0,\infty)$, and subsequences $(\phi_{k_j}(\omega))_{j\in\mathbb{N}}$ and $(\phi_{l_j}(\omega))_{j\in\mathbb{N}}$ such that $\phi_{k_j}(\omega)\to\mathtt{t}_1(\omega)$ and $\phi_{l_j}(\omega)\to\mathtt{t}_2(\omega)>\mathtt{t}_1(\omega)$. Then, $\lim_{j\to\infty}Q_{k_j}(p)(\omega)=Q_\infty(p)(\omega)=(1-\bar{\alpha})\mathtt{t}_1(\omega)<(1-\bar{\alpha})\mathtt{t}_2(\omega)=\lim_{j\to\infty}Q_{l_j}(p)(\omega)=Q_\infty(p)(\omega)$, a contradiction. It follows that $\mathtt{t}_1(\omega)=\mathtt{t}_2(\omega)$ and, in turn, $\phi_k(\omega)\to\mathtt{t}(\omega)$. Thus, for each $p\in\mathsf{S}$, $\phi_k(p)\to\mathtt{t}$ $\mathbb{P}$-a.s.

Since we assume that H is separable, [67, Prop 2.3(iii)] guarantees that there exists a set $\Omega_0\in\mathcal{F}$ with $\mathbb{P}(\Omega_0)=1$, and, for every $\omega\in\Omega_0$ and every $p\in\mathsf{S}$, the sequence $(\left\|X_k(\omega)-p\right\|)_{k\in\mathbb{N}}$ converges.

We next show that all weak limit points of $(X_k)_{k\in\mathbb{N}}$ are contained in S. Let $\omega\in\Omega$ such that $(X_k(\omega))_{k\in\mathbb{N}}$ is bounded. Thanks to [3, Lemma 2.45], we can find a weakly convergent subsequence $(X_{k_j}(\omega))_{j\in\mathbb{N}}$ with limit $\chi(\omega)$, i.e. for all $u\in\mathsf{H}$ we have $\lim_{j\to\infty}\left\langle X_{k_j}(\omega),u\right\rangle=\langle\chi(\omega),u\rangle$. This implies

$$\lim_{j\to\infty}\left\langle Z_{k_j}(\omega),u\right\rangle=\lim_{j\to\infty}\left\langle X_{k_j}(\omega),u\right\rangle+\lim_{j\to\infty}\alpha_{k_j}\left\langle X_{k_j}(\omega)-X_{k_{j-1}}(\omega),u\right\rangle=\langle\chi(\omega),u\rangle,$$

showing that $Z_{k_j}(\omega)\rightharpoonup\chi(\omega)$. Along this weakly converging subsequence, define

$$r_{k_j}(\omega)\triangleq Z_{k_j}(\omega)-J_{\lambda_{k_j}T}(Z_{k_j}(\omega)-\lambda_{k_j}V(Z_{k_j}(\omega))).$$

Clearly, $\mathsf{res}_{\lambda_{k_j}}(Z_{k_j}(\omega))=\left\|r_{k_j}(\omega)\right\|$, so that $\lim_{j\to\infty}r_{k_j}(\omega)=0$. By definition

$$\frac{1}{\lambda_{k_j}} r_{k_j}(\omega) - V(Z_{k_j}(\omega)) + V\Big(Z_{k_j}(\omega) - r_{k_j}(\omega)\Big) \in F(Z_{k_j}(\omega) - r_{k_j}(\omega)).$$

Since V and $F = T + V$ are maximally monotone, their graphs are sequentially closed in the weak-strong topology $\mathsf{H}^{\text{weak}} \times \mathsf{H}^{\text{strong}}$ [3, Prop. 20.33(ii)]. Therefore, by the strong convergence of the sequence $(r_{k_j}(\omega))_{j\in\mathbb{N}}$, we deduce weak convergence of the sequence $(Z_{k_j}(\omega) - r_{k_j}(\omega), Z_{k_j}(\omega))_{j\in\mathbb{N}} \rightharpoonup (\chi(\omega), \chi(\omega))$. Therefore $\frac{1}{\lambda} r_{k_j}(\omega) - V(Z_{k_j}(\omega)) + V\Big(Z_{k_j}(\omega) - r_{k_j}(\omega)\Big) \to 0$. Hence, $0 \in (T + V)(\chi(\omega))$, showing that $\chi(\omega) \in \mathsf{S}$. Invoking [67, Prop 2.3(iv)], we conclude that $(X_k)_{k\in\mathbb{N}}$ converges weakly $\mathbb{P}$-a.s to an S -valued random variable.

We now establish (iii). Let $q_k \triangleq \mathbb{E}[Q_k(p)]$, so that (42) yields the recursion

$$q_k \leq q_{k-1} - \mathbb{E}[\theta_k] + \psi_k.$$

By Assumption 5, and the definition of all sequences involved, we see that $\sum_{k=1}^{\infty} \psi_k < \infty$. Hence, a telescopian argument gives

$$q_k - q_0 = \sum_{i=1}^{k}(q_i - q_{i-1}) \leq -\sum_{i=1}^{k} \mathbb{E}[\theta_i] + \sum_{i=1}^{k} \psi_i \leq -\sum_{i=1}^{k} \mathbb{E}[\theta_i] + \sum_{i=1}^{\infty} \psi_i.$$

Hence, for all $k \geq 1$, rearranging the above reveals

$$\sum_{i=1}^{k} \mathbb{E}[\theta_i] \leq q_0 + \sum_{i=1}^{\infty} \psi_i < \infty.$$

Letting $k \to \infty$, we conclude $\big(\mathbb{E}[\theta_k]\big)_{k\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$. Classically, this implies $\theta_k \to 0$ $\mathbb{P}$ -a.s. By a simple majorization argument, we deduce that $\mathbb{P}$-a.s.

$$\begin{aligned}\infty &> \sum_{k=1}^{\infty} \left\{ \frac{\rho_k}{8} \mathsf{res}^2_{\lambda_k}(Z_k) - \left[2\alpha_k^2 + (1-\alpha_k)\left(1 - \frac{5(1-\alpha_k)}{4\rho_k(1+L\lambda_k)}\right)\right]\Delta_k \right\} \\ &\geq \sum_{k=1}^{\infty} \left[(1-\alpha_k)\left(\frac{5(1-\alpha_k)}{4\rho_k(1+L\lambda_k)} - 1\right) - 2\alpha_k^2\right]\Delta_k.\end{aligned}$$

□

Remark 2 The above result gives some indication of the balance between the inertial effect and the relaxation effect. Our analysis revealed that the maximal value of the relaxation parameter is $\rho \leq \frac{5(1-\bar{\varepsilon})(1-\alpha)^2}{4(1+L\lambda)(2\alpha^2-\alpha+1)}$. This is closely aligned with the maximal relaxation value exhibited in Remark 2.13 of [2]. Specifically, the function $\rho_m(\alpha, \varepsilon) = \frac{5(1-\varepsilon)(1-\alpha)^2}{4(1+L\lambda)(2\alpha^2-\alpha+1)}$. This function is decreasing in α. For this choice of parameters, one observes that for $\alpha \to 0$ we get $\rho \to \frac{5(1-\varepsilon)}{4(1+L\lambda)}$ and for $\alpha \to 1$ it is observed $\rho \to 0$.

As an immediate corollary of Theorem 2, we obtain a convergence result when all parameter sequences are constant.

Corollary 5 (Asymptotic convergence under constant inertia and relaxation) *Let the same Assumptions as in Theorem 2 hold. Consider Algorithm RISFBF with the constant parameter sequences* $\alpha_k \equiv \alpha \in (0,1), \lambda_k \equiv \lambda \in (0, \frac{1}{4L})$ *and* $\rho_k = \rho < \frac{5(1-\alpha)^2}{4(1+L\lambda)(2\alpha^2+1-\alpha)}$. *Then* $(X_k)_{k\in\mathbb{N}}$ *converges weakly* $\mathbb{P}$*-a.s. to a limiting random variable with values in* S.

In fact, the a.s. convergence with a larger λ_k is allowed as shown in the following corollary.

Corollary 6 (Asymptotic convergence under larger steplength) *Let the same Assumptions as in Theorem 2 hold. Consider Algorithm RISFBF with the constant parameter sequences* $\alpha_k \equiv \alpha \in (0,1), \lambda_k \equiv \lambda \in (0, \frac{1-\nu}{2L})$ *and* $\rho_k = \rho < \frac{(3-\nu)(1-\alpha)^2}{2(1+L\lambda)(2\alpha^2+1-\alpha)}$, *where* $0 < \nu < 1$. *Then* $(X_k)_{k\in\mathbb{N}}$ *converges weakly* $\mathbb{P}$*-a.s. to a limiting random variable with values in* S.

Proof First we make a slight modification to (27) that the following relation holds for $0 < \nu < 1$

$$
\begin{aligned}
&(1-\rho_k)\|Z_k - p\|^2 + \rho_k\|R_k - p\|^2 - \frac{1-\rho_k}{\rho_k}\|X_{k+1} - Z_k\|^2 \\
\leq\ & \|Z_k - p\|^2 - \frac{1-\rho_k}{\rho_k}\|X_{k+1} - Z_k\|^2 - \rho_k((1-\nu) - 2L^2\lambda_k^2)\|Z_k - Y_k\|^2 \\
&+ 2\lambda^2\rho_k\|\mathsf{e}_k\|^2 - 2\rho_k\lambda_k\langle W_k + p^*, Y_k - p\rangle - \rho_k\nu\|Y_k - Z_k\|^2 + 2\rho_k\lambda_k\langle V(Y_k) - V(p), p - Y_k\rangle.
\end{aligned}
$$

Then similarly with (29), we multiply both sides of (28) by $\frac{\rho_k((1-\nu)-2L\lambda_k)}{1+L\lambda_k}$, which is positive since $\lambda_k \in (0, \frac{1-\nu}{2L})$. The convergence follows in a similar fashion to Theorem 2. □

Another corollary of Theorem 2 is a strong convergence result, assuming that F is demiregular (cf. Definition 1).

Corollary 7 (Strong Convergence under demiregularity) *Let the same Assumptions as in Theorem 2 hold. If* $F = T + V$ *is demiregular, then* $(X_k)_{k\in\mathbb{N}}$ *converges strongly* $\mathbb{P}$*-a.s. to a* S*-valued random variable.*

Proof Set $y_{k_j}(\omega) \triangleq Z_{k_j}(\omega) - r_{k_j}(\omega)$, and $u_{k_j}(\omega) \triangleq \frac{1}{\lambda} r_{k_j}(\omega) - V(Z_{k_j}(\omega)) + V(Z_{k_j}(\omega) - r_{k_j}(\omega))$. We know from the proof of Theorem 2 that $y_{k_j}(\omega) \rightharpoonup \chi(\omega)$ and $u_{k_j}(\omega) \to 0$. If $F = T + V$ is demiregular then $y_{k_j}(\omega) \to \chi(\omega)$. Since we know $r_{k_j}(\omega) \to 0$, we conclude $Z_{k_j}(\omega) \to \chi(\omega)$. Since Z_k and X_k have the same limit points, it follows $X_k \to \chi$. □

4.2 Linear convergence

In this section, we derive a linear convergence rate and prove strong convergence of the last iterate in the case where the single-valued operator V is *strongly monotone*. Various linear convergence results in the context of stochastic approximation algorithms for solving fixed-point problems are reported in [68] in the context of the random sweeping processes. In a general structured monotone inclusion setting [69] derive rate statements for cocoercive mean operators in the context of forward-backward splitting methods. More recently, Cui and Shanbhag [27] provide linear and sublinear rates of convergence for a variance-reduced inexact proximal-point scheme for both strongly monotone and monotone inclusion problems. However, to the best of our knowledge, our results are the first published for a stochastic operator splitting algorithm, featuring *relaxation and inertial* effects. Notably, this result does not require imposing Assumption 4(i) (i.e. the noise process be conditionally unbiased.) Instead our derivations hold true under a weaker notion of an asymptotically unbiased SO.

Assumption 6 (Asymptotically unbiased SO) There exists a constant $\mathsf{s} > 0$ such that

$$\mathbb{E}[\|U_k\|^2|\mathcal{F}_k] \leq \frac{\mathsf{s}^2}{m_k} \text{ and } \mathbb{E}[\|W_k\|^2|\mathcal{F}_k] \leq \frac{\mathsf{s}^2}{m_k}, \qquad \mathbb{P}-\text{a.s.} \tag{43}$$

for all $k \geq 1$.

This definition is rather mild and is imposed in many simulation-based optimization schemes in finite dimensions. Amongst the more important ones is the simultaneous perturbation stochastic approximation (SPSA) method pioneered by Spall [70, 71]. In this scheme, it is required that the gradient estimator satisfies an asymptotic unbiasedness requirement; in particular, the bias in the gradient estimator needs to diminish at a suitable rate to ensure asymptotic convergence. In fact, this setting has been investigated in detail in the context of stochastic Nash games [72]. Further examples for stochastic approximation schemes in a Hilbert-space setting obeying Assumption 6 are [73, 74] and [35]. We now discuss an example that further clarifies the requirements on the estimator.

Example 3 Let $\{\hat{V}_k(x,\xi)\}_{k\in\mathbb{N}}$ be a collection of independent random H-valued vector fields of the form $\hat{V}_k(x,\xi) = V(x) + \varepsilon_k(x,\xi)$ such that

$$\mathbb{E}_\xi[\varepsilon_k(x,\xi)|x] = \frac{B_k}{\sqrt{m_k}} \text{ and } \mathbb{E}_\xi[\|\varepsilon_k(x,\xi)\|^2|x] \leq \hat{\sigma}^2 \qquad \mathbb{P}-\text{a.s.,}$$

where $\hat{\sigma} > 0$ and $\tilde{b} > 0$ such that $(B_k)_{k\in\mathbb{N}}$ is an H-valued sequence satisfying $\|B_k\|^2 \leq \hat{b}^2$ in an a.s. sense. These statistics can be obtained as

$$\begin{aligned}\mathbb{E}[\|U_k\|^2|\mathcal{F}_k] &= \mathbb{E}\left[\left\|\frac{1}{m_k}\sum_{t=1}^{m_k}\varepsilon_t(Z_k)\right\|^2|\mathcal{F}_k\right]\\ &= \frac{1}{m_k^2}\sum_{t=1}^{m_k}\mathbb{E}[\|\varepsilon_t(Z_k)\|^2|\mathcal{F}_k] + \frac{2}{m_k^2}\sum_{t=1}^{m_k}\sum_{l>t}\mathbb{E}[\langle\varepsilon_t(Z_k),\varepsilon_l(Z_k)\rangle|\mathcal{F}_k]\\ &\leq \frac{\hat{\sigma}^2}{m_k} + \frac{(m_k-1)}{m_k}\frac{\|B_k\|^2}{m_k} \leq \frac{\hat{\sigma}^2+\hat{b}^2}{m_k} \qquad \mathbb{P}-\text{a.s.}\end{aligned}$$

Setting $\mathsf{s}^2 \triangleq \hat{\sigma}^2 + \hat{b}^2$, we see that condition (43) holds. A similar estimate holds for the random noise $\|W_k\|^2$.

Assumption 7 $V : \mathsf{H} \to \mathsf{H}$ is μ-strongly monotone ($\mu > 0$), i.e.

$$\langle V(x) - V(y), x - y\rangle \geq \mu\|x-y\|^2 \qquad \forall x, y \in \operatorname{dom} V = \mathsf{H}. \tag{44}$$

Combined with Assumption 1, strong monotonicity implies that $\mathsf{S} = \{\bar{x}\}$ for some $\bar{x} \in \mathsf{H}$.

Remark 3 In the context of a structured operator $F = T + V$, the assumption that the single-valued part V is strongly monotone can be done without loss of generality. Indeed, if instead T is assumed to be μ-strongly monotone, then $(V + \mu\,\mathrm{Id}\,) + (T - \mu\,\mathrm{Id}\,)$ is maximally monotone and Lipschitz continuous while $\tilde{V} \triangleq V + \mu\,\mathrm{Id}$ may be seen to be μ-strongly monotone operator.

Our first result establishes a "perturbed linear convergence" rate on the anchor sequence $\left(\|X_k - \bar{x}\|^2\right)_{k\in\mathbb{N}}$, similar to the one derived in [68, Corollary 3.2] in the context of randomized fixed point iterations.

Theorem 8 (Perturbed linear convergence) *Consider RISFBF with $X_0 = X_1$. Suppose Assumptions 1-3, Assumption 6 and Assumption 7 hold. Let $\mathsf{S} = \{\bar{x}\}$ denotes the unique solution of (MI). Suppose $\lambda_k \equiv \lambda \leq \min\left\{\frac{a}{2\mu}, b\mu, \frac{1-a}{2\tilde{L}}\right\}$, where $0 < a, b < 1$, $\tilde{L}^2 \triangleq L^2 + \frac{1}{2}$, $\eta_k \equiv \eta \triangleq (1-b)\lambda\mu$. Define $\Delta M_k \triangleq 2\rho_k\|W_k\|^2 + \frac{(3-a)\rho_k\lambda_k^2}{1+\tilde{L}\lambda_k}\|\mathsf{e}_k\|^2$. Let $(\alpha_k)_{k\in\mathbb{N}}$ be a non-decreasing sequence such that $0 < \alpha_k \leq \bar{\alpha} < 1$, and define $\rho_k \triangleq \frac{(3-a)(1-\alpha_k)^2}{2(2\alpha_k^2-0.5\alpha_k+1)(1+\tilde{L}\lambda)}$ for every $k \in \mathbb{N}$. Set*

$$\begin{aligned}&H_k \triangleq \|X_k - \bar{x}\|^2 + (1-\alpha_k)\left(\frac{3-a}{2\rho_k(1+\tilde{L}\lambda)} - 1\right)\|X_k - X_{k-1}\|^2 - \alpha_k\|X_{k-1} - \bar{x}\|^2,\\ &c_k \triangleq \mathbb{E}[\Delta M_k|\mathcal{F}_k], \text{ and } \bar{c}_k \triangleq \sum_{i=1}^{k} q^{k-i}\mathbb{E}[c_i|\mathcal{F}_1],\end{aligned} \tag{45}$$

where $q = 1 - \rho\eta \in (0,1)$, $\rho = \frac{16(3-a)(1-\bar{\alpha})^2}{31(1+\tilde{L}\lambda)}$. Then the following hold:

(i) $(\bar{c}_k)_{k\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$.

(ii) *For all* $k \geq 1$

$$\mathbb{E}[H_{k+1}|\mathcal{F}_1] \leq q^k H_1 + \bar{c}_k. \tag{46}$$

In particular, this implies a perturbed linear rate of the sequence $(\|X_k - \bar{x}\|^2)_{k\in\mathbb{N}}$ as

$$\mathbb{E}[\|X_{k+1} - \bar{x}\|^2|\mathcal{F}_0] \leq q^k\left(\frac{2(1-\alpha_1)}{1-\bar{\alpha}}\|X_1 - \bar{x}\|^2\right) + \tfrac{2}{1-\bar{\alpha}}\bar{c}_k. \tag{47}$$

(iii) $\sum_{k=1}^{\infty}(1-\alpha_k)\left(\frac{3-a}{2\rho_k(1+\tilde{L}\lambda)} - 1\right)\|X_k - X_{k-1}\|^2 < \infty$ $\mathbb{P}$-a.s..

Proof Our point of departure for the analysis under the stronger Assumption 7 is eq. (23), which becomes

$$\langle Z_k - R_k, Y_k - p\rangle \geq \lambda_k\langle W_k + p^*, Y_k - p\rangle + \lambda_k\mu\|Y_k - p\|^2 \quad \forall (p, p^*) \in \mathrm{gr}\,(F).$$

Repeating the analysis of the previous section with reference point $p = \bar{x}$ and $p^* = 0$, the unique solution of (MI), yields the bound

$$\begin{aligned}\|R_k - \bar{x}\|^2 \leq \|Z_k - \bar{x}\|^2 &- (1 - 2L^2\lambda_k^2)\|Y_k - Z_k\|^2 + 2\lambda_k^2\|\mathsf{e}_k\|^2 \\ &+ 2\lambda_k\langle W_k, \bar{x} - Y_k\rangle - 2\lambda_k\mu\|\bar{x} - Y_k\|^2.\end{aligned}$$

The triangle inequality $\|Z_k - \bar{x}\|^2 \leq 2\|Y_k - Z_k\|^2 + 2\|Y_k - \bar{x}\|^2$ gives

$$\begin{aligned}\|R_k - \bar{x}\|^2 \leq \|Z_k - \bar{x}\|^2 - (1 - 2L^2\lambda_k^2)\|Y_k - Z_k\|^2 + 2\lambda_k^2\|\mathsf{e}_k\|^2 + 2\lambda_k\langle W_k, \bar{x} - Y_k\rangle \\ + 2\lambda_k\mu\|Y_k - Z_k\|^2 - \lambda_k\mu\|Z_k - \bar{x}\|^2.\end{aligned}$$

By (9), we have for all $c > 0$

$$\begin{aligned}\langle W_k, \bar{x} - Y_k\rangle &\leq \frac{1}{2c}\|W_k\|^2 + \frac{c}{2}\|Y_k - \bar{x}\|^2 \\ &\leq \frac{1}{2c}\|W_k\|^2 + c\left(\|Z_k - \bar{x}\|^2 + \|Z_k - Y_k\|^2\right).\end{aligned} \tag{48}$$

Observe that this estimate is crucial in weakening the requirement of conditional unbiasedness. Choose $c = \frac{\lambda_k}{2}$ to get

$$\begin{aligned}\|R_k - \bar{x}\|^2 &\leq \|Z_k - \bar{x}\|^2 - (1 - 2L^2\lambda_k^2)\|Y_k - Z_k\|^2 + 2\lambda_k^2\|\mathsf{e}_k\|^2 + 2\|W_k\|^2 + \lambda_k^2\|\bar{x} - Z_k\|^2 \\ &\quad + 2\lambda_k\mu\|Y_k - Z_k\|^2 - \lambda_k\mu\|Z_k - \bar{x}\|^2 + \lambda_k^2\|Z_k - Y_k\|^2 \\ &= (1 + \lambda_k^2 - \lambda_k\mu)\|Z_k - \bar{x}\|^2 + 2\lambda_k^2\|\mathsf{e}_k\|^2 + 2\|W_k\|^2 \\ &\quad - (1 - 2L^2\lambda_k^2 - 2\lambda_k\mu - \lambda_k^2)\|Y_k - Z_k\|^2.\end{aligned}$$

Assume that $\lambda_k\mu \leq \frac{a}{2} < 1$. Then,

$$1-2L^2\lambda_k^2-2\lambda_k\mu-\lambda_k^2 \geq (1-a)-2L^2\lambda_k^2-\lambda_k^2=(1-a)-2\tilde{L}^2\lambda_k^2,$$

where $\tilde{L}^2 \triangleq L^2+1/2$. Moreover, choosing $\lambda_k \leq b\mu$, we see

$$1+\lambda_k^2-\lambda_k\mu \leq 1-(1-b)\lambda_k\mu.$$

Using these bounds, we readily deduce for $0<\lambda_k \leq \min\{\frac{a}{2\mu}, b\mu\}$, that

$$\begin{aligned}\|R_k-\bar{x}\|^2 \leq &\big(1-(1-b)\lambda_k\mu\big)\|Z_k-\bar{x}\|^2-\big((1-a)-2\tilde{L}^2\lambda_k^2\big)\|Y_k-Z_k\|^2\\ &+2\lambda_k^2\|\mathsf{e}_k\|^2+2\|W_k\|^2.\end{aligned} \tag{49}$$

Proceeding as in the derivation of eq. (30), one sees first that

$$\frac{1}{2\rho_k^2}\|X_{k+1}-Z_k\|^2 \leq (1+\tilde{L}\lambda_k)^2\|Y_k-Z_k\|^2+\lambda_k^2\|\mathsf{e}_k\|^2,$$

and therefore,

$$\begin{aligned}-\rho_k((1-a)-2\tilde{L}^2\lambda_k^2)\|Y_k-Z_k\|^2 \leq &-\frac{(1-a)-2\tilde{L}\lambda_k}{2\rho_k(1+\tilde{L}\lambda_k)}\|X_{k+1}-Z_k\|^2\\ &+\frac{\rho_k\lambda_k^2((1-a)-2\tilde{L}\lambda_k)}{1+\tilde{L}\lambda_k}\|\mathsf{e}_k\|^2.\end{aligned} \tag{50}$$

Define $\eta_k=(1-b)\lambda_k\mu$. Using the equality (25),

$$\begin{aligned}\|X_{k+1}-\bar{x}\|^2 &= (1-\rho_k)\|Z_k-\bar{x}\|^2+\rho_k\|R_k-\bar{x}\|^2-\tfrac{1-\rho_k}{\rho_k}\|X_{k+1}-Z_k\|^2\\
&\overset{(49)}{\leq} (1-\rho_k\eta_k)\|Z_k-\bar{x}\|^2-\tfrac{1-\rho_k}{\rho_k}\|X_{k+1}-Z_k\|^2-\rho_k((1-a)-2\tilde{L}^2\lambda_k^2)\|Z_k-Y_k\|^2\\
&\quad+2\lambda_k^2\rho_k\|\mathsf{e}_k\|^2+2\rho_k\|W_k\|^2\\
&\overset{(50)}{\leq} (1-\rho_k\eta_k)\|Z_k-\bar{x}\|^2-\tfrac{(3-a)-2\rho_k(1+\tilde{L}\lambda_k)}{2\rho_k(1+\tilde{L}\lambda_k)}\|X_{k+1}-Z_k\|^2+2\rho_k\|W_k\|^2+\tfrac{(3-a)\rho_k\lambda_k^2}{1+\tilde{L}\lambda_k}\|\mathsf{e}_k\|^2\\
&\overset{(32),(31)}{\leq} (1-\rho_k\eta_k)[(1+\alpha_k)\|X_k-\bar{x}\|^2-\alpha_k\|X_{k-1}-\bar{x}\|^2+\alpha_k(1+\alpha_k)\|X_k-X_{k-1}\|^2]\\
&\quad-\tfrac{(3-a)-2\rho_k(1+\tilde{L}\lambda_k)}{2\rho_k(1+\tilde{L}\lambda_k)}[(1-\alpha_k)\|X_{k+1}-X_k\|^2+(\alpha_k^2-\alpha_k)\|X_k-X_{k-1}\|^2]\\
&\quad+2\rho_k\|W_k\|^2+\tfrac{(3-a)\rho_k\lambda_k^2}{(1+\tilde{L}\lambda_k)}\|\mathsf{e}_k\|^2\\
&\leq (1+\alpha_k)(1-\rho_k\eta_k)\|X_k-\bar{x}\|^2-\alpha_k(1-\rho_k\eta_k)\|X_{k-1}-\bar{x}\|^2+\Delta M_k\\
&\quad+\alpha_k\|X_k-X_{k-1}\|^2\Big[(1+\alpha_k)(1-\rho_k\eta_k)+(\alpha_k-1)+\tfrac{(3-a)(1-\alpha_k)}{2\rho_k(1+\tilde{L}\lambda_k)}\Big]\\
&\quad-(1-\alpha_k)\Big(\tfrac{3-a}{2\rho_k(1+\tilde{L}\lambda_k)}-1\Big)\|X_{k+1}-X_k\|^2,\end{aligned}$$

with stochastic error term $\Delta M_k \triangleq 2\rho_k\|W_k\|^2 + \frac{(3-a)\rho_k\lambda_k^2}{1+\tilde{L}\lambda_k}\|\mathsf{e}_k\|^2$. From here, it follows that

$$
\begin{aligned}
&\|X_{k+1}-\bar{x}\|^2 + (1-\alpha_k)\Big(\tfrac{3-a}{2\rho_k(1+\tilde{L}\lambda_k)} - 1\Big)\|X_{k+1}-X_k\|^2 - \alpha_k\|X_k-\bar{x}\|^2 \\
&\quad \leq (1-\rho_k\eta_k)\Big[\|X_k-\bar{x}\|^2 + (1-\alpha_k)\Big(\tfrac{3-a}{2\rho_k(1+\tilde{L}\lambda_k)} - 1\Big)\|X_k-X_{k-1}\|^2 - \alpha_k\|X_{k-1}-\bar{x}\|^2\Big] \\
&\qquad - \Big[(1-\rho_k\eta_k)(1-\alpha_k)\Big(\tfrac{3-a}{2\rho_k(1+\tilde{L}\lambda_k)} - 1\Big) \\
&\qquad -\alpha_k\Big((1+\alpha_k)(1-\rho_k\eta_k) + (\alpha_k-1) + \tfrac{(3-a)(1-\alpha_k)}{2\rho_k(1+\tilde{L}\lambda_k)}\Big)\Big]\|X_k-X_{k-1}\|^2 \\
&\qquad - \alpha_k\rho_k\eta_k\|X_k-\bar{x}\|^2 + \Delta M_k \\
&\quad = (1-\rho_k\eta_k)\Big[\|X_k-\bar{x}\|^2 + (1-\alpha_k)\Big(\tfrac{3-a}{2\rho_k(1+\tilde{L}\lambda_k)} - 1\Big)\|X_k-X_{k-1}\|^2 - \alpha_k\|X_{k-1}-\bar{x}\|^2\Big] \\
&\qquad - \underbrace{\Big[(1-\alpha_k-\rho_k\eta_k)\Big(\tfrac{(3-a)(1-\alpha_k)}{2\rho_k(1+\tilde{L}\lambda_k)} - 1\Big) - \alpha_k^2(2-\rho_k\eta_k)\Big]}_{\triangleq \tilde{I}}\|X_k-X_{k-1}\|^2 - \alpha_k\rho_k\eta_k\|X_k-\bar{x}\|^2 + \Delta M_k.
\end{aligned}
\tag{51}
$$

Since $\lambda_k = \lambda$, and $\rho_k = \frac{(3-a)(1-\alpha_k)^2}{2(2\alpha_k^2-0.5\alpha_k+1)(1+\tilde{L}\lambda)}$, we claim that $\rho_k \leq \frac{1-\alpha_k}{(1+4\alpha_k)\eta}$ for $\eta \equiv (1-b)\lambda\mu$. Indeed,[2]

$$
\frac{\frac{1-\alpha_k}{(1+4\alpha_k)\eta}}{\rho_k} = \frac{2(2\alpha_k^2-0.5\alpha_k+1)(1+\tilde{L}\lambda)}{(3-a)(1-\alpha_k)(1+4\alpha_k)\eta} \geq \frac{2(2\alpha_k^2-0.5\alpha_k+1)(1+\tilde{L}\lambda)}{(3-a)(1-\alpha_k)(1+4\alpha_k)\frac{a}{2}(1-b)} \geq \frac{2\cdot\frac{31}{32}\cdot 1}{\frac{25}{16}\cdot 1} = \frac{31}{25} > 1.
$$

In particular, this implies $\eta\rho_k \in (0,1)$ for all $k \in \mathbb{N}$. We then have

$$
\begin{aligned}
\tilde{I} &= (1-\alpha_k-\rho_k\eta)\Big(\frac{(3-a)(1-\alpha_k)}{2\rho_k(1+\tilde{L}\lambda)} - 1\Big) - \alpha_k^2(2-\rho_k\eta) \\
&\geq \Big(1-\alpha_k-\frac{1-\alpha_k}{1+4\alpha_k}\Big)\Big(\frac{2\alpha_k^2-0.5\alpha_k+1}{1-\alpha_k} - 1\Big) - 2\alpha_k^2 \\
&= \frac{(1-\alpha_k)4\alpha_k}{1+4\alpha_k}\cdot\frac{2\alpha_k^2+0.5\alpha_k}{1-\alpha_k} - \frac{2\alpha_k^2(1+4\alpha_k)}{1+4\alpha_k} \\
&= 0.
\end{aligned}
\tag{52}
$$

Next, we show that $H_k \geq \frac{1-\bar{\alpha}}{2}\|X_k-\bar{x}\|^2$, for H_k defined in (45). This can be seen from the next string of inequalities:

[2] To wit, the function $x \mapsto 2x^2 - 0.5x + 1$ is attains a global minumum at $x = 1/8$, which gives the global lower bound 31/32. Furthermore, the function $x \mapsto (1-x)(1+4x)$ attains a global maximum at $x = 3/8$, with corresponding value 25/16.

$$
\begin{aligned}
H_k &= \|X_k - \bar{x}\|^2 - \alpha_k \|X_{k-1} - \bar{x}\|^2 + (1-\alpha_k)\left(\frac{3-a}{2\rho_k(1+\tilde{L}\lambda)} - 1\right)\|X_k - X_{k-1}\|^2 \\
&\geq \|X_k - \bar{x}\|^2 + \left(\frac{(1-\alpha_k)(2\alpha_k^2 + 1 - 0.5\alpha_k)}{(1-\alpha_k)^2} - 1 + \alpha_k\right)\|X_k - X_{k-1}\|^2 \\
&\quad - \alpha_k \|X_{k-1} - \bar{x}\|^2 \\
&\geq \|X_k - \bar{x}\|^2 + \left(\frac{(1-\alpha_k)(2\alpha_k^2 + 1 - \alpha_k)}{(1-\alpha_k)^2} - 1 + \alpha_k\right)\|X_k - X_{k-1}\|^2 \\
&\quad - \alpha_k \|X_{k-1} - \bar{x}\|^2 \\
&= \left(\alpha_k + \frac{1-\alpha_k}{2}\right)\|X_k - \bar{x}\|^2 + \left(\alpha_k + \frac{2\alpha_k^2}{1-\alpha_k}\right)\|X_k - X_{k-1}\|^2 - \alpha_k \|X_{k-1} - \bar{x}\|^2 \\
&\quad + \frac{1-\alpha_k}{2}\|X_k - \bar{x}\|^2 \\
&\geq \alpha_k\left(\|X_k - \bar{x}\|^2 + \|X_k - X_{k-1}\|^2\right) - \alpha_k \|X_{k-1} - \bar{x}\|^2 \\
&\quad + 2\alpha_k \|X_k - \bar{x}\| \cdot \|X_k - X_{k-1}\| + \frac{1-\alpha_k}{2}\|X_k - \bar{x}\|^2 \\
&\geq \alpha_k\left(\|X_k - \bar{x}\| + \|X_k - X_{k-1}\|\right)^2 - \alpha_k \|X_{k-1} - \bar{x}\|^2 \\
&\quad + \frac{1-\alpha_k}{2}\|X_k - \bar{x}\|^2 \geq \frac{1-\alpha_k}{2}\|X_k - \bar{x}\|^2 \\
&\geq \frac{1-\bar{\alpha}}{2}\|X_k - \bar{x}\|^2.
\end{aligned}
$$

In this derivation we have used the (9) to estimate $\frac{1-\alpha_k}{2}\|X_k - \bar{x}\|^2 + \frac{2\alpha_k^2}{1-\alpha_k}\|X_k - X_{k-1}\|^2 \geq 2\alpha_k \|X_k - \bar{x}\| \cdot \|X_k - X_{k-1}\|$, and the specific choice $\rho_k = \frac{(3-a)(1-\alpha_k)^2}{2(2\alpha_k^2 - \frac{1}{2}\alpha_k + 1)(1+\tilde{L}\lambda)}$.

By recalling (51) and invoking (52), we are left with the stochastic recursion

$$
H_{k+1} \leq q_k H_k - \tilde{b}_k + \Delta M_k. \tag{53}
$$

where $q_k \triangleq 1 - \rho_k \eta$ and $\tilde{b}_k \triangleq \alpha_k \rho_k \eta_k \|X_k - \bar{x}\|^2$. Since $\rho_k = \frac{(3-a)(1-\alpha_k)^2}{2(2\alpha_k^2 - \frac{1}{2}\alpha_k + 1)(1+\tilde{L}\lambda)} \geq \rho = \frac{16(3-a)(1-\bar{\alpha})^2}{31(1+\tilde{L}\lambda)}$ for every k, we have that $q_k \leq q = 1 - \eta\rho$ for every k. Furthermore, $1 > \eta\rho_k \geq \eta\rho$, so that $q \in (0,1)$. Taking conditional expectations on both sides on (53), we get

$$
\mathbb{E}[H_{k+1}|\mathcal{F}_k] + \tilde{b}_k \leq q H_k + c_k \quad \mathbb{P}\text{-a.s.}
$$

using the notation $c_k \triangleq \mathbb{E}[\Delta M_k|\mathcal{F}_k]$. Applying the operator $\mathbb{E}[\cdot|\mathcal{F}_{k-1}]$ and using the tower property of conditional expectations, this gives

$$
\mathbb{E}[H_{k+1}|\mathcal{F}_{k-1}] \leq q^2 H_{k-1} - q\mathbb{E}[\tilde{b}_{k-1}|\mathcal{F}_{k-1}] - \mathbb{E}[\tilde{b}_k|\mathcal{F}_{k-1}] + q\mathbb{E}[c_{k-1}|\mathcal{F}_{k-1}] + \mathbb{E}[c_k|\mathcal{F}_{k-1}].
$$

Proceeding inductively, we see that

$$\mathbb{E}[H_{k+1}|\mathcal{F}_1] \leq q^k H_1 + \sum_{i=1}^{k-1} q^{k-i}\mathbb{E}[c_i|\mathcal{F}_1] = q^k H_1 + \bar{c}_k.$$

This establishes eq. (46). To validate eq. (47), recall that we assume $X_1 = X_0$, so that $H_1 = (1-\alpha_1)\|X_1 - \bar{x}\|^2$. Furthermore, $H_{k+1} \geq \frac{1-\bar{\alpha}}{2}\|X_{k+1} - \bar{x}\|^2$, so that

$$\mathbb{E}[\|X_{k+1} - \bar{x}\|^2|\mathcal{F}_1] \leq q^k\left(\frac{2(1-\alpha_1)}{1-\bar{\alpha}}\|X_1 - \bar{x}\|^2\right) + \frac{2}{1-\bar{\alpha}}\bar{c}_k.$$

We now show that $(\bar{c}_k)_{k\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$. Simple algebra, combined with Assumption 6, gives

$$\begin{aligned} c_k = \mathbb{E}[\Delta M_k|\mathcal{F}_k] &\leq 2\rho_k\left(1 + \frac{(3-a)\lambda^2}{1+\tilde{L}\lambda}\right)\mathbb{E}[\|W_k\|^2|\mathcal{F}_k] + \frac{2(3-a)\rho_k\lambda^2}{1+\tilde{L}\lambda}\mathbb{E}[\|U_k\|^2|\mathcal{F}_1] \\ &\leq \frac{2\mathrm{s}^2\rho_k}{m_k}\left(1 + \frac{2(3-a)\lambda^2}{1+\tilde{L}\lambda}\right) \equiv \frac{\rho_k \mathrm{s}^2}{m_k}\kappa. \end{aligned} \tag{54}$$

Hence, since $(\rho_k)_{k\in\mathbb{N}}$ is bounded, Assumption 5 gives $\lim_{k\to\infty} c_k = 0$ a.s. Using again the tower property, we see $\mathbb{E}[c_k|\mathcal{F}_1] = \mathbb{E}\big[\mathbb{E}(c_k|\mathcal{F}_k)|\mathcal{F}_1\big] \leq \kappa\frac{\rho_k \mathrm{s}^2}{m_k} \leq \kappa\frac{\bar{\rho}\mathrm{s}^2}{m_k}$, where $\rho_k = \frac{(3-a)(1-\alpha_k)^2}{2(2\alpha_k^2 - \frac{1}{2}\alpha_k + 1)(1+\tilde{L}\lambda)} \leq \bar{\rho} = \frac{3-a}{2(1+\tilde{L}\lambda)}$ for every k. Consequently, the discrete convolution $\left(\sum_{i=1}^{k-1} q^{k-i}\mathbb{E}[c_i|\mathcal{F}_1]\right)_{k\in\mathbb{N}}$ is summable. Therefore $\sum_{k\geq 1}\mathbb{E}[H_k] < \infty$ and $\sum_{k\geq 1}\mathbb{E}[\tilde{b}_k] < \infty$. Clearly, this implies $\lim_{k\to\infty}\mathbb{E}[\tilde{b}_k] = 0$, and consequently the subsequently stated two implication follow as well:

$$\begin{aligned} &\lim_{k\to\infty}\|X_k - \bar{x}\| = 0 \quad \mathbb{P}\text{-a.s.,} \qquad \text{and} \\ &\sum_{k=1}^{\infty}(1-\alpha_k)\left(\tfrac{3-a}{2\rho_k(1+\tilde{L}\lambda)} - 1\right)\|X_k - X_{k-1}\|^2 < \infty \quad \mathbb{P}\text{-a.s..} \end{aligned}$$

□

Remark 4 It is worth remarking that the above proof does not rely on unbiasedness of the random estimators. The reason why we can lift this rather typical assumption lies in our application Young's inequality in the estimate (48). The only assumption needed is a summable oracle variance as formulated in Assumption 6 to get the above result working.

Remark 5 The above result illustrates again nicely the well-known trade-off between relaxation and inertial effects (cf. Remark 2). Indeed, up to constant factors, the coupling between inertia and relaxation is expressed by the function $\alpha \mapsto \frac{(1-\alpha)^2}{2\alpha^2 - \frac{1}{2}\alpha + 1}$. Basic calculus reveals that this function is decreasing for α increasing. In the extreme case when $\alpha \uparrow 1$, it is necessary to let $\rho \downarrow 0$, and vice versa. When $\alpha \to 0$ then the limiting value of our specific relaxation policy is $\frac{3-a}{1+\tilde{L}\lambda}$. In practical applications, it is advisable to choose b small in order to make q large. The value a must be calibrated in a disciplined way in order to allow for a sufficiently large step size λ. This requires some knowledge of the condition number of the problem μ/L. As a heuristic argument, a good strategy, anticipating that b should be close to 0, is to set $\frac{a}{2\mu} = \frac{1-a}{2\tilde{L}}$. This means $a = \frac{\mu}{\tilde{L}+\mu}$.

We obtain a full linear rate of convergence when a more aggressive sample rate is employed in the SO. We achieve such global linear rates, together with tuneable iteration and oracle complexity estimates in two settings: First, we consider an aggressive simulation strategy, where the sample size grows over time geometrically. Such a sampling frequency can be quite demanding in some applications. As an alternative, we then move on and consider a more modest simulation strategy under which only polynomial growth of the batch size is required. Whatever simulation strategy is adopted, key to the assessment of the iteration and oracle complexity is to bound the stopping time

$$K_\epsilon \triangleq \inf\{k \in \mathbb{N} \mid \mathbb{E}\left(\|X_{k+1} - \bar{x}\|^2\right) \leq \epsilon\}. \tag{55}$$

In order to understand the definition of this stopping time, recall that RISFBF computes the last iterate X_{K+1} by extrapolating between the current base point Z_k and the correction step involving $Y_k + \lambda_K(A_k - B_k)$, which requires $2m_k$ iid realizations from the law P. In total, when executing the algorithm until the terminal time K_ϵ, we therefore need to simulate $2\sum_{k=1}^{K_\epsilon} m_k$ random variables. We now estimate the integer K_ϵ under a geometric sampling strategy.

Proposition 9 (Non-asymptotic linear convergence under geometric sampling) *Suppose the conditions of Theorem 8 hold. Let* $p \in (0,1), \mathtt{B} = 2\bar{\rho}\mathtt{s}^2\left(1 + \frac{2(3-a)\lambda^2}{1+\tilde{L}\lambda}\right)$, *and choose the sampling rate* $m_k = \lfloor p^{-k} \rfloor$. *Let* $\hat{p} \in (p,1)$, *and define*

$$C(p,q) \triangleq \tfrac{2(1-\alpha_1)}{1-\bar{\alpha}}\|X_1 - \bar{x}\|^2 + \tfrac{4\mathtt{B}}{(1-\bar{\alpha})(1-\min\{p/q,q/p\})} \quad \text{if } p \neq q, \text{ and} \tag{56}$$

$$\hat{C} \triangleq \tfrac{2(1-\alpha_1)}{1-\bar{\alpha}}\|X_1 - \bar{x}\|^2 + \tfrac{4\mathtt{B}}{(1-\bar{\alpha})\exp(1)\ln(\hat{p}/q)} \quad \text{if } p = q. \tag{57}$$

Then, whenever $p \neq q$, *we see that*

$$\mathbb{E}\left(\|X_{k+1} - \bar{x}\|^2\right) \leq C(p,q)\max\{p,q\}^k,$$

and whenever $p = q$,

$$\mathbb{E}\left(\|X_{k+1} - \bar{x}\|^2\right) \leq \hat{C}\hat{p}^k.$$

In particular, the stochastic process $(X_k)_{k\in\mathbb{N}}$ *converges strongly and* $\mathbb{P}$*-a.s. to the unique solution* $\bar{x}$ *at a linear rate.*

Proof Departing from (53), ignoring the positive term $\tilde{b}_k$ from the right-hand side, and taking expectations on both sides leads to

$$\frac{1-\bar{\alpha}}{2}\mathbb{E}(\|X_{k+1}-\bar{x}\|^2) \leq h_{k+1} \triangleq \mathbb{E}(H_{k+1}) \leq q\mathbb{E}(H_k) + c_k = qh_k + c_k, \tag{58}$$

where the equality follows from c_k being deterministic. The sequence $(c_k)_{k\in\mathbb{N}}$ is further upper bounded by the following considerations: First, the relaxation sequence is bounded by $\rho_k \leq \bar{\rho} = \frac{3-a}{2(1+\tilde{L}\lambda)}$; Second, the sample rate is bounded by $m_k = \lfloor p^{-k} \rfloor \geq \left\lceil \frac{1}{2}p^{-k} \right\rceil \geq \frac{1}{2}p^{-k}$. Using these facts, eq. (54) yields

$$c_k \leq \frac{\rho_k \mathsf{s}^2 \kappa}{m_k} \leq 2\mathsf{B}p^k \qquad \forall k \geq 1, \tag{59}$$

where $\mathsf{B} = 2\bar{\rho}\mathsf{s}^2\left(1 + \frac{2(3-a)\lambda^2}{1+\tilde{L}\lambda}\right)$. Iterating the recursion above, one readily sees that

$$h_{k+1} \leq q^k h_1 + \sum_{i=1}^{k} q^{k-i} c_i \quad \forall k \geq 1. \tag{60}$$

Consequently, by recalling that $h_1 = (1-\alpha_1)\|X_1 - \bar{x}\|^2$ and $h_k \geq \frac{1-\bar{\alpha}}{2}\mathbb{E}(\|X_k - \bar{x}\|^2)$, the bound (59) allows us to derive the recursion

$$\mathbb{E}\left(\|X_{k+1}-\bar{x}\|^2\right) \leq q^k\left(\frac{2(1-\alpha_1)}{1-\bar{\alpha}}\|X_1 - \bar{x}\|^2\right) + \frac{4\mathsf{B}}{1-\bar{\alpha}}\sum_{i=1}^{k} q^{k-i}p^i. \tag{61}$$

We consider three cases.

(i) $0<q<p<1$: Defining $\mathsf{c}_1 \triangleq \frac{2(1-\alpha_1)}{1-\bar{\alpha}}\|X_1-\bar{x}\|^2 + \frac{4\mathsf{B}}{(1-\bar{\alpha})(1-q/p)}$, we obtain from (61)

$$\mathbb{E}(\|X_{k+1}-\bar{x}\|^2) \leq q^k\left(\frac{2(1-\alpha_1)}{1-\bar{\alpha}}\|X_1-\bar{x}\|^2\right) + \frac{4\mathsf{B}}{1-\bar{\alpha}}\sum_{i=1}^{k}(q/p)^{k-i}p^k \leq \mathsf{c}_1 p^k.$$

(ii) $0<p<q<1$. Akin to (i) and defining $\mathsf{c}_2 \triangleq \frac{2(1-\alpha_1)}{1-\bar{\alpha}}\|X_1-\bar{x}\|^2 + \frac{4\mathsf{B}}{(1-\bar{\alpha})(1-p/q)}$, we arrive as above at the bound $\mathbb{E}(\|X_k-\bar{x}\|^2) \leq q^k\mathsf{c}_2$.

(iii) $p=q<1$. Choose $\hat{p}\in(q,1)$ and $\mathsf{c}_3 \triangleq \frac{1}{\exp(1)\ln(\hat{p}/q)}$, so that Lemma 18 yields $kq^k \leq \mathsf{c}_3\hat{p}^k$ for all $k\geq 1$. Therefore, plugging this estimate in eq. (61), we see

$$\begin{aligned}\mathbb{E}(\|X_k-\bar{x}\|^2) &\leq q^k\left(\frac{2(1-\alpha_1)}{1-\bar{\alpha}}\|X_1-\bar{x}\|^2\right) + \frac{4\mathsf{B}}{1-\bar{\alpha}}\sum_{i=1}^{k} q^k \\ &\leq \hat{p}^k\left(\frac{2(1-\alpha_1)}{1-\bar{\alpha}}\|X_1-\bar{x}\|^2\right) + \frac{4\mathsf{B}}{1-\bar{\alpha}}\mathsf{c}_3\hat{p}^k \\ &= \mathsf{c}_4\hat{p}^k,\end{aligned}$$

after setting $\mathsf{c}_4 \triangleq \frac{2(1-\alpha_1)}{1-\bar{\alpha}}\left\|X_1-\bar{x}\right\|^2+\frac{4\mathsf{B}\mathsf{c}_3}{1-\bar{\alpha}}$. Collecting these three cases together, verifies the first part of the proposition.

□

Proposition 10 (*Oracle and Iteration Complexity under geometric sampling*) *Given* $\epsilon > 0$, *define the stopping time* K_ϵ *as in eq.* (55). *Define*

$$\tau_\epsilon(p,q) \triangleq \begin{cases} \lceil \frac{\ln(C(p,q)\epsilon^{-1})}{\ln(1/\max\{p,q\})} \rceil & \text{if } p \neq q, \\ \lceil \frac{\ln(\hat{C}\epsilon^{-1})}{\ln(1/\hat{p})} \rceil & \text{if } p = q \end{cases} \tag{62}$$

and the same hypothesis as in Theorem 8 *hold true. Then,* $K_\varepsilon \leq \tau_\epsilon(p,q) = \mathcal{O}(\ln(\varepsilon^{-1}))$. *The corresponding oracle complexity of RISFBF is upper bounded as* $2\sum_{i=1}^{\tau_\epsilon(p,q)} m_i = \mathcal{O}\big((1/\epsilon)^{1+\delta(p,q)}\big)$, *where*

$$\delta(p,q) \triangleq \begin{cases} 0 & \text{if } p > q, \\ \frac{\ln(p)}{\ln(q)} - 1 & \text{if } p \in (0,q), \\ \frac{\ln(p)}{\ln(\hat{p})} - 1 & \text{if } p = q. \end{cases}$$

Proof First, let us recall that the total oracle complexity of the method is assessed by

$$2\sum_{i=1}^{K_\epsilon} m_i = 2\sum_{i=1}^{K_\epsilon} \lfloor p^{-i} \rfloor \leq 2\sum_{i=1}^{K_\epsilon} p^{-i}.$$

If $p \neq q$ define $\tau_\epsilon \equiv \tau_\epsilon(p,q) = \lceil \frac{\ln(C(p,q)\epsilon^{-1})}{\ln(1/\max\{p,q\})} \rceil$. Then, $\mathbb{E}(\left\|X_{\tau_\epsilon+1}-\bar{x}\right\|^2) \leq \epsilon$, and hence $K_\epsilon \leq \tau_\epsilon$. We now compute

$$\sum_{i=1}^{\tau_\epsilon}(1/p)^i = \frac{1}{p}\frac{(1/p)^{\lceil \frac{\ln(C(p,q)\epsilon^{-1})}{\ln(1/\max\{p,q\})} \rceil} - 1}{1/p-1} \leq \frac{1}{p^2}\frac{(1/p)^{\frac{\ln(C(p,q)\epsilon^{-1})}{\ln(1/\max\{p,q\})}}}{1/p-1}$$
$$= \frac{\big(\epsilon^{-1}C(p,q)\big)^{\ln(1/p)/\ln(1/\max\{p,q\})}}{p(1-p)}.$$

This gives the oracle complexity bound

$$2\sum_{i=1}^{\tau_\epsilon} m_i \leq 2\frac{\big(\epsilon^{-1}C(p,q)\big)^{\ln(1/p)/\ln(1/\max\{p,q\})}}{p(1-p)}.$$

If $p = q$, we can replicate this calculation, after setting $\tau_\epsilon = \lceil \frac{\ln(\epsilon^{-1}\hat{C})}{\ln(1/\hat{p})} \rceil$. After so many iterations, we can be ensured that $\mathbb{E}(\left\|X_{\tau_\epsilon+1}-\bar{x}\right\|^2) \leq \epsilon$, with an oracle complexity

$$2\sum_{i=1}^{\tau_\epsilon} m_i \leq \frac{2}{\hat{p}(1-\hat{p})}\left(\frac{\hat{C}}{\epsilon}\right)^{\ln(p)/\ln(\hat{p})}.$$

□

To the best of our knowledge, the provided non-asymptotic linear convergence guarantee appears to be amongst the first in relaxed and inertial splitting algorithms. In particular, by leveraging the increasing nature of mini-batches, this result no longer requires the unbiasedness assumption on the SO, a crucial benefit of the proposed scheme.

There may be settings where geometric growth of m_k is challenging to adopt. To this end, we provide a result where the sampling rate is polynomial rather than geometric. A polynomial sampling rate arises if $m_k = \lceil a_k(k+k_0)^\theta + b_k \rceil$ for some parameters $a_k, b_k, \theta > 0$. Such a regime has been adopted in related mini-batch approaches [75, 76]. This allows for modulating the growth rate by changing the exponent in the sampling rate. We begin by providing a supporting result. We make the specific choice $a_k = b_k = 1$ for all $k \geq 1$, and $k_0 = 0$, leaving essentially the exponent $\theta > 0$ as a free parameter in the design of the stochastic oracle.

Proposition 11 (Polynomial rate of convergence under polynomially increasing m_k) *Suppose the conditions of Theorem 8 hold. Choose the sampling rate $m_k = \lfloor k^\theta \rfloor$ where $\theta > 0$. Then, for any $k \geq 1$,*

$$\mathbb{E}(\|X_{k+1} - \bar{x}\|^2) \leq q^k \left(\frac{2(1-\alpha_1)}{1-\bar{\alpha}} \|X_1 - \bar{x}\|^2 + \frac{2}{1-\bar{\alpha}} \frac{q^{-1}\exp(2\theta) - 1}{1-q} \right) + \frac{4\mathtt{B}}{(1-\bar{\alpha})q\ln(1/q)}(k+1)^{-\theta} \tag{63}$$

Proof From the relation (60), we obtain

$$\begin{aligned} h_{k+1} &\leq q^k h_1 + \sum_{i=1}^{k} q^{k-i} c_i \leq q^k h_1 + \mathtt{B}\sum_{i=1}^{k} q^{k-i} i^{-\theta} \\ &= q^k \left(h_1 + \mathtt{B}\sum_{i=1}^{k} q^{-i} i^{-\theta} \right) \\ &= q^k \left(h_1 + \mathtt{B} \sum_{i=1}^{\lceil 2\theta/\ln(1/q) \rceil} q^{-i} i^{-\theta} + \mathtt{B} \sum_{i=\lceil 2\theta/\ln(1/q)\rceil+1}^{k} q^{-i} i^{-\theta} \right). \end{aligned}$$

A standard bound based on the integral criterion for series with non-negative summands gives

$$\sum_{i=\lceil 2\theta/\ln(1/q)\rceil+1}^{k} q^{-i} i^{-\theta} \leq \int_{\lceil 2\theta/\ln(1/q)\rceil}^{k+1} \frac{(1/q)^t}{t^\theta}\, dt.$$

The upper bounding integral can be evaluated using integration-by-parts, as follows:

$$\int_{\lceil 2\theta/\ln(1/q)\rceil}^{k+1} \frac{(1/q)^t}{t^\theta}\, dt = t^\theta \frac{e^{t\ln(1/q)}}{\ln(1/q)} \Big|_{t=\lceil 2\theta/\ln(1/q)\rceil}^{t=k+1} + \int_{\lceil 2\theta/\ln(1/q)\rceil}^{k+1} \theta t^{-(\theta+1)} \frac{e^{t\ln(1/q)}}{\ln(1/q)}\, dt.$$

Note that $\frac{\theta}{t\ln(1/q)} \leq \frac{1}{2}$ when $t \geq \lceil 2\theta/\ln(1/q)\rceil$. Therefore, we can attain a simpler bound from the above by

$$\int_{\lceil 2\theta/\ln(1/q)\rceil}^{k+1} \frac{(1/q)^t}{t^\theta}\,dt \leq \frac{(1/q)^{k+1}}{\ln(1/q)(k+1)^\theta} + \frac{1}{2}\int_{\lceil 2\theta/\ln(1/q)\rceil}^{k+1} \frac{(1/q)^t}{t^\theta}\,dt$$

Consequently,

$$\int_{\lceil 2\theta/\ln(1/q)\rceil}^{k+1} \frac{(1/q)^t}{t^\theta}\,dt \leq \frac{2(1/q)^{k+1}(k+1)^{-\theta}}{\ln(1/q)}.$$

Furthermore,

$$\sum_{i=1}^{\lceil 2\theta/\ln(1/q)\rceil} q^{-i} i^{-\theta} \leq \sum_{i=1}^{\lceil 2\theta/\ln(1/q)\rceil} q^{-i} = \frac{1}{q}\frac{(1/q)^{\lceil 2\theta/\ln(1/q)\rceil}-1}{1/q-1} \leq \frac{1}{q}\frac{(1/q)^{2\theta/\ln(1/q)+1}-1}{1/q-1}.$$

Note that $(1/q)^{2\theta/\ln(1/q)} = (\exp(\ln(1/q)))^{2\theta/\ln(1/q)} = \exp(2\theta)$. Hence,

$$\sum_{i=1}^{\lceil 2\theta/\ln(1/q)\rceil} q^{-i} i^{-\theta} \leq \frac{1}{q}\frac{q^{-1}\exp(2\theta)-1}{1/q-1} = \frac{q^{-1}\exp(2\theta)-1}{1-q}.$$

Plugging this into the opening string of inequalities shows

$$\begin{aligned} h_{k+1} &\leq q^k\left(h_1 + \mathtt{B}\sum_{i=1}^{\lceil 2\theta/\ln(1/q)\rceil} q^{-i} + \frac{2\mathtt{B}(1/q)^{k+1}(k+1)^{-\theta}}{\ln(1/q)}\right) \\ &\leq q^k\left(h_1 + \mathtt{B}\frac{q^{-1}\exp(2\theta)-1}{1-q} + \frac{2\mathtt{B}(1/q)^{k+1}(k+1)^{-\theta}}{\ln(1/q)}\right) \\ &= q^k\left(h_1 + \mathtt{B}\frac{q^{-1}\exp(2\theta)-1}{1-q}\right) + \frac{2\mathtt{B}/q}{\ln(1/q)}(k+1)^{-\theta}. \end{aligned}$$

Since $h_1 = (1-\alpha_1)\|X_1 - \bar{x}\|^2$ and $h_{k+1} \geq \frac{1-\bar{\alpha}}{2}\mathbb{E}\left(\|X_{k+1}-\bar{x}\|^2\right)$, we finally arrive at the desired expression (63). □

Proposition 12 (*Oracle and Iteration complexity under polynomial sampling*) *Let all Assumptions as in Theorem 8 hold. Given $\epsilon > 0$, define K_ϵ as in (55). Then the iteration and oracle complexity to obtain an ϵ-solution are $\mathcal{O}(\theta\varepsilon^{-1/\theta})$ and $\mathcal{O}(\exp(\theta)\theta^\theta(1/\epsilon)^{1+1/\theta})$, respectively.*

Proof We first note that $(k+1)^{-\theta} \leq k^{-\theta}$ for all $k \geq 1$. Hence, the bound established in Proposition 11 yields

$$\begin{aligned}\mathbb{E}(\|X_{k+1}-\bar{x}\|^2) \le \quad & q^k\left(\frac{2(1-\alpha_1)}{1-\bar{\alpha}}\|X_1-\bar{x}\|^2+\frac{2}{1-\bar{\alpha}}\frac{q^{-1}\exp(2\theta)-1}{1-q}\right)\\ & +\frac{4\mathsf{B}}{(1-\bar{\alpha})q\ln(1/q)}k^{-\theta}\end{aligned}$$

Consider the function $\psi(t) \triangleq t^\theta q^t$ for $t > 0$. Then, straightforward calculus shows that $\psi(t)$ is unimodal on $(0,\infty)$, with unique maximum $t^* = \frac{\theta}{\ln(1/q)}$ and associated value $\psi(t^*) = \exp(-\theta)\left(\frac{\theta}{\ln(1/q)}\right)^\theta$. Hence, for all $t > 0$, we have $t^\theta q^t \le \exp(-\theta)\left(\frac{\theta}{\ln(1/q)}\right)^\theta$, and consequently, $q^k \le \exp(-\theta)\left(\frac{\theta}{\ln(1/q)}\right)^\theta k^{-\theta}$ for all $k \ge 1$. This allows us to conclude

$$\begin{aligned}\mathbb{E}(\|X_{k+1}-\bar{x}\|^2) \le \exp(-\theta)\left(\frac{\theta}{\ln(1/q)}\right)^\theta & k^{-\theta}\\ & \left(\frac{2(1-\alpha_1)}{1-\bar{\alpha}}\|X_1-\bar{x}\|^2+\frac{2}{1-\bar{\alpha}}\frac{q^{-1}\exp(2\theta)-1}{1-q}\right)\\ & +\frac{4\mathsf{B}}{(1-\bar{\alpha})q\ln(1/q)}k^{-\theta} = \mathsf{c}_{q,\theta}k^{-\theta},\end{aligned}$$

where

$$\begin{aligned}\mathsf{c}_{q,\theta} \triangleq \exp(-\theta)\left(\frac{\theta}{\ln(1/q)}\right)^\theta & \left(\frac{2(1-\alpha_1)}{1-\bar{\alpha}}\|X_1-\bar{x}\|^2+\frac{2}{1-\bar{\alpha}}\frac{q^{-1}\exp(2\theta)-1}{1-q}\right)\\ +\frac{4\mathsf{B}}{(1-\bar{\alpha})q\ln(1/q)} &\end{aligned} \tag{64}$$

Then, for any $k \ge K_\epsilon \triangleq \lceil(\mathsf{c}_{q,\theta}/\epsilon)^{1/\theta}\rceil$, we are ensured that $\mathbb{E}(\|X_{k+1}-\bar{x}\|^2) \le \varepsilon$. Since $(\mathsf{c}_{q,\theta})^{1/\theta} = \mathcal{O}(\exp(-1)\theta)$, we conclude that $K_\epsilon = \mathcal{O}(\theta\epsilon^{-1/\theta})$. The corresponding oracle complexity is bounded as follows:

$$2\sum_{i=1}^{K_\epsilon} m_i \le 2\sum_{i=1}^{K_\epsilon} i^\theta \le 2\int_1^{K_\epsilon+1} t^\theta\, dt \le \frac{2}{1+\theta}\left(\lceil\frac{\mathsf{c}_{q,\theta}}{\epsilon}\rceil^{1/\theta}+1\right)^{1+\theta} = \mathcal{O}(\exp(\theta)\theta^\theta(1/\epsilon)^{1+1/\theta}).$$

□

Remark 6 It may be observed that if the $\theta = 1$ or $m_k = k$, there is a worsening of the rate and complexity statements from their counterparts when the sampling rate is geometric; in particular, the iteration complexity worsens from $\mathcal{O}(\ln(\frac{1}{\epsilon}))$ to $\mathcal{O}(\frac{1}{\epsilon})$ while the oracle complexity degrades from the optimal level of $\mathcal{O}(\frac{1}{\epsilon})$ to $\mathcal{O}(\frac{1}{\epsilon^2})$. But this deterioration comes with the advantage that the sampling rate is far slower and this may be of significant consequence in some applications.

4.3 Rates in terms of merit functions

In this subsection we estimate the iteration and oracle complexity of RISFBF with the help of a suitably defined *gap function*. Generally, a gap function associated with the monotone inclusion problem (MI) is a function $\mathsf{Gap} : \mathsf{H} \to \mathbb{R}$ such that (i) Gap is sign restricted on H; and (ii) $\mathsf{Gap}(x) = 0$ if and only if $x \in \mathsf{S}$. The *Fitzpatrick function* [3, 30, 31, 77] is a useful tool to construct gap functions associated with a set-valued operator $F : \mathsf{H} \to 2^{\mathsf{H}}$. It is defined as the function $G_F : \mathsf{H} \times \mathsf{H} \to [-\infty, \infty]$ given by

$$G_F(x, x^*) = \langle x, x^* \rangle - \inf_{(y,y^*) \in \operatorname{gr}(F)} \langle x - y, x^* - y^* \rangle. \tag{65}$$

This function allows us to recover the operator F, by means of the following result (cf. [3, Prop. 20.58]): If $F : \mathsf{H} \to 2^{\mathsf{H}}$ is maximally monotone, then $G_F(x, x^*) \geq \langle x, x^* \rangle$ for all $(x, x^*) \in \mathsf{H} \times \mathsf{H}$, with equality if and only if $(x, x^*) \in \operatorname{gr}(F)$. In particular, $\operatorname{gr}(F) = \{(x, x^*) \in \mathsf{H} \times \mathsf{H} |\ G_F(x, x^*) \geq \langle x, x^* \rangle\}$. In fact, it can be shown that the Fitzpatrick function is minimal in the family of convex functions $f : \mathsf{H} \times \mathsf{H} \to (-\infty, \infty]$ such that $f(x, x^*) \geq \langle x, x^* \rangle$ for all $(x, x^*) \in \mathsf{H} \times \mathsf{H}$, with equality if $(x, x^*) \in \operatorname{gr}(F)$ [77].

Our gap function for the structured monotone operator $F = V + T$ is derived from its Fitzpatrick function by setting $\mathsf{Gap}(x) \triangleq G_F(x, 0)$ for $x \in \mathsf{H}$. This reads explicitly as

$$\mathsf{Gap}(x) \triangleq \sup_{(y,y^*) \in \operatorname{gr}(F)} \langle y^*, x - y \rangle = \sup_{p \in \operatorname{dom} T} \sup_{p^* \in T(p)} \langle V(p) + p^*, x - p \rangle \qquad \forall x \in \mathsf{H}. \tag{66}$$

It immediately follows from the definition that $\mathsf{Gap}(x) \geq 0$ for all $x \in \mathsf{H}$. It is also clear, that $x \mapsto \mathsf{Gap}(x)$ is convex and lower semi-continuous and $\mathsf{Gap}(x) = 0$ if and only if $x \in \mathsf{S} = \mathsf{Zer}(F)$. Let us give some concrete formulae for the gap function.

Example 4 (Variational Inequalities) We reconsider the problem described in Example 2. Let $V : \mathsf{H} \to \mathsf{H}$ be a maximally monotone and L-Lipschitz continuous map, and $T(x) = \mathsf{N}_{\mathsf{C}}(x)$ the normal cone of a given closed convex set $\mathsf{C} \subset \mathsf{H}$. Then, by [77, Prop. 3.3], the gap function (66) reduces to the well-known *dual gap function*, due to [78],

$$\mathsf{Gap}(x) = \sup_{p \in \mathsf{C}} \langle V(p), x - p \rangle.$$

Example 5 (Convex Optimization) Reconsider the general non-smooth convex optimization problem in Example 1, with primal objective function $\mathsf{H}_1 \ni u \mapsto f(u) + g(Lu) + h(u)$. Let us introduce the convex-concave function

$$\mathcal{L}(u, v) \triangleq f(u) + h(u) - g^*(v) + \langle Lu, v \rangle \qquad \forall (u, v) \in \mathsf{H}_1 \times \mathsf{H}_2.$$

Define

$$\Gamma(x') \triangleq \sup_{u \in \mathsf{H}_1, v \in \mathsf{H}_2} \big(\mathcal{L}(u', v) - \mathcal{L}(u, v') \big) \qquad \forall x' = (u', v') \in \mathsf{H} = \mathsf{H}_1 \times \mathsf{H}_2. \tag{67}$$

It is easy to check that $\Gamma(x') \geq 0$, and equality holds only for a primal-dual pair (saddle-point) $\bar{x} \in \mathsf{S}$. Hence, $\Gamma(\cdot)$ is a gap function for the monotone inclusion derived from the Karush-Kuhn-Tucker conditions (5). In fact, the function (67) is a standard merit function for saddle-point problems (see e.g. [79]). To relate this gap function to the Fitzpatrick function, we exploit the maximally monotone operators V and T introduced Example 1. In terms of these mappings, first observe that for $p = (\tilde{u}, \tilde{v}), x = (u, v)$ we have

$$\langle V(p), x - p \rangle = \langle \nabla h(\tilde{u}), u - \tilde{u} \rangle + \langle \tilde{v}, Lu \rangle - \langle L\tilde{u}, v \rangle$$

Since h is convex differentiable, the classical gradient inequality reads as $h(u) - h(\tilde{u}) \geq \langle \nabla h(\tilde{u}), u - \tilde{u} \rangle$. Using this estimate in the previous display shows

$$\langle V(p), x - p \rangle \leq h(u) - h(\tilde{u}) - \langle L\tilde{u}, v \rangle + \langle \tilde{v}, Lu \rangle.$$

For $p^* = (\tilde{u}^*, \tilde{v}^*) \in T(p)$, we again employ convexity to get

$$\begin{aligned} f(u) &\geq f(\tilde{u}) + \langle \tilde{u}^*, u - \tilde{u} \rangle \qquad \forall u \in H_1, \\ g^*(v) &\geq g^*(\tilde{v}) + \langle \tilde{v}^*, v - \tilde{v} \rangle \qquad \forall v \in H_2. \end{aligned}$$

Hence,

$$\langle \tilde{u}^*, u - \tilde{u} \rangle + \langle \tilde{v}^*, v - \tilde{v} \rangle \leq (f(u) - f(\tilde{u})) + (g^*(v) - g^*(\tilde{v})).$$

Therefore, we see

$$\begin{aligned} \langle V(p) + p^*, x - p \rangle &\leq (f(u) + h(u) - g^*(\tilde{v}) + \langle \tilde{v}, Lu \rangle) - (f(\tilde{u}) + h(\tilde{u}) - g^*(v) + \langle v, L\tilde{u} \rangle) \\ &= \mathcal{L}(u, \tilde{v}) - \mathcal{L}(\tilde{u}, v). \end{aligned}$$

Hence,

$$\mathsf{Gap}(x) = \sup_{(p,p^*) \in \operatorname{gr}(T)} \langle V(p) + p^*, x - p \rangle \leq \sup_{(\tilde{u},\tilde{v}) \in \mathsf{H}_1 \times \mathsf{H}_2} (\mathcal{L}(u, \tilde{v}) - \mathcal{L}(\tilde{u}, v)) = \Gamma(x).$$

It is clear from the definition that a convex gap function can be extended-valued and its domain is contingent on the boundedness properties of $\operatorname{dom} T$. In the setting where $T(x)$ is bounded for all $x \in \mathsf{H}$, the gap function is clearly globally defined. However, the case where $\operatorname{dom} T$ is unbounded has to be handled with more care. There are potentially two approaches to cope with such a situation: One would be to introduce a perturbation-based termination criterion as defined in [80], and recently used in [81] to solve a class of structured stochastic variational inequality problems. The other solution strategy is based on the notion of *restricted merit functions*, first introduced in [82], and later on adopted in [83]. We follow the latter strategy.

Let $x^s \in \operatorname{dom} T$ denote an arbitrary reference point and $D > 0$ a suitable constant. Define the closed set $\mathsf{C} \triangleq \operatorname{dom} T \cap \{x \in \mathsf{H} | \, \|x - x^s\| \leq D\}$, and the *restricted gap function*

$$\mathsf{Gap}(x|\mathsf{C}) \triangleq \sup\{\langle y^*, x - y \rangle | y \in \mathsf{C}, y^* \in F(y)\}. \tag{68}$$

Clearly, $\mathsf{Gap}(x|\operatorname{dom} T) = \mathsf{Gap}(x)$. The following result explains in a precise way the meaning of the restricted gap function. It extends the variational case in [82, Lemma 1] and [83, Lemma 3] to the general monotone inclusion case.

Lemma 13 *Let* $\mathsf{C} \subset \mathsf{H}$ *be nonempty closed and convex. The function* $\mathsf{H} \ni x \mapsto \mathsf{Gap}(x|\mathsf{C})$ *is well-defined and convex on* H*. For any* $x \in \mathsf{C}$ *we have* $\mathsf{Gap}(x|\mathsf{C}) \geq 0$*. Moreover, if* $\bar{x} \in \mathsf{C}$ *is a solution to* (*MI*)*, then* $\mathsf{Gap}(\bar{x}|\mathsf{C}) = 0$*. Moreover, if* $\mathsf{Gap}(\bar{x}|\mathsf{C}) = 0$ *for some* $\bar{x} \in \operatorname{dom} T$ *such that* $\|\bar{x} - x^s\| < D$*, then* $\bar{x} \in \mathsf{S}$.

Proof The convexity and non-negativity for $x \in \mathsf{C}$ of the restricted function is clear. Since $\mathsf{Gap}(x|C) \leq \mathsf{Gap}(x)$ for all $x \in \mathsf{H}$, we see

$$\bar{x} \in \mathsf{S} \Leftrightarrow \mathsf{Gap}(\bar{x}) = 0 \Rightarrow \mathsf{Gap}(\bar{x}|\mathsf{C}) = 0.$$

To show the converse implication, suppose $\mathsf{Gap}(\bar{x}|\mathsf{C}) = 0$ for some $\bar{x} \in \mathsf{C}$ with $\|\bar{x} - x^s\| < D$. Without loss of generality we can choose $\bar{x} \in \mathsf{C}$ in this particular way, since we may choose the radius of the ball as large as desired. It follows that $\langle y^*, \bar{x} - y\rangle \leq 0$ for all $y \in \mathsf{C}, y^* \in F(y)$. Hence, $\bar{x} \in \mathsf{C}$ is a Minty solution to the Generalized Variational inequality with maximally monotone operator $F(x) + \mathsf{N}_{\mathsf{C}}(x)$. Since F is upper semi-continuous and monotone, Minty solutions coincide with Stampacchia solutions, implying that there exists $\bar{x}^* \in F(\bar{x})$ such that $\langle \bar{x}^*, y - \bar{x}\rangle \geq 0$ for all $y \in \mathsf{C}$ (see e.g. [84]). Consider now the gap program

$$g_{\mathsf{C}}(\bar{x}, \bar{x}^*) \triangleq \inf\{\langle \bar{x}^*, y - \bar{x}\rangle | y \in \mathsf{C}\}.$$

This program is solved at $y = \bar{x}$, which is a point for which $\|x - x^s\| < D$. Hence, the constraint can be removed, and we conclude $\langle \bar{x}^*, y - \bar{x}\rangle \geq 0$ for all $y \in \operatorname{dom}(F)$. By monotonicity of F, it follows

$$\langle y^*, y - \bar{x}\rangle \geq \langle \bar{x}^*, y - \bar{x}\rangle \geq 0 \quad \forall (y, y^*) \in \operatorname{gr}(F).$$

Hence, $\mathsf{Gap}(\bar{x}) = 0$ and we conclude $\bar{x} \in \mathsf{S}$. □

In order to state and prove our complexity results in terms of the proposed merit function, we start with the first preliminary result.

Lemma 14 *Consider the sequence* $(X_k)_{k\in\mathbb{N}}$ *generated by RISFBF with the initial condition* $X_0 = X_1$*. Suppose* $\lambda_k = \lambda \in (0, 1/(2L))$ *for every* $k \in \mathbb{N}$*. Moreover, suppose* $(\alpha_k)_{k\in\mathbb{N}}$ *is a non-decreasing sequence such that* $0 < \alpha_k \leq \bar{\alpha} < 1$, $\rho_k = \frac{3(1-\bar{\alpha})^2}{2(2\alpha_k^2 - \alpha_k + 1)(1 + L\lambda)}$ *for every* $k \in \mathbb{N}$*. Define*

$$\Delta M_k \triangleq \frac{3\rho_k \lambda_k^2}{1 + L\lambda_k} \|\mathsf{e}_k\|^2 + \frac{\rho_k \lambda_k^2}{2} \|U_k\|^2 \tag{69}$$

and for $(p, p^*) \in \operatorname{gr}(F)$*, we define* $\Delta N_k(p, p^*)$ *as in* (21)*. Then, for all* $(p, p^*) \in \operatorname{gr}(F)$*, we have*

$$\sum_{k=1}^{K} 2\rho_k \lambda \langle p^*, Y_k - p\rangle \leq (1-\alpha_1)\|X_1 - p\|^2 + \sum_{k=1}^{K} \Delta M_k + \sum_{k=1}^{K} \Delta N_k(p, 0). \tag{70}$$

Proof For $(p, p^*) \in$ gr $(V+T)$, we know from eq. (23)

$$\begin{aligned}\langle Z_k - R_k, Y_k - p\rangle &\geq \lambda_k \langle W_k + p^*, Y_k - p\rangle + \lambda_k \langle V(Y_k) - V(p), Y_k - p\rangle \\ &\geq \langle p^*, Y_k - p\rangle + \lambda_k \langle W_k, Y_k - p\rangle,\end{aligned}$$

where the last inequality uses the monotonicity of V. We first derive a recursion which is similar to the fundamental recursion in Lemma 4. Invoking (25) and (26), we get

$$\begin{aligned}\|X_{k+1} - p\|^2 \leq\ &\|Z_k - p\|^2 - \frac{1-\rho_k}{\rho_k}\|X_{k+1} - Z_k\|^2 + 2\lambda^2 \rho_k \|\mathsf{e}_k\|^2 \\ &- 2\rho_k \lambda_k \langle W_k + p^*, Y_k - p\rangle \\ &- \rho_k(1 - 2L^2\lambda_k^2)\|Y_k - Z_k\|^2 + \frac{\rho_k \lambda_k^2}{2}\|U_k\|^2 + 2\rho_k \lambda_k \langle V(Y_k) - V(p), p - Y_k\rangle.\end{aligned} \tag{71}$$

Multiplying both sides of (28) and noting that $(1-2L\lambda_k)(1+L\lambda_k) \leq 1 - 2L^2\lambda_k^2$, we obtain the following inequality

$$-\rho_k(1 - 2L^2\lambda_k^2)\|Y_k - Z_k\|^2 \leq -\frac{1-2L\lambda_k}{2\rho_k(1+L\lambda_k)}\|X_{k+1} - Z_k\|^2 + \frac{\rho_k \lambda_k^2 (1-2L\lambda_k)}{1+L\lambda_k}\|\mathsf{e}_k\|^2.$$

Inserting the above inequality to (71) and using the same fashion in deriving (33), we arrive at

$$\begin{aligned}\|X_{k+1} - p\|^2 \leq &(1+\alpha_k)\|X_k - p\|^2 - \alpha_k \|X_{k-1} - p\|^2 + \Delta M_k \\ &+ \Delta N_k(p, p^*) - 2\rho_k \lambda_k \langle V(Y_k) - V(p), Y_k - p\rangle \\ &+ \alpha_k \|X_k - X_{k-1}\|^2 \left(2\alpha_k + \frac{3(1-\alpha_k)}{2\rho_k(1+L\lambda_k)}\right) \\ &- (1-\alpha_k)\left(\frac{3}{2\rho_k(1+L\lambda_k)} - 1\right)\|X_{k+1} - X_k\|^2.\end{aligned} \tag{72}$$

Invoking the monotonicity of V and rearranging (72), it follows that

$$\begin{aligned}
&\|X_{k+1}-p\|^2+(1-\alpha_k)\Big(\tfrac{3}{2\rho_k(1+L\lambda_k)}-1\Big)\|X_{k+1}-X_k\|^2-\alpha_k\|X_k-p\|^2\\
&\quad\le \|X_k-p\|^2+(1-\alpha_k)\Big(\tfrac{3}{2\rho_k(1+L\lambda_k)}-1\Big)\|X_k-X_{k-1}\|^2\\
&\qquad-\alpha_k\|X_{k-1}-p\|^2+\Delta M_k+\Delta N_k(p,p^*)\\
&\qquad+\underbrace{\Big(2\alpha_k^2+(1-\alpha_k)\Big(1-\tfrac{3(1-\alpha_k)}{2\rho_k(1+L\lambda_k)}\Big)\Big)}_{\le 0}\|X_k-X_{k-1}\|^2\\
&\quad\le \|X_k-p\|^2+(1-\alpha_k)\Big(\tfrac{3}{2\rho_k(1+L\lambda_k)}-1\Big)\|X_k-X_{k-1}\|^2\\
&\qquad-\alpha_k\|X_{k-1}-p\|^2+\Delta M_k+\Delta N_k(p,p^*).
\end{aligned}$$

We define β_{k+1} as

$$\beta_{k+1}\triangleq(1-\alpha_k)\left(\frac{3}{2\rho_k(1+L\lambda_k)}-1\right)-(1-\alpha_{k+1})\left(\frac{3}{2\rho_{k+1}(1+L\lambda_{k+1})}-1\right),$$

and similarly with (36), we can show $\{\beta_k\}$ is non-increasing by choosing $\rho_k=\frac{3(1-\bar{\alpha})^2}{2(2\alpha_k^2-\alpha_k+1)(1+L\lambda_k)}$ and $\lambda_k\equiv\lambda$. Thus, $(1-\alpha_{k+1})\Big(\frac{3}{2\rho_{k+1}(1+L\lambda_{k+1})}-1\Big)\le(1-\alpha_k)\Big(\frac{3}{2\rho_k(1+L\lambda_k)}-1\Big)$. Together with $\alpha_{k+1}\ge\alpha_k$, the last inequality gives

$$\begin{aligned}
&\|X_{k+1}-p\|^2+(1-\alpha_{k+1})\Big(\tfrac{3}{2\rho_{k+1}(1+L\lambda)}-1\Big)\|X_{k+1}-X_k\|^2-\alpha_{k+1}\|X_k-p\|^2\\
&\quad\le \|X_k-p\|^2+(1-\alpha_k)\Big(\tfrac{3}{2\rho_k(1+L\lambda)}-1\Big)\|X_k-X_{k-1}\|^2\\
&\qquad-\alpha_k\|X_{k-1}-p\|^2+\Delta M_k+\Delta N_k(p,p^*).
\end{aligned}$$

Recall that $\Delta N_k(p,p^*)=\Delta N_k(p,0)+2\rho_k\lambda\langle p^*,p-Y_k\rangle$. Hence, after setting $\Delta N_k(p,0)=\Delta N_k(p)$, rearranging the expression given in the previous display shows that

$$\begin{aligned}
2\rho_k\lambda\langle p^*,Y_k-p\rangle\le&\Big(X_k-p^2+(1-\alpha_k)\Big(\tfrac{3}{2\rho_k(1+L\lambda)}-1\Big)X_k\\
&-X_{k-1}^2-\alpha_kX_{k-1}-p^2\Big)\\
&-\Big(X_{k+1}-p^2+(1-\alpha_{k+1})\Big(\tfrac{3}{2\rho_{k+1}(1+L\lambda)}-1\Big)X_{k+1}\\
&-X_k^2-\alpha_{k+1}X_k-p^2\Big)+\Delta M_k+\Delta N_k(p)
\end{aligned}$$

Summing over $k=1,\dots,K$, we obtain

$$
\begin{aligned}
\sum_{k=1}^{K} 2\rho_k\lambda\langle p^*, Y_k - p\rangle \leq \sum_{k=1}^{K} \Big[&\Big(\|X_k - p\|^2 + (1-\alpha_k)\Big(\tfrac{3}{2\rho_k(1+L\lambda)} - 1\Big)\|X_k \\
&-X_{k-1}\|^2 - \alpha_k\|X_{k-1} - p\|^2\Big) \\
&-\Big(\|X_{k+1} - p\|^2 + (1-\alpha_{k+1})\Big(\tfrac{3}{2\rho_{k+1}(1+L\lambda)} - 1\Big)\|X_{k+1} - X_k\|^2 - \alpha_{k+1}\|X_k - p\|^2\Big)\Big] \\
&+ \sum_{k=1}^{K}\Delta M_k + \sum_{k=1}^{K}\Delta N_k(p) \\
\leq \|X_1 - p\|^2 + (1-\alpha_1)&\Big(\tfrac{3}{2\rho_1(1+L\lambda)} - 1\Big)\|X_1 - X_0\|^2 - \alpha_1\|X_0 - p\|^2 \\
&+ \sum_{k=1}^{K}\Delta M_k + \sum_{k=1}^{K}\Delta N_k(p) \\
= (1-\alpha_1)\|X_1 - p\|^2 &+ \sum_{k=1}^{K}\Delta M_k + \sum_{k=1}^{K}\Delta N_k(p),
\end{aligned}
$$

where we notice $X_1 = X_0$ in the last inequality. □

Next, we derive a rate statement in terms of the gap function, using the averaged sequence

$$\bar{X}_K \triangleq \frac{\sum_{k=1}^{K}\rho_k Y_k}{\sum_{k=1}^{K}\rho_k}. \tag{73}$$

Theorem 15 (Rate and oracle complexity under monotonicity of V) *Consider the sequence $(X_k)_{k\in\mathbb{N}}$ generated RISFBF. Suppose Assumptions 1-5 hold. Suppose $m_k \triangleq \lfloor k^a \rfloor$ and $\lambda_k = \lambda \in (0, 1/(2L))$ for every $k \in \mathbb{N}$ where $a > 1$. Suppose $(\alpha_k)_{k\in\mathbb{N}}$ is a non-decreasing sequence such that $0 < \alpha_k \leq \bar{\alpha} < 1$, $\rho_k = \frac{3(1-\bar{\alpha})^2}{2(2\alpha_k^2 - \alpha_k + 1)(1+L\lambda)}$ for every $k \in \mathbb{N}$. Then the following hold for any $K \in \mathbb{N}$:*

(i) $\mathbb{E}[\mathsf{Gap}(\bar{X}_K|\mathsf{C})] \leq \mathcal{O}\Big(\frac{1}{K}\Big)$.

(ii) *Given* $\varepsilon > 0$, *define* $K_\varepsilon \triangleq \{k \in \mathbb{N} | \mathbb{E}[\mathsf{Gap}(\bar{X}_k|\mathsf{C})] \leq \varepsilon\}$, *then* $\sum_{k=1}^{K_\varepsilon} m_k \leq \mathcal{O}\Big(\frac{1}{\varepsilon^{1+a}}\Big)$.

The proof of this Theorem builds on an idea which is frequently used in the analysis of stochastic approximation algorithms, and can at least be traced back to the robust stochastic approximation approach of [49]. In order to bound the expectation of the gap function, we construct an auxiliary process which allows us to majorize the gap via a quantity which is independent of the reference points. Once this is achieved, a simple variance bound completes the result.

Proof of Theorem 15 We define an auxiliary process $(\Psi_k)_{k\in\mathbb{N}}$ such that

$$\Psi_{k+1} \triangleq \Psi_k + \rho_k \lambda_k W_k, \quad \Psi_1 \in \operatorname{dom}(T). \tag{74}$$

Then,

$$\|\Psi_{k+1} - p\|^2 = \|(\Psi_k - p) + \rho_k \lambda_k W_k\|^2 = \|\Psi_k - p\|^2 + \rho_k^2 \lambda_k^2 \|W_k\|^2 + 2\rho_k \lambda_k \langle \Psi_k - p, W_k\rangle,$$

so that

$$2\rho_k \lambda_k \langle W_k, p - \Psi_k\rangle = \|\Psi_k - p\|^2 - \|\Psi_{k+1} - p\|^2 + \rho_k^2 \lambda_k^2 \|W_k\|^2.$$

Introducing the iterate Y_k, the above implies

$$\begin{aligned} 2\rho_k \lambda_k \langle W_k, p - Y_k\rangle &= 2\rho_k \lambda_k \langle W_k, p - \Psi_k\rangle + 2\rho_k \lambda_k \langle W_k, \Psi_k - Y_k\rangle \\ &= \|\Psi_k - p\|^2 - \|\Psi_{k+1} - p\|^2 + \rho_k^2 \lambda_k^2 \|W_k\|^2 + 2\rho_k \lambda_k \langle W_k, \Psi_k - Y_k\rangle. \end{aligned}$$

As $\Delta N_k(p) = 2\rho_k \lambda_k \langle W_k, p - Y_k\rangle$, this implies via a telescopian sum argument

$$\sum_{k=1}^{K} \Delta N_k(p) \le \|\Psi_1 - p\|^2 + \sum_{k=1}^{K} \rho_k^2 \lambda_k^2 \|W_k\|^2 + \sum_{k=1}^{K} 2\rho_k \lambda_k \langle W_k, \Psi_k - Y_k\rangle. \tag{75}$$

Using Lemma 14 and setting $\lambda_k \equiv \lambda$, for any $(p, p^*) \in \operatorname{gr}(F)$ it holds true that

$$\sum_{k=1}^{K} 2\rho_k \lambda \langle p^*, Y_k - p\rangle \le (1 - \alpha_1)\|X_1 - p\|^2 + \sum_{k=1}^{K} \Delta M_k + \sum_{k=1}^{K} \Delta N_k(p).$$

Define $\mathsf{c}_1 \triangleq (1 - \alpha_1)\|X_1 - p\|^2$, divide both sides by $\sum_{k=1}^{K} \rho_k$ and using our definition of an ergodic average (73), this gives

$$2\lambda \left\langle p^*, \bar{X}_K - p \right\rangle \le \frac{1}{\sum_{k=1}^{K} \rho_k} \left\{ \mathsf{c}_1 + \sum_{k=1}^{K} \Delta M_k + \sum_{k=1}^{K} \Delta N_k(p) \right\}.$$

Using the bound established in eq. (75), it follows

$$\begin{aligned} 2\lambda \left\langle p^*, \bar{X}_K - p \right\rangle \le \frac{1}{\sum_{k=1}^{K} \rho_k} \Bigg\{ & \mathsf{c}_1 + \sum_{k=1}^{K} \Delta M_k + \|\Psi_1 - p\|^2 \\ & + \sum_{k=1}^{K} \rho_k^2 \lambda^2 \|W_k\|^2 + \sum_{k=1}^{K} 2\rho_k \lambda \langle W_k, \Psi_k - Y_k\rangle \Bigg\}. \end{aligned}$$

Choosing $\Psi_1, p \in \mathsf{C}$ and introducing $\mathsf{c}_2 \triangleq \mathsf{c}_1 + 4D^2$, we see that the above can be bounded by a random quantity which is independent of p:

$$2\lambda \left\langle p^*, \bar{X}_K - p \right\rangle \le \frac{1}{\sum_{k=1}^{K} \rho_k} \left\{ \mathsf{c}_2 + \sum_{k=1}^{K} \Delta M_k + \sum_{k=1}^{K} \rho_k^2 \lambda^2 \|W_k\|^2 + \sum_{k=1}^{K} 2\rho_k \lambda_k \langle W_k, \Psi_k - Y_k\rangle \right\}.$$

Taking the supremum over pairs (p, p^*) such that $p \in \mathcal{C}$ and $p^* \in F(y)$, it follows

$$2\lambda\mathsf{Gap}(\bar{X}_K|\mathsf{C}) \leq \frac{\mathsf{c}_2}{\sum_{k=1}^K \rho_k} + \frac{\sum_{k=1}^K \Delta M_k + \sum_{k=1}^K \rho_k^2\lambda^2\|W_k\|^2 + \sum_{k=1}^K 2\rho_k\lambda_k\langle W_k, \Psi_k - Y_k\rangle}{\sum_{k=1}^K \rho_k} \tag{76}$$

In order to proceed, we bound the first moment of the process ΔM_k in the same way as in (34), in order to get

$$\begin{aligned}\mathbb{E}[\Delta M_k|\mathcal{F}_k] &\leq \frac{6\rho_k\lambda_k^2}{1+L\lambda_k}\mathbb{E}[\|W_k\|^2|\mathcal{F}_k] + \lambda_k^2\left(\frac{6\rho_k}{1+L\lambda_k} + \frac{\rho_k\lambda_k^2}{2}\right)\mathbb{E}[\|U_k\|^2|\mathcal{F}_k] \\ &= \frac{\left(\frac{12\rho_k\lambda_k^2}{1+L\lambda_k}\sigma^2 + \frac{\rho_k\lambda_k^2}{2}\sigma^2\right)}{m_k} \triangleq \frac{\mathsf{a}_k\sigma^2}{m_k}.\end{aligned}$$

Next, we take expectations on both sides of inequality (76), and use the bound (17), and $\mathbb{E}[\langle W_k, \Psi_k - Y_k\rangle] = \mathbb{E}\big[\mathbb{E}\big(\langle W_k, \Psi_k - Y_k\rangle|\hat{\mathcal{F}}_k\big)\big] = 0$. This yields

$$2\lambda\mathbb{E}\big[\mathsf{Gap}(\bar{X}_K|\mathsf{C})\big] \leq \frac{\mathsf{c}_2}{\sum_{k=1}^K \rho_k} + \frac{1}{\sum_{k=1}^K \rho_k}\left(\sum_{k=1}^K \frac{\mathsf{a}_k\sigma^2}{m_k} + \sum_{k=1}^K \rho_k^2\lambda^2\frac{\sigma^2}{m_k}\right).$$

Since $\alpha_k \uparrow \bar{\alpha} \in (0,1)$, we know that $\rho_k \geq \tilde{\rho} \triangleq \frac{3(1-\bar{\alpha}^2)}{2(1+L\lambda)(2\bar{\alpha}^2+1)}$. Similarly, since $2\alpha_k^2 - \alpha_k + 1 \geq 7/8$ for all k, it follows $\rho_k \leq \bar{\rho} \triangleq \frac{12(1-\bar{\alpha})^2}{7}$. Using this upper and lower bound on the relaxation sequence, we also see that $\mathsf{a}_k \leq \lambda^2\left(\frac{12\bar{\rho}}{1+L\lambda} + \frac{\bar{\rho}}{2}\right) \equiv \bar{\mathsf{a}}$, so that

$$2\lambda\mathbb{E}\big[\mathsf{Gap}(\bar{X}_K|\mathsf{C})\big] \leq \frac{\mathsf{c}_2}{\tilde{\rho}K} + \frac{1}{\tilde{\rho}K}\left(\bar{\mathsf{a}}\sigma^2 + \bar{\rho}^2\lambda^2\sigma^2\right)\sum_{k=1}^K \frac{1}{m_k} \leq \frac{\mathsf{c}_3}{K}$$

where $\mathsf{c}_3 \triangleq \frac{\mathsf{c}_2}{\tilde{\rho}} + \frac{1}{\tilde{\rho}}\left(\bar{\mathsf{a}}\sigma^2 + \bar{\rho}\lambda^2\sigma^2\right)\sum_{k=1}^\infty \frac{1}{m_k}$. Hence, defining the deterministic stopping time $K_\varepsilon = \{k \in \mathbb{N}|\mathbb{E}[\mathsf{Gap}(\bar{X}_k|\mathsf{C})] \leq \varepsilon\}$, we see $K_\varepsilon \geq \frac{\mathsf{c}_3}{2\lambda\varepsilon} = \frac{\mathsf{c}_4}{\varepsilon}$.

(ii). Suppose $m_k = \lfloor k^a \rfloor$, for $a > 1$. Then the oracle complexity to compute an $\bar{X}_K$ such that $\mathbb{E}[\mathsf{Gap}(\bar{X}_k|\mathsf{C})] \leq \epsilon$ is bounded as

$$\sum_{k=1}^K m_k \leq \sum_{k=1}^{\lceil(\mathsf{c}_4/\varepsilon)\rceil} m_k \leq \sum_{k=1}^{\lceil(\mathsf{c}_4/\varepsilon)\rceil} k^a \leq \int_{k=1}^{(\mathsf{c}_4/\varepsilon)+1} x^a dx \leq \frac{((\mathsf{c}_4/\varepsilon)+1)^{a+1}}{a+1} \leq \left(\frac{\mathsf{c}_4}{\varepsilon^{a+1}}\right).$$

□

Remark 7 In the prior result, we employ a sampling rate $m_k = \lfloor k^a \rfloor$ where $a > 1$. This achieves the optimal rate of convergence. In contrast, the authors in [32] employ a sampling rate, loosely given by $m_k = \lfloor k^{1+a}(\ln(k))^{1+b} \rfloor$ where $a > 0, b \geq -1$ or $a = 0, b > 0$. We observe that when $a > 0$ and $b \geq -1$, the mini-batch size grows faster than our proposed m_k while it is comparable in the other case.

5 Applications

In this section, we compare the proposed scheme with its SA counterparts on a class of monotone two-stage stochastic variational inequality problems (Sec. 5.1) and a supervised learning problem (Sec. 5.2) and discuss the resulting performance.

5.1 Two-stage stochastic variational inequality problems

In this section, we describe some preliminary computational results obtained from Algorithm 1 when applied to a class of two-stage stochastic variational inequality problems, recently introduced by [85].

Consider an imperfectly competitive market with N firms playing a two-stage game. In the first stage, the firms decide upon their capacity level $x_i \in [l_i, u_i]$, anticipating the expected revenues to be obtained in the second stage in which they compete by choosing quantities à la Cournot. The second-stage market is characterized by uncertainty as the per-unit cost $h_i(\xi_i)$ is realized on the spot and cannot be anticipated. To compute an equilibrium in this game, we assume that each player is able to take stochastic recourse by determining production levels $y_i(\xi)$, contingent on random convex costs and capacity levels x_i. In order to bring this into the terminology for our problem, let use define the feasible set for capacity decisions of firm i as $\mathcal{X}_i \triangleq [l_i, u_i] \subset \mathbb{R}_+$. The joint profile of capacity decisions is denoted by an N-tuple $x = (x_1, \ldots, x_N) \in \mathcal{X} \triangleq \prod_{i=1}^N \mathcal{X}_i = \mathcal{X}$. The capacity choice of player i is then determined as a solution to the parametrized problem (Play$_i(x_{-i})$)

$$\min_{x_i \in \mathcal{X}_i} c_i(x_i) - \big(p(X)x_i - \mathbb{E}_\xi[\mathcal{Q}_i(x_i, \xi)]\big), \qquad (\text{ Play}_i(x_{-i}))$$

where $c_i : \mathcal{X}_i \to \mathbb{R}_+$ is a $\tilde{L}_i^c$-smooth and convex cost function and $p(\cdot)$ denotes the inverse-demand function defined as $p(X) = d - rX,\ d, r > 0$. The function $\mathcal{Q}_i(\cdot, \xi)$ denotes the optimal cost function of firm i in scenario $\xi \in \Xi$, assuming a value $\mathcal{Q}_i(x_i, \xi)$ when the capacity level x_i is chosen. The recourse function $\mathbb{E}_\xi[\mathcal{Q}_i(\cdot, \xi)]$ denotes the expectation of the optimal value of the player i's second stage problem and is defined as

$$\begin{aligned} \mathcal{Q}_i(x_i, \xi) &\triangleq \min\{h_i(\xi)y_i(\xi) | y_i(\xi) \in [0, x_i]\} \\ &= \max\{\pi_i(\xi)x_i | \pi_i(\xi) \le 0, h_i(\xi) - \pi_i(\xi) \ge 0\}. \qquad (\text{ Rec}_i(x_{-i})) \end{aligned}$$

A Nash equilibrium of this game is given by a tuple $(x_1^*, \cdots, x_N^*)$ where x_i^* solves (Play$_i(x_{-i}^*)$) for each $i = 1, 2, \ldots, N$. A simple computation shows that $\mathcal{Q}_i(x_i, \xi) = \min\{0, h_i(\xi)x_i\}$, and hence it is nonsmooth. In order to obtain a smoothed variant, we introduce $\mathcal{Q}_i^\epsilon(\cdot, \xi_i)$, defined as

$$\mathcal{Q}_i^\epsilon(x_i, \xi) \triangleq \max\{x_i\pi_i(\xi) - \tfrac{\epsilon}{2}(\pi_i(\xi))^2 | \pi_i(\xi) \le 0, \pi_i(\xi) \le h_i(\xi)\}, \quad \epsilon > 0.$$

This is the value function of a quadratic program, requiring the maximization of an ϵ-strongly concave function. Hence, $\mathcal{Q}_i^\epsilon(x_i, \xi)$ is single-valued and $\nabla_{x_i} \mathcal{Q}_i^\epsilon(\cdot, \xi)$ is

$\frac{1}{\epsilon}$-Lipschitz and ϵ-strongly monotone [86, Prop.12.60] for all $\xi \in \Xi$. The latter is explicitly given by

$$\nabla_{x_i} \mathcal{Q}_i^{\epsilon}(x_i, \xi) \triangleq \operatorname{argmax}\{x_i \pi_i(\xi) - \tfrac{\epsilon}{2}(\pi_i(\xi))^2 | \pi_i(\xi) \le 0, \pi_i(\xi) \le h_i(\xi)\}.$$

Employing this smoothing strategy in our two-stage noncooperative game yields the individual decision problem

$$(\forall i \in \{1, \dots, N\}) : \min_{x_i \in \mathcal{X}_i} \; c_i(x_i) - p(X)x_i + \mathbb{E}_{\xi}[\mathcal{Q}_i^{\epsilon}(x_i, \xi)]. \tag{$\text{Play}_i^{\epsilon}(x_{-i})$}$$

The necessary and sufficient equilibrium conditions of this ϵ-smoothed game can be compactly represented as

$$\begin{aligned} &0 \in F^{\epsilon}(x) \triangleq V^{\epsilon}(x) + T(x), \text{ where} \\ &V^{\epsilon}(x) = C(x) + R(x) + D^{\epsilon}(x), \text{ and } T(x) = \mathsf{N}_{\mathcal{X}}(x), \end{aligned} \tag{SGE$^{\epsilon}$}$$

and C, R, and D^{ϵ} are single-valued maps given by

$$C(x) \triangleq \begin{pmatrix} c_1'(x_1) \\ \vdots \\ c_N'(x_N) \end{pmatrix}, \quad R(x) \triangleq r(X\mathbf{1} + x) - d, \text{ and } D^{\epsilon}(x) \triangleq \begin{pmatrix} \mathbb{E}_{\xi}[\nabla_{x_1} \mathcal{Q}_1^{\epsilon}(x_1, \xi)] \\ \vdots \\ \mathbb{E}_{\xi}[\nabla_{x_N} \mathcal{Q}_N^{\epsilon}(x_N, \xi)] \end{pmatrix}.$$

We note that the interchange between the expectation and the gradient operator can be invoked based on smoothness requirements (cf. [87, Th. 7.47]). The problem (SGE$^{\epsilon}$) aligns perfectly with the structured inclusion (MI), in which T is a maximal monotone map and V is an expectation-valued maximally monotone map. In addition, we can quantify the Lipschitz constant of V as $L_V = L_C + L_R + L_D^{\epsilon}$,where $L_C = \max_{1 \le i \le N} \tilde{L}_i^c$, $L_R = r\| \operatorname{Id} + \mathbf{1}\mathbf{1}^{\top} \|_2 = r(N+1)$ and $L_D^{\epsilon} = \frac{1}{\epsilon}$. Here, Id is the $N \times N$ identity matrix, and $\mathbf{1}$ is the $N \times 1$ vector consisting only of ones.

Problem parameters for 2-stage SVI. Our numerics are based on specifying $N = 10$, $r = 0.1$, and $d = 1$. We consider four problem settings of L_V ranging from $10, \cdots, 10^4$ (See Table 1). For each setting, the problem parameters are defined as follows.

(i) *Specification of $h_i(\xi)$.* The cost parameters $h_i(\xi_i) \triangleq \xi_i$ where $\xi_i \sim \texttt{Uniform}[-5, 0]$ and $i = 1, \cdots, N$.

(ii) *Specification of $L_V, L_R, L_D^{\epsilon}, L_C$, and $\hat{b}_1$.* Since $\| \operatorname{Id} + \mathbf{1}\mathbf{1}^{\top} \|_2 = 11$ when $N = 10$, $L_R = r\| \operatorname{Id} + \mathbf{1}\mathbf{1}^{\top} \| = 1.1$. Let ϵ be defined as $\epsilon = \frac{10}{L_V}$ and $L_D^{\epsilon} = \frac{1}{\epsilon} = \frac{L_V}{10}$. It follows that $L_C = L_V - L_R - L_D^{\epsilon}$ and $\hat{b}_1 = L_C$.

(iii) *Specification of $c_i(x_i)$.* The cost function c_i is defined as $c_i(x_i) = \frac{1}{2}\hat{b}_i x_i^2 + a_i x_i$ where $a_1, \dots, a_N \sim \texttt{Uniform}[2, 3]$ and $\hat{b}_2, \cdots, \hat{b}_N \sim \texttt{Uniform}[0, \hat{b}_1]$. Further, $a \triangleq [a_1, \dots, a_N]^{\top} \in \mathbb{R}^N$ and $B \triangleq \operatorname{diag}(\hat{b}_1, \dots, \hat{b}_N)$ is a diagonal matrix with nonnegative elements.

Algorithm specifications We compare Algorithm 1 (RISFBF) with a stochastic forward-backward (SFB) scheme and a stochastic forward-backward-forward

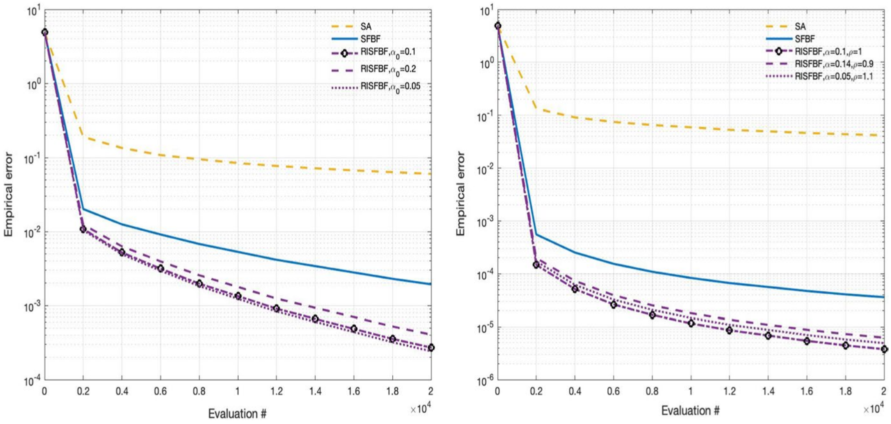

Fig. 1 Trajectories for (SFB), (SFBF), and (RISFBF) (left: monotone, right: s-monotone)

(SFBF) scheme. Solution quality is compared by estimating the residual function $\mathsf{res}(x) = \|x - \Pi_{\mathcal{X}}(x - \lambda V^{\epsilon}(x))\|$. All of the schemes were implemented in `MATLAB` on a PC with 16GB RAM and 6-Core Intel Core i7 processor (2.6GHz).

(i) *(SFB)*: The (SFB) scheme is defined as the recursion

$$X_{k+1} := \Pi_{\mathcal{X}}\Big[X_k - \lambda_k \widehat{V}^{\epsilon}(X_k, \xi_k)\Big], \tag{SFB}$$

where $V^{\epsilon}(X_k) = \mathbb{E}_{\xi}[\widehat{V}^{\epsilon}(X_k, \xi)]$ and $\lambda_k = \frac{1}{\sqrt{k}}$. The operator $\Pi_{\mathcal{X}}[\cdot]$ means the orthogonal projection onto the set $\mathcal{X}$. Note that x_0 is randomly generated in $[0, 1]^N$.

(ii) *(SFBF)*: The Variance-reduced stochastic modified forward-backward scheme we employ is defined by the updates

$$\begin{cases} Y_k = \Pi_{\mathcal{X}}[X_k - \lambda_k A_k(X_k)], \\ X_{k+1} = Y_k - \lambda_k(B_k(Y_k) - A_k(X_k)). \end{cases} \tag{SFBF}$$

where $A_k(X_k) = \frac{1}{m_k}\sum_{t=1}^{m_k} \widehat{V}^{\epsilon}(X_k, \xi_k)$, $B_k(Y_k) = \frac{1}{m_k}\sum_{t=1}^{m_k} \widehat{V}^{\epsilon}(Y_k, \eta_k)$. We choose a constant $\lambda_k \equiv \lambda = \frac{1}{4L_V}$. We assume $m_k = \lfloor k^{1.01} \rfloor$ for merely monotone problems and $m_k = \lfloor 1.01^k \rfloor$ for strongly monotone problems.

(iii) *(RISFBF)*: In the implementation of Algorithm 1 we choose a constant steplength $\lambda_k \equiv \lambda = \frac{1}{4L_V}$. In merely monotone settings, we utilize an increasing sequence $\alpha_k = \alpha_0(1 - \frac{1}{k+1})$, where $\alpha_0 = 0.1$, the relaxation parameter sequence ρ_k defined as $\rho_k = \frac{3(1-\alpha_0)^2}{2(2\alpha_k^2 - \alpha_k + 1)(1 + L_V\lambda)}$, and $m_k = \lfloor k^{1.01} \rfloor$. In strongly monotone regimes, we choose a constant inertial parameter $\alpha_k \equiv \alpha = 0.1$, a constant relaxation parameter $\rho_k \equiv \rho = 1$, and $m_k = \lfloor 1.01^k \rfloor$.

Table 1 Comparison of (RISFBF) with (SFB) and (SFBF) under various Lipschitz constant

merely monotone, 20000 evaluations									
L_V	RISFBF			SFBF			SFB		
	error	time	CI	error	time	CI	error	time	CI
1e1	2.2e-4	2.7	[2.0e-4,2.5e-4]	1.6e-3	2.6	[1.3e-3,1.8e-3]	5.3e-2	2.7	[5.0e-2,5.7e-2]
1e2	2.7e-4	2.7	[2.5e-4,3.0e-4]	1.9e-3	2.6	[1.6e-3,2.1e-3]	6.1e-2	2.7	[5.8e-2,6.4e-2]
1e3	6.9e-4	2.7	[6.7e-3,7.1e-4]	2.2e-3	2.6	[2.0e-3,2.5e-3]	7.6e-2	2.5	[7.3e-2,7.9e-2]
1e4	2.7e-3	2.7	[2.5e-3,3.0e-3]	5.9e-3	2.6	[5.4e-3,6.2e-3]	9.4e-2	2.6	[9.0e-1,9.7e-1]
strongly monotone, 20000 evaluations									
L_V	RISFBF			SFBF			SFB		
	error	time	CI	error	time	CI	error	time	CI
1e1	1.5e-6	2.6	[1.3e-6,1.7e-6]	1.5e-5	2.6	[1.2e-5,1.7e-5]	2.9e-2	2.5	[2.7e-2,3.1e-2]
1e2	3.7e-6	2.6	[3.5e-6,3.9e-6]	3.6e-5	2.5	[3.3e-5,3.9e-5]	4.1e-2	2.5	[3.8e-2,4.4e-2]
1e3	4.5e-6	2.6	[4.3e-6,4.7e-6]	5.6e-5	2.5	[4.2e-6,4.7e-6]	5.5e-2	2.4	[5.2e-2,5.7e-2]
1e4	1.4e-5	2.6	[1.1e-5,1.7e-5]	7.4e-5	2.5	[7.1e-5,7.7e-5]	6.0e-2	2.5	[5.7e-2,6.3e-2]

In Fig. 1, we compare the three schemes under maximal monotonicity and strong monotonicity, respectively and examine their sensitivities to inertial and relaxation parameters. Both sets of plots are based on selecting $L_V = 10^2$.

Key insights Several insights may be drawn from Table 1 and Figure 1.

(a) First, from Table 1, one may conclude that on this class of problems, (RISFBF) and (SFBF) significantly outperform (SFB) schemes, which is less surprising given that both schemes employ an increasing mini-batch sizes, leading to performance akin to that seen in deterministic schemes. We should note that when $\mathcal{X}$ is somewhat more complicated, the difference in run-times between SA schemes and mini-batch variants becomes more pronounced; in this instance, the set $\mathcal{X}$ is relatively simple to project onto and there is little difference in run-time across the three schemes.

(b) Second, we observe that while both (SFBF) and (RISFBF) schemes can contend with poorly conditioned problems, as seen by noting that as L_V grows, their performance does not degenerate significantly in terms of empirical error; However, in both monotone and strongly monotone regimes, (RISFBF) provides consistently better solutions in terms of empirical error over (SFBF). Figure 1 displays the range of trajectories obtained for differing relaxation and inertial parameters and in the instances considered, (RISFBF) shows consistent benefits over (SFBF).

(c) Third, since such schemes display geometric rates of convergence for strongly monotone inclusion problems, this improvement is reflected in terms of the empirical errors for strongly monotone vs monotone regimes.

5.2 Supervised learning with group variable selection

Our second numerical example considers the following population risk formulation of a composite absolute penalty (CAP) problem arising in supervised statistical learning [7]

$$\min_{w \in \mathcal{W}} \tfrac{1}{2}\mathbb{E}_{(a,b)}[(a^\top w - b)^2] + \eta \sum_{g \in \mathcal{S}} \|w_g\|_2, \tag{CAP}$$

where the feasible set $\mathcal{W} \subseteq \mathbb{R}^d$ is a Euclidean ball with $\mathcal{W} \triangleq \{w \in \mathbb{R}^d \mid \|w\|_2 \leq D\}$, $\xi = (a, b) \in \mathbb{R}^d \times \mathbb{R}$ denotes the random variable consisting of a set of predictors a and output b. The parameter vector w is the sparse linear hypothesis to be learned. The sparsity structure of w is represented by group $\mathcal{S} \in 2^{\{1,\dots,l\}}$. When the groups in $\mathcal{S}$ do not overlap, $\sum_{g \in \mathcal{S}} \|w_g\|_2$ is referred to as the group lasso penalty [6, 88]. When the groups in $\mathcal{S}$ form a partition of the set of predictors, then $\sum_{g \in \mathcal{S}} \|w_g\|_2$ is a norm afflicted by singularities when some components w_g are equal to zero. For any $g \in \{1, \cdots, l\}$, w_g is a sparse vector constructed by components of x whose indices are in g, i.e., $w_g := (w_i)_{i \in g}$ with few non-zero components in w_g. Here, we assume that each group $g \in \mathcal{S}$ consists of k elements. Introduce the linear operator $L : \mathbb{R}^d \to \mathbb{R}^k \underbrace{\times \cdots \times}_{l-\text{times}} \mathbb{R}^k$, given by $Lw = [\eta w_{g_1}, \dots, \eta w_{g_l}]$. Let us also define

$$Q = \mathbb{E}_\xi[aa^\top], q = \mathbb{E}_\xi[ab], c = \frac{1}{2}\mathbb{E}_\xi[b^2],$$
$$h(w) \triangleq \frac{1}{2}w^\top Qw - w^\top q + c, \text{ and } f(w) \triangleq \delta_{\mathcal{W}}(w),$$

where $\delta_{\mathcal{W}}(\cdot)$ denotes the indicator function with respect to the set $\mathcal{W}$. Then (CAP) becomes

$$\min_{w \in \mathbb{R}^d} \{h(w) + g(Lw) + f(w)\}, \quad \text{where } g(y_1, \dots, y_l) \triangleq \sum_{i=1}^{l} \|y_i\|.$$

This is clearly seen to be a special instance of the convex programming problem (2). Specifically, we let $\mathsf{H}_1 = \mathbb{R}^d$ with the standard Euclidean norm, and $\mathsf{H}_2 = \mathbb{R}^k \underbrace{\times \cdots \times}_{l-\text{times}} \mathbb{R}^k$ with the product norm

$$\|(y_1, \dots, y_l)\|_{\mathsf{H}_2} \triangleq \sum_{i=1}^{l} \|y_i\|_2.$$

Since

$$g^*(v_1, \dots, v_l) = \sum_{i=1}^{l} \delta_{\mathbb{B}(0,1)}(v_i) \qquad \forall v = (v_1, \dots, v_l) \in \mathbb{R}^k \underbrace{\times \cdots \times}_{l-\text{times}} \mathbb{R}^k,$$

the Fenchel-dual takes the form (3). Accordingly, a primal-dual pair for (CAP) is a root of the monotone inclusion (MI) with

Table 2 The comparison of the RISFBF, SFBF and SEG algorithms in solving (CAP)

Iteration	RISFBF		SFBF		SEG	
N	Rel. error	CPU	Rel. error	CPU	Rel. error	CPU
v400	5.4e-1	0.1	34.6	0.1	34.7	0.1
v800	8.1e-3	0.5	1.1e-1	0.5	1.5e-1	0.5
1200	6.0e-3	1.1	2.4e-2	1.1	2.4e-2	1.1
1600	5.2e-3	2.0	2.0e-2	2.0	1.9e-2	2.0
2000	4.6e-3	3.1	1.6e-2	3.1	1.5e-2	3.1

The relative error and CPU time in the table is the average results of 20 runs

$$V(w,v) = (\nabla h(w) + L^* v, -Lw) \text{ and } T(w,v) \triangleq \partial f(w) \times \partial g^*(v)$$

involving $d + kl$ variables.

Problem parameters for (CAP) We simulated data with $d = 82$, covered by 10 groups of 10 variables with 2 variables of overlap between two successive groups: $\{1,\dots,10\}, \{9,\dots,18\}, \dots, \{73,\dots,82\}$. We assume the nonzeros of w_{true} lie in the union of groups 4 and 5 and sampled from i.i.d. Gaussian variables. The operator $V(w, v)$ is estimated by the mini-batch estimator using m_k iid copies of the random input-output pair $\xi = (a, b) \in \mathbb{R}^d \times \mathbb{R}$. Specifically, we draw each coordinate of the random vector a from the standard Gaussian distribution $\mathtt{N}(0, 1)$ and generate $b = a^\top w_{\text{true}} + \varepsilon$, for $\varepsilon \sim \mathtt{N}(0, \sigma_\varepsilon^2)$. In the concrete experiment reported here, the error variance is taken as $\sigma_\varepsilon = 0.1$. In all instances, the regularization parameter is chosen as $\eta = 10^{-4}$. The accuracy of feature extraction of algorithm output w is evaluated by the relative error to the ground truth, defined as

$$\frac{\|w - w_{\text{true}}\|_2}{\|w_{\text{true}}\|_2}.$$

Algorithm specifications We compare (RISFBF) with stochastic extragradient (SEG) and stochastic forward-backward-forward (SFBF) schemes and specify their algorithm parameters. Again, all the schemes are run on MATLAB 2018b on a PC with 16GB RAM and 6-Core Intel Core i7 processor (2.6×8GHz).

(i) *(SEG)*: Set $\mathcal{X} \triangleq \mathcal{W} \times \operatorname{dom}(g^*)$. The (SEG) scheme [32] utilizes the updates

$$\begin{aligned} Y_k &:= \Pi_{\mathcal{X}}\big[X_k - \lambda_k A_k(X_k)\big], \\ X_{k+1} &:= \Pi_{\mathcal{X}}\big[X_k - \lambda_k B_k(Y_k)\big], \end{aligned} \tag{SEG}$$

where $A_k(X_k) = \frac{1}{m_k} \sum_{t=1}^{m_k} V(X_k, \xi_k)$, $B_k(Y_k) = \frac{1}{m_k} \sum_{t=1}^{m_k} V(Y_k, \eta_k)$. In this scheme, $\lambda_k \equiv \lambda$ is chosen to be $\frac{1}{4L_V}$ (L_V is the Lipschitz constant of V). We assume $m_k = \left\lfloor \frac{k^{1.1}}{n} \right\rfloor$.

(ii) *(SFBF)*: We employ the algorithm parameters employed in (i). Specifically, we choose a constant $\lambda_k \equiv \lambda = \frac{1}{4L_V}$ and $m_k = \left\lfloor \frac{k^{1.1}}{n} \right\rfloor$.

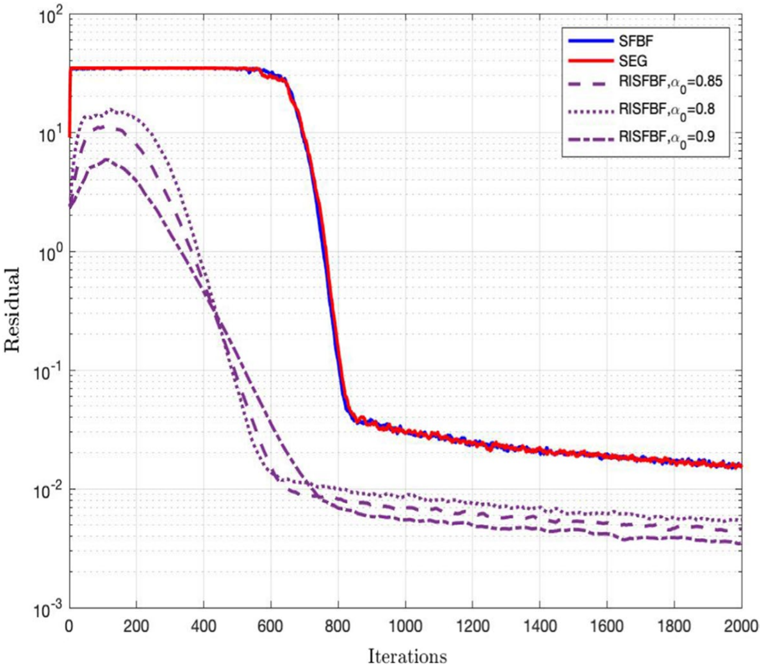

Fig. 2 Trajectories for (SEG), (SFBF), and (RISFBF) for problem (CAP)

(iii) *(RISFBF)*: Here, we employ a constant step-length $\lambda_k \equiv \lambda = \frac{1}{4L_V}$, an increasing sequence $\alpha_k = \alpha_0(1 - \frac{1}{k+1})$, where $\alpha_0 = 0.85$, a relaxation parameter sequence $\rho_k = \frac{3(1-\alpha_0)^2}{2(2\alpha_k^2-\alpha_k+1)(1+L_V\lambda)}$, and assume $m_k = \left\lfloor \frac{k^{1.1}}{n} \right\rfloor$.

Insights We compare the performance of the schemes in Table 2 and observe that (RISFBF) outperforms its competitors others in extracting the underlying feature of the datasets. In Fig. 2, trajectories for (RISFBF), (SFBF) and (SEG) are presented where a consistent benefit of employing (RISFBF) can be seen for a range of choices of α_0.

6 Conclusion

In a general structured monotone inclusion setting in Hilbert spaces, we introduce a relaxed inertial stochastic algorithm based on Tseng's forward-backward-forward splitting method. Motivated by the gaps in convergence claims and rate statements in both deterministic and stochastic regimes, we develop a variance-reduced framework and make the following contributions: (i) Asymptotic convergence guarantees are provided under both increasing and constant mini-batch sizes, the latter requiring somewhat stronger assumptions on V; (ii) When V is monotone, rate statements provided in terms of a restricted gap function, inspired by the Fitzpatrick function for inclusions, show that the expected gap of an averaged sequence diminishes at the rate of $\mathcal{O}(1/k)$ and oracle complexity of computing an ϵ-solution is $\mathcal{O}(1/\epsilon^{1+a})$ where $a > 1$; (iii) When V is strongly monotone, a non-asymptotic linear rate statement can be proven with an oracle complexity of $\mathcal{O}(\log(1/\epsilon))$ of computing an ϵ-solution. In addition, a perturbed linear rate is also developed. It is worth emphasizing that the rate statements in the strongly monotone regime accommodate the possibility of a biased stochastic oracle. Unfortunately, the growth rates in batch-size may be onerous in some situations, motivating the analysis of a polynomial growth rate in sample-size which is easily modulated. This leads to an associated polynomial rate of convergence.

Various open questions arise from our analysis. First, we exclusively focused on a variance reduction technique based on increasing mini-batches. From the point of view of computations and oracle complexity, this approach can become quite costly. Exploiting different variance reduction techniques, taking perhaps special structure of the single-valued operator V into account (as in [57]), has the potential of improving the computational complexity of our proposed method. At the same time, this will complicate the analysis of the variance of the stochastic estimators considerably and consequently, we leave this as an important question for future research.

Second, our analysis needs knowledge about the Lipschitz constant L. While in deterministic regimes, line search techniques have obviated such a need, such avenues are far more challenging to adopt in stochastic regimes. Efforts to address this in variational regimes have centered around leveraging empirical process theory [33]. This remains a goal of future research. Another avenue emerges in applications where we can gain a reasonably good estimate about this quantity via some pre-processing of the data (see e.g. Section 6 in [62]). Developing such an adaptive framework robust to noise is an important topic for future research.

Appendix

Appendix A Auxiliary results

Lemma 16 *For $x, y \in \mathsf{H}$ and scalars $\alpha, \beta \geq 0$ with $\alpha + \beta = 1$, it holds that*

$$\|\alpha x + \beta y\|^2 = \alpha\|x\|^2 + \beta\|y\|^2 - \alpha\beta\|x - y\|^2. \tag{A1}$$

We recall the Minkowski inequality: For $X, Y \in L^p(\Omega, \mathcal{F}, \mathbb{P}; \mathsf{H}), \mathcal{G} \subseteq \mathcal{F}$ and $p \in [1, \infty]$,

$$\mathbb{E}[\|X + Y\|^p|\mathcal{G}]^{1/p} \leq \mathbb{E}[\|X\|^p|\mathcal{G}]^{1/p} + \mathbb{E}[\|Y\|^p|\mathcal{G}]^{1/p}. \tag{A2}$$

In the convergence analysis, we use the Robbins-Siegmund Lemma [38, Lemma 11, pg. 50].

Lemma 17 (Robbins-Siegmund) *Let* $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$ *be a discrete stochastic basis. Let* $(v_n)_{n\geq 1}, (u_n)_{n\geq 1} \in \ell^0_+(\mathbb{F})$ *and* $(\theta_n)_{n\geq 1}, (\beta_n)_{n\geq 1} \in \ell^1_+(\mathbb{F})$ *be such that for all* $n \geq 0$,

$$\mathbb{E}[v_{n+1}|\mathcal{F}_n] \leq (1+\theta_n)v_n - u_n + \beta_n \qquad \mathbb{P} - \text{a.s.}\ .$$

Then $(v_n)_{n\geq 0}$ *converges a.s. to a random variable* v, *and* $(u_n)_{n\geq 1} \in \ell^1_+(\mathbb{F})$.

Lemma 18 *Let* $z \geq 0$ *and* $0 < q < p < 1$. *Then, if* $D \geq \frac{1}{\exp(1)\ln(p/q)}$, *it holds true that* $zq^z \leq Dp^z$ *for all* $z \geq 0$.

Proof We want to find a positive constant $D_{\min} > 0$ such that $D_{\min}\exp(z\ln(p)) = z\exp(z\ln(q))$ for all $z > 0$. Choosing D larger than this, gives a valid value. Rearranging, this is equivalent to $D = z\left(\frac{q}{p}\right)^z \geq 0$ for all $z \geq 0$, or, which is still equivalent to $\ln(D) - \ln(z) - z\ln(q/p) = 0$. Define the extended-valued function $f : [0,\infty) \to [-\infty, \infty]$ by $f(z) = \ln(D) - \ln(z) - \ln(q/p)$ if $z > 0$, and $f(z) = \infty$ if $z = 0$. Then, for all $z > 0$, simple calculus show $f'(z) = -1/z - \ln(q/p)$ and $f''(z) = 1/z^2$. Hence, $z \mapsto f(z)$ is a convex function with a unique minimum $z_{\min} = \frac{1}{\ln(p/q)} > 0$ and a corresponding function value $f(z_{\min}) = \ln(D) + \ln(\ln(p/q)) + 1$. Hence, for $D \geq D_{\min} = \frac{1}{\exp(1)\ln(p/q)}$, we see that $f(z_{\min}) > 0$, and thus $zq^z \leq Dp^z$ for all $z \geq 0$. □

Acknowledgements M. Staudigl acknowledges support from the COST Action CA16228 “European Network for Game Theory”.

Author Contributions All authors contributed equally to the completion of this manuscript.

Availability of data materials All data generated or analysed during this study are included in this published article.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article

References

1. Attouch, H., Cabot, A.: Convergence of a relaxed inertial forward-backward algorithm for structured monotone inclusions. Appl. Math. Optim. **80**(3), 547–598 (2019). https://doi.org/10.1007/s00245-019-09584-z
2. Attouch, H., Cabot, A.: Convergence of a relaxed inertial proximal algorithm for maximally monotone operators. Math. Program. **184**(1), 243–287 (2020). https://doi.org/10.1007/s10107-019-01412-0
3. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, (2016)
4. Facchinei, F., Pang, J.-S.: Finite-Dimensional Variational Inequalities and Complementarity Problems - Volume I and Volume II. Springer, (2003)
5. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**(1), 259–268 (1992). https://doi.org/10.1016/0167-2789(92)90242-F
6. Jacob, L., Obozinski, G., Vert, J.-P.: Group lasso with overlap and graph lasso. Proceedings of the 26th annual international conference on machine learning, pp. 433–440 (2009)
7. Zhao, P., Rocha, G., Yu, B.: The composite absolute penalties family for grouped and hierarchical variable selection. Ann. Stat. **37**(6A), 3468–3497 (2009). https://doi.org/10.1214/07-AOS584
8. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. J. Royal Statistical Soc.: Series B (Statistical Methodology) **67**(1), 91–108 (2005). https://doi.org/10.1111/j.1467-9868.2005.00490.x
9. Tibshirani, R.J., Taylor, J.: The solution path of the generalized lasso. Ann. Statist. **39**(3), 1335–1371 (2011). https://doi.org/10.1214/11-AOS878
10. Attouch, H., Briceno-Arias, L.M., Combettes, P.L.: A parallel splitting method for coupled monotone inclusions. SIAM J. Control. Optim. **48**(5), 3246–3270 (2010). https://doi.org/10.1137/090754297
11. Latafat, P., Freris, N.M., Patrinos, P.: A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization. IEEE Trans. Autom. Control **64**(10), 4050–4065 (2019). https://doi.org/10.1109/TAC.2019.2906924
12. Rockafellar, R.T.: Conjugate Duality and Optimization. Society for Industrial and Applied Mathematics (1974)
13. Combettes, P.L., Pesquet, J.-C.: Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators. Set-Valued and variational analysis **20**(2), 307–330 (2012)
14. Jiang, H., Xu, H.: Stochastic approximation approaches to the stochastic variational inequality problem. IEEE Trans. Autom. Control **53**(6), 1462–1475 (2008). https://doi.org/10.1109/TAC.2008.925853
15. Shanbhag, U.V.: Chapter 5. Stochastic Variational Inequality Problems: Applications, Analysis, and Algorithms, pp. 71–107 (2013). https://doi.org/10.1287/educ.2013.0120
16. Staudigl, M., Mertikopoulos, P.: Convergent noisy forward-backward-forward algorithms in non-monotone variational inequalities. IFAC-PapersOnLine **52**(3), 120–125 (2019)
17. Mertikopoulos, P., Staudigl, M.: Convergence to Nash Equilibrium in Continuous Games with Noisy First-order Feedback. In: 56th IEEE Conference on Decision and Control (2017)
18. Briceno-Arias, L.M., Combettes, P.L.: Monotone operator methods for Nash equilibria in non-potential games, pp. 143–159. Springer, ??? (2013)
19. Yi, P., Pavel, L.: An operator splitting approach for distributed generalized Nash equilibria computation. Automatica **102**, 111–121 (2019). https://doi.org/10.1016/j.automatica.2019.01.008
20. Franci, B., Staudigl, M., Grammatico, S.: Distributed forward-backward (half) forward algorithms for generalized nash equilibrium seeking. In: 2020 European Control Conference (ECC), pp. 1274–1279 (2020). https://doi.org/10.23919/ECC51009.2020.9143676
21. Friesz, T.L., Bernstein, D., Smith, T.E., Tobin, R.L., Wie, B.W.: Variational inequality formulation of the dynamic network user equilibrium. Oper. Res. **41**(1), 179–191 (1993)
22. Fukushima, M.: The primal Douglas-Rachford splitting algorithm for a class of monotone mappings with application to the traffic equilibrium problem. Math. Program. **72**(1), 1–15 (1996). https://doi.org/10.1007/BF02592328

23. Han, K., Eve, G., Friesz, T.L.: Computing dynamic user equilibria on large-scale networks with software implementation. Netw. Spat. Econ. **19**(3), 869–902 (2019). https://doi.org/10.1007/s11067-018-9433-y
24. Börgens, E., Kanzow, C.: ADMM-type methods for generalized Nash equilibrium problems in Hilbert spaces. SIAM J. Optim., 377–403 (2021). https://doi.org/10.1137/19M1284336
25. Tseng, P.: A modified forward-backward splitting method for maximal monotone mappings. SIAM J. Control. Optim. **38**(2), 431–446 (2000). https://doi.org/10.1137/S0363012998338806
26. Boţ, R.I., Mertikopoulos, P., Staudigl, M., Vuong, P.T.: Minibatch forward-backward-forward methods for solving stochastic variational inequalities. Stochastic Syst. (2021) https://doi.org/10.1287/stsy.2019.0064. https://doi.org/10.1287/stsy.2019.0064
27. Cui, S., Shanbhag, U.V.: On the computation of equilibria in monotone and potential stochastic hierarchical game. arXiv preprint arXiv:2104.07860 (2021)
28. Thong, D.V., Gibali, A., Staudigl, M., Vuong, P.T.: Computing dynamic user equilibrium on large-scale networks without knowing global parameters. Netw. Spat. Econ. **21**, 735–768 (2021)
29. Diakonikolas, J., Daskalakis, C., Jordan, M.: Efficient methods for structured nonconvex-nonconcave min-max optimization. International Conference on Artificial Intelligence and Statistics, pp. 2746–2754 (2021)
30. Fitzpatrick, S.: Representing monotone operators by convex functions. In: Workshop/Miniconference on Functional Analysis and Optimization, pp. 59–65 (1988). Centre for Mathematics and its Applications, Mathematical Sciences Institute ..
31. Simons, S., Zalinescu, C.: A new proof for Rockafellar's characterization of maximal monotone operators **132**(10), 2969–2972 (2004)
32. Iusem, A., Jofré, A., Oliveira, R.I., Thompson, P.: Extragradient method with variance reduction for stochastic variational inequalities. SIAM J. Optim. **27**(2), 686–72410526234 (2017)
33. Iusem, A.N., Jofré, A., Oliveira, R.I., Thompson, P.: Variance-based Extragradient methods with line search for stochastic variational inequalities. SIAM J. Optim. **29**(1), 175–206 (2019). https://doi.org/10.1137/17M1144799
34. Geiersbach, C., Pflug, G.C.: Projected stochastic gradients for convex constrained problems in Hilbert spaces. SIAM J. Optim. **29**(3), 2079–2099 (2019). https://doi.org/10.1137/18M1200208
35. Geiersbach, C., Wollner, W.: A stochastic gradient method with mesh refinement for PDE-constrained optimization under uncertainty. SIAM J. Sci. Comput. **42**(5), 2750–2772 (2020). https://doi.org/10.1137/19M1263297
36. Polyak, B.T.: Some methods of speeding up the convergence of iteration methods. USSR Comput. Math. Math. Phys. **4**(5), 1–17 (1964). https://doi.org/10.1016/0041-5553(64)90137-5
37. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Applied Optimization, vol. 87. Kluwer Academic Publishers, (2004)
38. Polyak, B.T.: Introduction to Optimization. Optimization Software, (1987)
39. Attouch, H., Maingé, P.-E.: Asymptotic behavior of second-order dissipative evolution equations combining potential with non-potential effects. ESAIM: Control, Opt. Calculus of Variations **17**(3), 836–857 (2011)
40. Boţ, R.I., Csetnek, E.: Second order forward-backward dynamical systems for monotone inclusion problems. SIAM J. Control. Optim. **54**(3), 1423–1443 (2016). https://doi.org/10.1137/15M1012657
41. Attouch, H., Peypouquet, J.: Convergence of inertial dynamics and proximal algorithms governed by maximally monotone operators. Math. Program. **174**(1), 391–432 (2019). https://doi.org/10.1007/s10107-018-1252-x
42. Su, W., Boyd, S., Candes, E.J.: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. J. Mach. Learn. Res. (2016)
43. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $o(1/k^2)$. Soviet Math. Doklady **27**(2), 372–376 (1983)
44. Gadat, S., Panloup, F., Saadane, S.: Stochastic heavy ball. Electron. J. Statistics **12**(1), 461–529 (2018). https://doi.org/10.1214/18-EJS1395
45. Lorenz, D.A., Pock, T.: An inertial forward-backward algorithm for monotone inclusions. J. Math. Imag. Vis. **51**(2), 311–325 (2015). https://doi.org/10.1007/s10851-014-0523-2
46. Briceño-Arias, L.M., Combettes, P.L.: A monotone+skew splitting model for composite monotone inclusions in duality. SIAM J. Optim. **21**(4), 1230–1250 (2011). https://doi.org/10.1137/10081602X
47. Bot, R.I., Csetnek, E.R.: An inertial forward-backward-forward primal-dual splitting algorithm for solving monotone inclusion problems. Num. Algorithms **71**(3), 519–540 (2016). https://doi.org/10.1007/s11075-015-0007-5

48. Bot, R.I., Sedlmayer, M., Vuong, P.T.: A relaxed inertial forward-backward-forward algorithm for solving monotone inclusions with application to GANS. arXiv preprint arXiv:2003.07886 (2020)
49. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
50. Juditsky, A., Nemirovski, A., Tauvel, C.: Solving variational inequalities with stochastic mirror-prox algorithm, pp. 17–58 (2011). https://doi.org/10.1214/10-SSY011
51. Yousefian, F., Nedić, A., Shanbhag, U.V.: On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems. Math. Program. **165**(1), 391–431 (2017). https://doi.org/10.1007/s10107-017-1175-y
52. Gidel, G., Berard, H., Vignoud, G., Vincent, P., Lacoste-Julien, S.: A variational inequality perspective on generative adversarial networks. arXiv preprint arXiv:1802.10551 (2018)
53. Mishchenko, K., Kovalev, D., Shulgin, E., Richtárik, P., Malitsky, Y.: Revisiting stochastic extragradient. In: International Conference on Artificial Intelligence and Statistics, pp. 4573–4582 (2020). PMLR
54. Kannan, A., Shanbhag, U.V.: Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. Comput. Optim. Appl. **74**(3), 779–820 (2019)
55. Cui, S., Shanbhag, U.V.: On the analysis of variance-reduced and randomized projection variants of single projection schemes for monotone stochastic variational inequality problems. Set-Valued and Variational Analysis (to appear) (2021)
56. Rosasco, L., Villa, S., Vũ, B.C.: A stochastic inertial forward-backward splitting algorithm for multivariate monotone inclusions. Optimization **65**(6), 1293–1314 (2016). https://doi.org/10.1080/02331934.2015.1127371
57. Palaniappan, B., Bach, F.: Stochastic variance reduction methods for saddle-point problems. In: Advances in Neural Information Processing Systems, pp. 1416–1424 (2016)
58. Gower, R.M., Schmidt, M., Bach, F., Richtárik, P.: Variance-reduced methods for machine learning. Proc. IEEE **108**(11), 1968–1983 (2020). https://doi.org/10.1109/JPROC.2020.3028013
59. Friedlander, M.P., Schmidt, M.: Hybrid deterministic-stochastic methods for data fitting. SIAM J. Sci. Comput. **34**(3), 1380–1405 (2012)
60. Jalilzadeh, A., Shanbhag, U.V., Blanchet, J.H., Glynn, P.W.: Smoothed variable sample-size accelerated proximal methods for nonsmooth stochastic convex programs. arXiv preprint arXiv:1803.00718 (2018)
61. Jofré, A., Thompson, P.: On variance reduction for stochastic smooth convex optimization with multiplicative noise. Math. Program. **174**(1–2), 253–292 (2019)
62. Ghadimi, S., Lan, G., Zhang, H.: Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. Math. Program. **155**(1–2), 267–305 (2016)
63. Gunzburger, M.D., Webster, C.G., Zhang, G.: Stochastic finite element methods for partial differential equations with random input data. Acta Numer. **23**, 521–650 (2014)
64. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer, Dordrecht (2004)
65. Davis, D., Drusvyatskiy, D.: Stochastic model-based minimization of weakly convex functions. SIAM J. Optim. **29**(1), 207–239 (2019)
66. Lan, G.: First-order and Stochastic Optimization Methods for Machine Learning. Springer Series in the Data Sciences. Springer, (2020)
67. Combettes, P.L., Pesquet, J.-C.: Stochastic Quasi-Fejér block-coordinate fixed point iterations with random sweeping. SIAM J. Optim. **25**(2), 1221–1248 (2015). https://doi.org/10.1137/140971233
68. Combettes, P.L., Pesquet, J.-C.: Stochastic Quasi-Fejér block-coordinate fixed point iterations with random sweeping ii: mean-square and linear convergence. Math. Program. **174**(1), 433–451 (2019). https://doi.org/10.1007/s10107-018-1296-y
69. Rosasco, L., Villa, S., Vũ, B.C.: Stochastic Forward-Backward splitting for monotone inclusions. J. Optim. Theory Appl. **169**(2), 388–406 (2016). https://doi.org/10.1007/s10957-016-0893-2
70. Spall, J.C.: Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Trans. Autom. Control **37**(3), 332–341 (1992)
71. Spall, J.C.: A one-measurement form of simultaneous perturbation stochastic approximation. Automatica **33**(1), 109–112 (1997)
72. Duvocelle, B., Mertikopoulos, P., Staudigl, M., Vermeulen, D.: Learning in time-varying games. arXiv preprint arXiv:1809.03066 (2018)

73. Barty, K., Roy, J.-S., Strugarek, C.: Hilbert-valued perturbed subgradient algorithms. Math. Oper. Res. **32**(3), 551–562 (2007). https://doi.org/10.1287/moor.1070.0253
74. Barty, K., Roy, J.-S., Strugarek, C.: A stochastic gradient type algorithm for closed-loop problems. Math. Program. **119**(1), 51–78 (2009). https://doi.org/10.1007/s10107-007-0201-x
75. Lei, J., Shanbhag, U.V.: Distributed variable sample-size gradient-response and best-response schemes for stochastic Nash equilibrium problems over graphs. arXiv:1811.11246 (2019)
76. Lei, J., Shanbhag, U.V.: Asynchronous variance-reduced block schemes for composite non-convex stochastic optimization: block-specific steplengths and adapted batch-sizes. Optimization Methods and Software, pp. 1–31 (2020)
77. Borwein, J.M., Dutta, J.: Maximal monotone inclusions and Fitzpatrick functions. J. Optim. Theory Appl. **171**(3), 757–784 (2016)
78. Auslender, A., Gourgand, M., Guillet, A.: Resolution numerique d'inegalites variationnelles. In: Lecture Notes in Economics and Mathematical Systems (Mathematical Economics) (1974)
79. Chen, Y., Lan, G., Ouyang, Y.: Optimal primal-dual methods for a class of saddle point problems. SIAM J. Optim. **24**(4), 1779–1814 (2014). https://doi.org/10.1137/130919362
80. Monteiro, R.D.C., Svaiter, B.F.: On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. SIAM J. Optim. **20**(6), 2755–2787 (2010). https://doi.org/10.1137/090753127
81. Chen, Y., Lan, G., Ouyang, Y.: Accelerated schemes for a class of variational inequalities. Math. Program. (2017). https://doi.org/10.1007/s10107-017-1161-4
82. Nesterov, Y.: Dual extrapolation and its applications to solving variational inequalities and related problems. Math. Program. **109**(2), 319–344 (2007). https://doi.org/10.1007/s10107-006-0034-z
83. Malitsky, Y.: Golden ratio algorithms for variational inequalities. Math. Program. **184**(1), 383–410 (2020). https://doi.org/10.1007/s10107-019-01416-w
84. Burachik, R.S., Millán, R.D.: A projection algorithm for non-monotone variational inequalities. Set-Valued and Variational Anal. **28**(1), 149–166 (2020). https://doi.org/10.1007/s11228-019-00517-0
85. Rockafellar, R.T., Wets, R.J.: Stochastic variational inequalities: single-stage to multistage. Math. Program. **165**(1), 331–360 (2017). https://doi.org/10.1007/s10107-016-0995-5
86. Rockafellar, T.R., Wets, R.J.-B.: Variational Analysis. Springer, (1998)
87. Shapiro, A., Dentcheva, D., Ruszczyński, A..X.: Lectures on Stochastic Programming: Modeling and Theory. SIAM, (2009)
88. Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning vol. 1. Springer, (2001)

Authors and Affiliations

Shisheng Cui[1] · Uday Shanbhag[1] · Mathias Staudigl[2] · Phan Vuong[3]

Shisheng Cui
suc256@psu.edu

Uday Shanbhag
udaybag@psu.edu

Phan Vuong
T.V.Phan@soton.ac.uk

[1] Department of Industrial and Manufacturing Engineering, Pennsylvania State University,
University Park, PA 16802, USA

[2] Department of Advanced Computing Sciences (DACS), Maastricht University,
Paul-Henri-Spaaklaan 1, 6229 EN Maastricht, The Netherlands

[3] Mathematical Sciences, University of Southampton, Highfield Southampto,
Southampton SO17 1BJ, UK