Stochastic Proximal Point Methods for Monotone Inclusions under Expected Similarity

Abdurakhmon Sadiev
Laurent Condat
Peter Richtárik
Center of Excellence for Generative AI, King Abdullah University of Science and Technology (KAUST) & SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI) Thuwal, Kingdom of Saudi Arabia

Abstract

Monotone inclusions have a wide range of applications, including minimization, saddle-point, and equilibria problems. We introduce new stochastic algorithms, with or without variance reduction, to estimate a root of the expectation of possibly set-valued monotone operators, using at every iteration one call to the resolvent of a randomly sampled operator. We also introduce a notion of similarity between the operators, which holds even for discontinuous operators. We leverage it to derive linear convergence results in the strongly monotone setting.

1. Introduction

We consider stochastic monotone inclusions in a given finite-dimensional real Hilbert space $\mathcal{X}$, which are problems of the form

$$\text{Find } x^\star \in \mathcal{X} \text{ such that } 0 \in A(x^\star), \text{ where } A := \mathrm{E}_{\xi\sim\mathcal{D}}[A_\xi] \tag{1}$$

and A_ξ is a possibly set-valued monotone operator for every random sample ξ of a distribution $\mathcal{D}$. We recall basic notions of monotone operator theory in Section 2 and refer to the textbook Bauschke and Combettes [6] for more details. For instance, when $\mathcal{D}$ is the uniform distribution over $[n] := \{1, \ldots, n\}$ for some $n \geq 2$, (1) becomes the finite-sum monotone inclusion

$$\text{Find } x^\star \in \mathcal{X} \text{ such that } 0 \in A(x^\star) := \frac{1}{n}\sum_{i=1}^{n} A_i(x^\star). \tag{2}$$

We introduce randomized algorithms, with or without variance reduction, to solve (1). They use at every iteration the resolvent of one randomly chosen A_ξ.

1.1. Motivation

Monotone inclusions [6, 73] have a wide range of applications [19, 32, 46], in mechanics [36, 37], partial differential equations [2, 34, 55, 64], mean field games [12, 38], control [75], communications [61, 80], traffic equilibrium [3, 33], optimal transport [62], Nash equilibria and game theory [11, 14, 52, 58, 81], and are of utmost importance in machine learning. Primarily, they encompass optimization problems [4, 15, 17, 66, 74]: minimizing a convex function $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is equivalent to (1) with $A = \partial f$, the subdifferential of f, and finding a stationary point of a

© A. Sadiev, L. Condat & P. Richtárik.

smooth but possibly nonconvex function f is equivalent to (1) with $A = \nabla f$, the gradient of f. Nonconvex nonsmooth optimization problems have variational formulations, too [57]. Moreover, splitting algorithms to solve structured optimization problems can be derived by formulating the problem as a monotone inclusion in a higher-dimensional lifted space. For instance, minimizing $f + \sum_{i=1}^n g_i(K_i x)$, for linear operators $K_i : \mathcal{W} \to \mathcal{U}_i$ and functions g and h_i, can be formulated as (1) with $\mathcal{X} = \mathcal{W} \times \mathcal{U}_1 \times \cdots \times \mathcal{U}_n$ and the monotone operator

$$A = \begin{pmatrix} \partial f & K_1^* u_1 & \cdots & K_n^* u_n \\ -K_1 x & (\partial g_1)^{-1}(u_1) & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ -K_n x & 0 & 0 & (\partial g_n)^{-1}(u_n) \end{pmatrix}, \tag{3}$$

where $\cdot^*$ denotes the adjoint operator. For a suitable preconditioning linear operator P, that is symmetric and positive definite, $P^{-1}A$ is monotone in $\mathcal{X}$ endowed with the modified inner product $\langle \cdot, P \cdot \rangle$, and one can design iterative algorithms to solve $0 \in P^{-1}A(x)$ [5, 13, 20–24, 26, 27, 42, 71, 72].

Besides minimization problems, monotone inclusions allow us to formulate saddle-point problems [16, 28, 29, 50, 54, 60, 67], which have many applications in machine learning [9, 40, 56], e.g. for adversarial training [39, 53], GANs [35], and distributionally robust optimization [59].

We propose different algorithms in the framework of the Stochastic Proximal Point Method (SPPM), with or without variance reduction. Even in the optimization setting, our study under a similarity assumption, which is weaker than smoothness, is new, to the best of our knowledge.

2. Definitions and Properties of Monotone Operators

Let $B : \mathcal{X} \to 2^{\mathcal{X}}$ be a set-valued operator on $\mathcal{X}$. We define its graph $\mathrm{gra}(B) \coloneqq \{(x, u) \in \mathcal{X}^2 \ : \ u \in B(x)\}$ and its inverse B^{-1} as the set-valued operator whose graph is $\mathrm{gra}(B^{-1}) \coloneqq \{(u, x) \in \mathcal{X}^2 \ : \ u \in B(x)\}$. $x \in \mathcal{X}$ is a *zero* of B if $0 \in B(x)$.

2.1. Monotone Operators

B is *monotone* if for every (x, u) and (y, v) in $\mathrm{gra}(B)$,

$$\langle u - v, x - y \rangle \geq 0.$$

B is *maximally monotone* if there exists no monotone operator whose graph strictly contains $\mathrm{gra}(B)$. B is (maximally) monotone if and only if B^{-1} is (maximally) monotone. The subdifferential ∂f of a proper lower semicontinuous convex function f is maximally monotone.

B is *μ-strongly monotone* for some $\mu > 0$ if, for every (x, u) and (y, v) in $\mathrm{gra}(B)$,

$$\langle u - v, x - y \rangle \geq \mu \|x - y\|^2. \tag{4}$$

In that case, γB is $\gamma\mu$-strongly monotone, for every $\gamma > 0$. If B is μ-strongly maximally monotone, its zero exists and is unique.

The following assumption on the operators in (1) will be considered to analyze the proposed algorithms.

Assumption 1 (strong monotonicity) *There exists $\mu > 0$ such that A_ξ is μ-strongly maximally monotone for every $\xi \sim \mathcal{D}$. Therefore, $A \coloneqq \mathrm{E}_{\xi \sim \mathcal{D}}[A_\xi]$ is μ-strongly maximally monotone as well and the solution $x^\star$ to (1) exists and is unique.*

A single valued operator $C : \mathcal{X} \to \mathcal{X}$ is β-*cocoercive* for some $\beta > 0$ if, for every $(x, y) \in \mathcal{X}^2$,

$$\langle x - y, C(x) - C(y)\rangle \geq \beta\|C(x) - C(y)\|^2.$$

A function f is L-*smooth* for some $L > 0$ if it is differentiable and its gradient ∇f is L-Lipschitz continuous; that is, for every $(x, y) \in \mathcal{X}^2$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

In that case, ∇f is L^{-1}-cocoercive, according to the Baillon–Haddad theorem. This equivalence between Lipschitz-continuity and cocoercivity only holds for operators which are gradients of convex functions. In general, a monotone operator can be Lipschitz-continuous without being cocoercive. A prominent example is the skew operator $(x, y) \in \mathcal{X}^2 \mapsto (K^*y, Kx)$ for any linear operator K on $\mathcal{X}$, which is $\|K\|$-Lipschitz continuous but not cocoercive. Thus, monotone inclusions are much more general than optimization problems. In particular, the forward algorithm generalizing gradient descent, which iterates $x^{k+1} := x^k - \gamma A(x^k)$ for a maximally monotone single-valued operator A and a stepsize $\gamma > 0$, converges if A is cocoercive, but not if it is merely Lipschitz-continuous (take $-I$ as an example, where I denotes the identity: the iteration diverges for every $\gamma > 0$ and $x^0 \neq 0$). This is why robust iterative fixed-point algorithms to solve monotone inclusions use the resolvent of the monotone operators, as we describe in the next section.

2.2. The Resolvent and the Proximal Point Method

The *resolvent* of B is the operator $(I + B)^{-1}$. According to the Minty theorem, if B is maximally monotone, its resolvent is defined everywhere and single-valued. The resolvent of a strongly monotone operator is contractive:

Lemma 1 (contractivity of the resolvent) *Let $B : \mathcal{X} \to 2^{\mathcal{X}}$ be a μ-strongly maximally monotone operator, for some $\mu > 0$. Then its resolvent is $(1 + \mu)^{-1}$-contractive; that is, for every $(x, y) \in \mathcal{X}$,*

$$\|x^+ - y^+\| \leq \frac{1}{1 + \mu}\|x - y\|, \tag{5}$$

where $x^+ = (I + B)^{-1}(x)$, $y^+ = (I + B)^{-1}(y)$.

The resolvent of the subdifferential ∂f of a function f is its proximity operator $\mathrm{prox}_f = (I + \partial f)^{-1} : x \in \mathcal{X} \mapsto \arg\min_y \big(f(y) + \frac{1}{2}\|y - x\|^2\big)$. Optimization algorithms making use of proximity operators are called proximal algorithms [27, 63, 70]. The iteration $x^{k+1} := \mathrm{prox}_f(x^k)$ to minimize a function f, and by extension the iteration $x^{k+1} := (I + B)^{-1}(x^k)$ to find a zero of the operator B, is called the *proximal point algorithm*, or *proximal point method* (PPM) [68]. It follows from Lemma 1 that if B is μ-strongly maximally monotone, the PPM converges linearly to its zero $x^\star = (I + B)^{-1}(x^\star)$, which exists and is unique, since $\|x^{k+1} - x^\star\| \leq \frac{1}{1+\mu}\|x^k - x^\star\|$ for every $k \geq 0$.

2.3. Similarity Between Operators

It is natural to consider that there exists some level of similarity or homogeneity between the operators A_i, in particular in machine learning where they express characteristics of underlying data [18, 43, 76]. To capture this property, we define two notions of similarity.

Assumption 2 (expected similarity) *In Problem* (1)*, there exist a solution* $x^\star$ *and a constant* $\delta > 0$ *such that, for every* $x \in \mathcal{X}$*,* $\mathcal{D}$*-almost every* ξ *and every* $a_\xi \in A_\xi(x)$*, there exists* $a_\xi^\star \in A_\xi(x^\star)$ *such that* $\mathrm{E}_{\xi' \sim \mathcal{D}}\left[a_{\xi'}^\star\right] = 0$ *and*

$$\mathrm{E}_{\xi \sim \mathcal{D}}\left[\left\|a_\xi - \mathrm{E}_{\xi' \sim \mathcal{D}}\left[a_{\xi'}\right] - a_\xi^\star\right\|^2\right] \leq \delta^2 \|x - x^\star\|^2. \tag{6}$$

This assumption can be satisfied by set-valued operators with discontinuities and is even weaker than assuming every $A_\xi - A$ to be Lipschitz-continuous at $x^\star$ (see an example in Appendix A).

Assumption 3 (average similarity) *In Problem* (2)*, there exist a solution* $x^\star$ *and* $\tilde{\delta} > 0$ *such that, for every* $x_i \in \mathcal{X}$ *and* $a_i \in A_i(x_i)$*,* $i \in [n]$*, there exist* $a_i^\star \in A_i(x^\star)$*,* $i \in [n]$*, such that* $\sum_{i=1}^n a_i^\star = 0$*, and*

$$\frac{1}{n}\sum_{i=1}^n \left\|a_i - \frac{1}{n}\sum_{j=1}^n a_j - a_i^\star\right\|^2 \leq \frac{\tilde{\delta}^2}{n}\sum_{i=1}^n \|x_i - x^\star\|^2. \tag{7}$$

Assumption 3 is stronger than Assumption 2 with $\mathcal{D}$ the uniform distribution over $[n]$, since (7) with $x_1 = \cdots = x_n = x$ implies (6).

Related definitions of similarities have been considered in several works [43, 47, 49, 51, 76, 77]. For instance, the property that for every $(x, y) \in \mathcal{X}^2$

$$\frac{1}{n}\sum_{i=1}^n \|A_i(x) - A(x) - A_i(y) + A(y)\|^2 \leq \delta^2 \|x - y\|^2,$$

in the case where the $A_i = \nabla f_i$ are gradients of smooth functions f_i, is called Hessian variance in Szlendak et al. [77] and δ-average second-order similarity in Lin et al. [51]. Indeed, if the functions f_i are twice differentiable, this property is equivalent to the one that, for every $x \in \mathcal{X}$,

$$\frac{1}{n}\sum_{i=1}^n \left\|\nabla^2 f_i(x) - \nabla^2 f(x)\right\|^2 \leq \delta^2;$$

that is, the variance of the Hessians $\nabla^2 f_i$ is uniformly bounded.

3. The Stochastic Proximal Point Method (SPPM)

The Stochastic Proximal Point Method (SPPM; Algorithm 1, Appendix D) consists of iterating the resolvent of an operator A_{ξ^k} chosen randomly at every iteration k. Under Assumption 1, it converges linearly to a neighborhood of $x^\star$.

Theorem 1 *In Problem* (1)*, let Assumption 1 hold, and for every* $\xi \sim \mathcal{D}$*, let* $a_\xi^\star \in A_\xi(x^\star)$*, such that* $\mathrm{E}_{\xi \sim \mathcal{D}}\left[a_\xi^\star\right] = 0$*. Such* $a_\xi^\star$ *exist by definition of* $x^\star$*. If they are not unique, we define them as ones minimizing*

$$\sigma_\star^2 := \mathrm{E}_{\xi \sim \mathcal{D}}\left[\left\|a_\xi^\star\right\|^2\right]. \tag{8}$$

Then in SPPM *with any stepsize* $\gamma > 0$ *and initial estimate* $x^0 \in \mathcal{X}$*, we have, for every* $k \geq 0$*,*

$$\mathrm{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq \left(\frac{1}{1+\gamma\mu}\right)^{2k} \left\|x^0 - x^\star\right\|^2 + \frac{1 - (1+\gamma\mu)^{-2k}}{(1+\gamma\mu)^2 - 1}\gamma^2\sigma_\star^2 \tag{9}$$

$$\leq \left(\frac{1}{1+\gamma\mu}\right)^{2k} \left\|x^0 - x^\star\right\|^2 + \frac{\gamma\sigma_\star^2}{2\mu + \gamma\mu^2}. \tag{10}$$

Our result is tight: (9) is satisfied with an equality with the operators $A_\xi(x) = \mu(x - x^\star) + a_\xi^\star$ for some $\mu > 0$, $x^\star \in \mathcal{X}$, and $a_\xi^\star \in \mathcal{X}$ such that $\mathrm{E}_{\xi\sim\mathcal{D}}\left[a_\xi^\star\right] = 0$.

Even in the optimization setting, Theorem 1 is new. In Bertsekas [7], the SPPM, called *incremental proximal algorithm*, was studied to minimize a finite sum of functions, but the convergence bounds depend on the number of functions, so they are not applicable to our setting where the distribution $\mathcal{D}$ is arbitrary. In Bianchi [10] and Toulis et al. [78], convergence results with decreasing stepsizes are derived. SPPM-type algorithms have been studied for stochastic optimization in Asi and Duchi [1], with a focus on stability in the case of inexact computation of the proximity operator. In Davis and Drusvyatskiy [30] the SPPM is studied for optimization, but their convergence analysis (Theorem 4.4) relies on the decay of the function values, so it is not applicable to our setting.

In Ryu and Boyd [69, Theorem 7], in the convex optimization setting, by simply using the triangular inequality $\|x^{k+1} - x^\star\| \leq \| \left(I + \gamma A_{\xi^k}\right)^{-1}(x^k) - \left(I + \gamma A_{\xi^k}\right)^{-1}(x^\star)\| + \| \left(I + \gamma A_{\xi^k}\right)^{-1}(x^\star) - x^\star\| \leq (1+\gamma\mu)^{-1}\|x^k - x^\star\| + \gamma\|\tilde{a}^\star_{\xi^k}\|$, where $\tilde{a}_\xi^\star$ is the minimum-norm element of $A_\xi(x^\star)$, they obtain

$$\mathrm{E}\left[\|x^k - x^\star\|\right] \leq \left(\frac{1}{1+\gamma\mu}\right)^k \|x^0 - x^\star\| + \frac{(1+\gamma\mu)\tilde{\sigma}_\star}{\mu},$$

where $\tilde{\sigma}_\star = \mathrm{E}_{\xi\sim\mathcal{D}}\left[\|\tilde{a}_\xi^\star\|\right]$. The neighborhood size does not tend to zero when $\gamma \to 0$, as is the case in (10). In Patrascu and Necoara [65, Theorem 10], the following result is obtained in the convex optimization setting with *smooth* functions:

$$\mathrm{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq 2\left(\frac{1}{1+\gamma\mu}\right)^{2k} \|x^0 - x^\star\|^2 + \frac{2(1+\gamma\mu)^2\sigma_\star^2}{\gamma^2}.$$

The neighborhood size tends to $+\infty$ when $\gamma \to 0$, whereas it should tend to zero. In the same setting, Khaled and Jin [47, eq. 19] derived

$$\mathrm{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq \left(\frac{1}{1+\gamma\mu}\right)^{k} \|x^0 - x^\star\|^2 + \frac{\gamma\sigma_\star^2}{\mu}.$$

The rate and the neighborhood size are larger than in (10). Thus, even in the optimization setting, our result is new and tight, with a simple and elegant proof.

4. The SPPM with Variance Reduction

SPPM with Operator Correction (SPPM-OC) The SPPM does not converge to the exact solution $x^\star$ of (1) but only to its neighborhood. To correct this shortcoming, we propose a new algorithm, the SPPM with Operator Correction (SPPM-OC; Algorithm 2 in the Appendix). It is variance-reduced [41]; that is, it converges to the exact solution under Assumptions 1 and 2. This is achieved by adding a shift to x^k before applying the resolvent of a randomly chosen A_{ξ^k}, to correct for the difference between A_{ξ^k} and its expectation A.

Theorem 2 *In Problem* (1)*, let Assumptions 1 and 2 hold. Then, with a suitable selection of a stepsize, the iteration complexity of* SPPM-OC *to achieve ϵ-accuracy for any $\epsilon > 0$ is*

$$\mathcal{O}\left(\left(\frac{\delta^2}{\mu^2} + 1\right)\log\left(\frac{\left\|x^0 - x^\star\right\|^2}{\epsilon}\right)\right).$$

Thus, SPPM-OC converges linearly to the solution $x^\star$. But it requires to select an element a^k in $A(x^k)$ at every iteration, which can be costly or even impractical. Therefore, in the next section, we study another algorithm, in which this selection is performed with a small probability only.

The Loopless Stochastic Variance-Reduced Proximal Point Method (L-SVRP) In the optimization setting with convex *differentiable* functions, the Stochastic Variance-Reduced Proximal Point Method (SVRP) was proposed in Khaled and Jin [47]. It was discovered independently in Traoré et al. [79], with an analysis based on the decay of the function values, which is not applicable to our setting. This algorithm is a proximal analog of the Stochastic Variance-Reduced Gradient Method (SVRG) [45, 82], hence its name. More precisely, it is a proximal analog of loopless versions of SVRG called L-SVRG [44, 48]. That is why we call the algorithm the Loopless Stochastic Variance-Reduced Proximal Point Method (L-SVRP), to emphasize its loopless nature. We introduce and study L-SVRP (Algorithm 3, Appendix F) in the much more general setting of set-valued monotone inclusions.

Theorem 3 *(Convergence of* L-SVRP*; informal) In Problem* (1)*, let Assumptions 1 and 2 hold. Then, with an appropriate selection of stepsizes,* L-SVRP *(Algorithm 3) solves Problem* (1) *in*

$$\mathcal{O}\left(\left(\frac{\delta^2}{\mu^2}+\frac{1}{p}\right)\log\left(\frac{V^0}{\epsilon}\right)\right).$$

The best value of p depends on how much more costly it is to pick an element $a^k \in A(x^k)$ than to apply the resolvent of an A_ξ. In any case, there is no interest in choosing p larger than $\frac{\mu^2}{\delta^2}$, which is typically very small. Hence, L-SVRP can be orders of magnitude faster than SPPM-OC, which corresponds to the particular case of L-SVRP with $p = 1$.

In the case of minimizing a sum of n differentiable functions f_i, i.e. Problem (2) with $A_i = \nabla f_i$, with $p = \frac{1}{n}$, we recover the same iteration complexity as in Khaled and Jin [47].

Point-SAGA for Monotone Inclusion Problem Point-SAGA (Algorithm 4, Appendix G) is an algorithm proposed by Defazio [31] for the minimization of a sum of convex functions, using at every iteration the proximity operator of one randomly chosen function. It was also studied as a randomized primal–dual algorithm in Condat and Richtárik [25]. The algorithm converges linearly when all functions are smooth and strongly convex. We introduce and study Point-SAGA in the general setting of set-valued monotone inclusions. Point-SAGA is an alternative to the snapshot algorithm L-SVRP that never requires invoking the average operator A. As a counterpart, Point-SAGA is limited to the finite-sum problem (2), since n elements of $\mathcal{X}$ are stored in a memory table.

Theorem 4 *(Convergence of* Point-SAGA*; informal) In Problem* (2)*, let Assumptions 1 and 3 hold. Then, with an appropriate selection of stepsizes, the iteration complexity of* Point-SAGA *to achieve ϵ-accuracy for any $\epsilon > 0$ is*

$$\mathcal{O}\left(\left(\frac{\widetilde{\delta}^2}{\mu^2}+n\right)\log\frac{1}{\epsilon}\right).$$

To the best of our knowledge, the analysis of Point-SAGA under a similarity assumption is new, even in the particular case of minimizing convex functions.

Acknowledgement. The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST): i) KAUST Baseline Research Scheme, ii) Center of Excellence for Generative AI, under award number 5940, iii) SDAIA-KAUST Center of Excellence in Artificial Intelligence and Data Science.

References

[1] H. Asi and J. C. Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM J. Optim.*, 29(3):2257–2290, 2019.

[2] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Alternating proximal algorithms for weakly coupled convex minimization problems—applications to dynamical games and PDE's. *J. Convex Anal.*, 15:485–506, 2008.

[3] H. Attouch, L. M. Briceño-Arias, and P. L. Combettes. A parallel splitting method for coupled monotone inclusions. *SIAM J. Control Optim.*, 48:3246–3270, 2010.

[4] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Found. Trends Mach. Learn.*, 4(1):1–106, 2012.

[5] H. H. Bauschke and P. L. Combettes. A Dykstra-like algorithm for two monotone operators. *Pacific J. Optim.*, 4:383–391, 2008.

[6] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, 2nd edition, 2017.

[7] D. P. Bertsekas. Incremental proximal methods for large scale convex optimization. *Math. Program.*, 129:163–195, 2011.

[8] D. P. Bertsekas. *Convex optimization algorithms*. Athena Scientific, Belmont, MA, USA, 2015.

[9] A. Beznosikov, B. Polyak, E. Gorbunov, D. Kovalev, and A. Gasnikov. Smooth monotone stochastic variational inequalities and saddle point problems: A survey. *European Mathematical Society Magazine*, (127):15–28, 2023.

[10] P. Bianchi. Ergodic convergence of a stochastic proximal point algorithm. *SIAM Journal on Optimization*, 26(4):2235–2260, 2016.

[11] M. Bravo, D. Leslie, and P. Mertikopoulos. Bandit learning in concave N-person games. In *Proc. of Conf. Neural Information Processing Systems (NeurIPS)*, 2018.

[12] L. M. Briceño-Arias and D. Davis. Forward-backward-half forward algorithm for solving monotone inclusions. *SIAM J. Optim.*, 28(4):2839–2871, 2018.

[13] L. M. Briceño-Arias and P. L. Combettes. A monotone+skew splitting model for composite monotone inclusions in duality. *SIAM J. Optim.*, 21(4):1230–1250, October 2011.

[14] Luis M. Briceño-Arias and Patrick L. Combettes. Monotone operator methods for nash equilibria in non-potential games. In David H. Bailey, Heinz H. Bauschke, Peter Borwein, Frank Garvan, Michel Théra, Jon D. Vanderwerff, and Henry Wolkowicz, editors, *Computational and Analytical Mathematics*, pages 143–159, New York, NY, 2013. Springer New York.

[15] S. Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8 (3–4):231–357, 2015.

[16] M. N. Bùi and P. L . Combettes. Multivariate monotone inclusions in saddle form. *Mathematics of Operations Research*, 47:1082–1109, 2022.

[17] V. Cevher, S. Becker, and M. Schmidt. Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics. *IEEE Signal Process. Mag.*, 31(5):32–43, 2014.

[18] E. M. Chayti and S. P. Karimireddy. Optimization with access to auxiliary information. *Transactions on Machine Learning Research*, 2024.

[19] P. L. Combettes. Systems of structured monotone inclusions: duality, algorithms, and applications. *SIAM J. Optim.*, 23:2420–2447, 2013.

[20] P. L. Combettes. Monotone operator theory in convex optimization. *Math. Program.*, 170(1): 177–206, July 2018.

[21] P. L. Combettes and J.-C. Pesquet. Primal–dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators. *Set-Val. Var. Anal.*, 20(2):307–330, 2012.

[22] P. L. Combettes and J.-C. Pesquet. Fixed point strategies in data science. *IEEE Trans. Signal Process.*, 69:3878–3905, 2021.

[23] P. L. Combettes, L. Condat, J.-C. Pesquet, and B. C. Vũ. A forward–backward view of some primal–dual optimization methods in image recovery. In *Proc. of IEEE ICIP*, Paris, France, October 2014.

[24] L. Condat. A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *J. Optim. Theory Appl.*, 158(2):460–479, 2013.

[25] L. Condat and P. Richtárik. RandProx: Primal-dual optimization algorithms with randomized proximal updates. In *Proc. of Int. Conf. Learning Representations (ICLR)*, 2023.

[26] L. Condat, G. Malinovsky, and P. Richtárik. Distributed proximal splitting algorithms with rates and acceleration. *Frontiers in Signal Processing*, 1, January 2022.

[27] L. Condat, D. Kitahara, A. Contreras, and A. Hirabayashi. Proximal splitting algorithms for convex optimization: A tour of recent advances, with new twists. *SIAM Review*, 65(2):375–435, 2023.

[28] D. Davis. SMART: the stochastic monotone aggregated root-finding algorithm. preprint arXiv:1601.00698, 2016.

[29] D. Davis. Variance reduction for root-finding problems. *Mathematical Programming*, 197: 375–410, 2023.

[30] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[31] A. Defazio. A simple practical accelerated method for finite sums. In *Proc. of 30st Conf. Neural Information Processing Systems (NIPS)*, volume 29, pages 676–684, 2016.

[32] F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2003.

[33] M. Fukushima. The primal Douglas–Rachford splitting algorithm for a class of monotone mappings with applications to the traffic equilibrium problem. *Math. Program.*, 72:1–15, 1996.

[34] N. Ghoussoub. *Self-dual Partial Differential Systems and Their Variational Principles*. Springer-Verlag, New York, 2009.

[35] G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *Proc. of Int. Conf. Learning Representations (ICLR)*, 2019.

[36] R. Glowinski. *Numerical Methods for Nonlinear Variational Problems*. Springer Series in Computational Physics. Springer, Berlin, 1984.

[37] R. Glowinski and P. Le Tallec. Augmented Lagrangian and operator-splitting methods in nonlinear mechanics. *SIAM Studies in Applied and Numerical Mathematics*, 1989.

[38] D. A. Gomes and J. Saúde. Numerical methods for finite-state mean-field games satisfying a monotonicity condition. *Appl Math Optim*, 83:51–82, 2021.

[39] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *Proc. of Int. Conf. Learning Representations (ICLR)*, 2015.

[40] E. Gorbunov, N. Loizou, and G. Gidel. Extragradient method: $O(1/K)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *Proc. of Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, volume PMLR 151, 2022.

[41] R. M. Gower, M. Schmidt, F. Bach, and P. Richtárik. Variance-reduced methods for machine learning. *Proc. of the IEEE*, 108(11):1968–1983, November 2020.

[42] B. S. He and X. M Yuan. Convergence analysis of primal–dual algorithms for a saddle-point problem: from contraction perspective. *SIAM J. Imaging Sci.*, 5:119–149, 2012.

[43] H. Hendrikx, L. Xiao, S. Bubeck, F. Bach, and L. Massoulie. Statistically preconditioned accelerated gradient method for distributed optimization. In *Proc. of 37th Int. Conf. Machine Learning (ICML)*, volume PMLR 119, pages 4203–4227, 2020.

[44] T. Hofmann, A. Lucchi, S. Lacoste-Julien, and B. McWilliams. Variance reduced stochastic gradient descent with neighbors. In *Proc. of 29th Conf. Neural Information Processing Systems (NIPS)*, pages 1509–1519, 2015.

[45] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Proc. of 27th Conf. Neural Information Processing Systems (NIPS)*, pages 315–323, 2013.

[46] A. Kaplan and R. Tichatschke. A general view on proximal point methods to variational inequalities in Hilbert spaces — iterative regularization and approximation. *Journal of Nonlinear and Convex Analysis*, 2:305–332, 2001.

[47] A. Khaled and C. Jin. Faster federated optimization under second-order similarity. In *Proc. of Int. Conf. Learning Representations (ICLR)*, 2023.

[48] D. Kovalev, S. Horváth, and P. Richtárik. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Proc. of 31st Int. Conf. Algorithmic Learning Theory (ALT)*, volume PMLR 117, pages 451–467, 2020.

[49] D. Kovalev, A. Beznosikov, E. D. Borodich, A. Gasnikov, and G. Scutari. Optimal gradient sliding and its application to optimal distributed optimization under similarity. In *Proc. of Conf. Neural Information Processing Systems (NeurIPS)*, 2022.

[50] V. N. Lebedev and N. T. Tynjanskii. Duality theory of concave-convex games. *Dokl. Akad. Nauk SSSR*, 174(6):1264–1267, 1967.

[51] D. Lin, Y. Han, H. Ye, and Z. Zhang. Stochastic distributed optimization under average second-order similarity: Algorithms and analysis. In *Proc. of Conf. Neural Information Processing Systems (NeurIPS)*, 2023.

[52] N. Loizou, H. Berard, G. Gidel, I. Mitliagkas, and S. Lacoste-Julien. Stochastic gradient descent ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. In *Proc. of Conf. Neural Information Processing Systems (NeurIPS)*, 2021.

[53] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *Proc. of Int. Conf. Learning Representations (ICLR)*, 2018.

[54] L. McLinden. An extension of Fenchel's duality theorem to saddle functions and dual minimax problems. *Pacific J. Math.*, 50:135–158, 1974.

[55] B. Mercier. *Topics in Finite Element Solution of Elliptic Problems*. Springer Berlin, Heidelberg, 1979.

[56] P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *Proc. of Int. Conf. Learning Representations (ICLR)*, 2019.

[57] B. S. Mordukhovich. *Variational analysis and generalized differentiation II: Applications*. Springer, 2006.

[58] O. Morgenstern and J. Von Neumann. *Theory of games and economic behavior*. Princeton university press, 1953.

[59] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Proc. of 30st Conf. Neural Information Processing Systems (NIPS)*, page 2208–2216, 2016.

[60] A. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[61] D. P. Palomar and Y. C. Eldar, editors. *Convex Optimization in Signal Processing and Communications*. Cambridge University Press, 2009.

[62] N. Papadakis, G. Peyré, and E. Oudet. Optimal transport with proximal splitting. *SIAM J. Imaging Sci.*, 7:212–238, 2014.

[63] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 3(1): 127–239, 2014.

[64] J. Park. Additive Schwarz methods for semilinear elliptic problems with convex energy functionals: Convergence rate independent of nonlinearity. *SIAM Journal on Scientific Computing*, 46(3):A1373–A1396, 2024.

[65] A. Patrascu and I. Necoara. Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *Journal of Machine Learning Research*, 18:1–42, 2018.

[66] N. G. Polson, J. G. Scott, and B. T. Willard. Proximal algorithms in statistics and machine learning. *Statist. Sci.*, 30(4):559–581, 2015.

[67] R. T. Rockafellar. Minimax theorems and conjugate saddle-functions. *Mathematica Scandinavica*, 14:151–173, 1964.

[68] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.*, 14(5):877–898, 1976.

[69] E. K. Ryu and S. Boyd. Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent. Research report on author's website, 2014.

[70] E. K. Ryu and S. Boyd. A primer on monotone operator methods. *Appl. Comput. Math.*, 1 (15):3–43, 2016.

[71] A. Salim, L. Condat, D. Kovalev, and P. Richtárik. An optimal algorithm for strongly convex minimization under affine constraints. In *Proc. of Int. Conf. Artif. Intell. Stat. (AISTATS), PMLR 151*, pages 4482–4498, 2022.

[72] A. Salim, L. Condat, K. Mishchenko, and P. Richtárik. Dualize, split, randomize: Toward fast nonsmooth optimization algorithms. *J. Optim. Theory Appl.*, July 2022.

[73] S. Simons. *From Hahn–Banach to Monotonicity*. Springer-Verlag, Berlin, 2008.

[74] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning*. The MIT Press, 2011.

[75] G. Stathopoulos, H. Shukla, A. Szucs, Y. Pu, and C. N. Jones. Operator splitting methods in control. *Foundations and Trends in Systems and Control*, 3(3):249–362, 2016.

[76] Y. Sun, G. Scutari, and A. Daneshmand. Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation. *SIAM Journal on Optimization*, 32(2): 354–385, 2022.

[77] R. Szlendak, A. Tyurin, and P. Richtárik. Permutation compressors for provably faster distributed nonconvex optimization. In *Proc. of Int. Conf. Learning Representations (ICLR)*, 2022.

[78] P. Toulis, D. Tran, and E. Airoldi. Towards stability and optimality in stochastic gradient descent. In *Proc. of Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, pages 1290–1298, 2016.

[79] C. Traoré, V. Apidopoulos, S. Salzo, and S. Villa. Variance reduction techniques for stochastic proximal point algorithms. preprint arXiv:2308.09310, 2023.

[80] R. Xin, S. Pu, A. Nedić, and U. A. Khan. A general framework for decentralized optimization with first-order methods. *Proceedings of the IEEE*, 108(11):1869–1889, November 2020.

[81] P. Yi and L. Pavel. An operator splitting approach for distributed generalized nash equilibria computation. *Automatica*, 102:111–121, 2019.

[82] L. Zhang, M. Mahdavi, and R. Jin. Linear convergence with condition number independent access of full gradients. In *Proc. of 27th Conf. Neural Information Processing Systems (NIPS)*, 2013.

Appendix

Contents

1 Introduction **1**
1.1 Motivation 1

2 Definitions and Properties of Monotone Operators **2**
2.1 Monotone Operators 2
2.2 The Resolvent and the Proximal Point Method 3
2.3 Similarity Between Operators 3

3 The Stochastic Proximal Point Method (SPPM) **4**

4 The SPPM with Variance Reduction **5**

A Simple Example **14**

B Experiments **14**

C Proof of Lemma 1 **16**

D SPPM: Convergence Analysis **16**
D.1 Proof of Theorem 1 16

E SPPM-OC: Convergence Analysis **17**
E.1 Proof of Theorem 5 18

F L-SVRP: Convergence Analysis **19**
F.1 Proof of Theorem 6 20

G Point-SAGA: Convergence Analysis **21**
G.1 Proof of Theorem 7 22

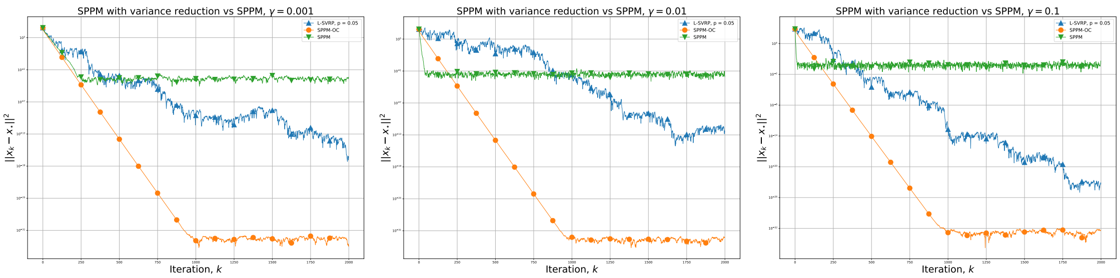

Figure 1: Performance comparison of SPPM with, from left to right, $\gamma = 10^{-3}, 10^{-2}, 10^{-1}$, SPPM-OC, L-SVRP. SPPM-OC and L-SVRP have the same parameter values in the 3 plots, the differences are only due to randomness.

Appendix A. Simple Example

Let us give a simple example of $n = 2$ maximally monotone operators $A_1 : x \in \mathbb{R} \mapsto (\{1\}$ if $x < 1$, $[1,3]$ if $x = 1$, $\{3\}$ if $x > 1)$ and $A_2 : x \in \mathbb{R} \mapsto (\{4x-7\}$ if $x < 1$, $[-3,-1]$ if $x = 1$, $\{4x-5\}$ if $x > 1)$ on $\mathcal{X} = \mathbb{R}$, with $\mathcal{D}$ the uniform distribution on $[n]$. We have $A = \frac{1}{2}(A_1 + A_2) : x \in \mathbb{R} \mapsto (\{2x-3\}$ if $x < 1$, $[-1,1]$ if $x = 1$, $\{2x-1\}$ if $x > 1)$, and $x^\star = 1$. For every $x < 1$, with $a_1 = 1 \in A_1(x)$, $a_2 = 4x - 7 \in A_2(x)$, $a_1^\star = 2$, $a_2^\star = -2$, we can check that (6) is satisfied with $\delta = 2$, as the left-hand side is $(2x-2)^2$. For every $x > 1$, with $a_1 = 3 \in A_1(x)$, $a_2 = 4x - 5 \in A_2(x)$, $a_1^\star = 2$, $a_2^\star = -2$, we can check that (6) is satisfied with $\delta = 2$, as the left-hand side is $(2x-2)^2$ as well. At $x = x^\star = 1$, for every $a_1 \in [1,3] = A_1(x)$ and $a_2 \in [-3,-1] = A_2(x)$, with $a_1^\star = \frac{1}{2}(a_1 - a_2) = -a_2^\star$, (6) is satisfied with any δ, as the left-hand side is zero. Overall, (6) is satisfied $\delta = 2$.

Appendix B. Experiments

We perform numerical experiments for the saddle-point problem

$$\min_{y \in \mathbb{R}^{d_y}} \max_{z \in \mathbb{R}^{d_z}} \frac{1}{n} \sum_{i}^{n} f_i(y, z),$$

for some vector dimensions $d_y \geq 1$ and $d_z \geq 1$, where each f_i is a strongly convex–strongly concave function defined as

$$f_i : (y, z) \mapsto \frac{1}{2} \langle y, M_i y \rangle + \langle b_i, y \rangle + \langle z, Q_i y \rangle - \langle c_i, z \rangle - \frac{1}{2} \langle z, N_i z \rangle,$$

with the following parameters:

- Each matrix $M_i \in \mathbb{R}^{d_y \times d_y}$ and $N_i \in \mathbb{R}^{d_z \times d_z}$ is generated randomly with apriori selected eigenvalues $\lambda_l(M_i) = 10^l$ and $\lambda_j(N_i) = 10^j$ respectively, where $l \in \{0, 1, \ldots, d_y - 1\}$ and $j \in \{0, 1, \ldots, d_z - 1\}$;
- The vectors $b_i \in \mathbb{R}^{d_y}$ and $c_i \in \mathbb{R}^{d_z}$ are sampled from normal distributions $\mathcal{N}(1, 5 \cdot I_{d_y})$ and $\mathcal{N}(1, 5 \cdot I_{d_z})$ respectively;

- Every element of the matrix Q_i is sampled from the standard normal distribution, then each column is normalized to have a full-rank matrix.

To formulate the problem as Problem (2), we define $x \coloneqq (y, z)$ and the single-valued monotone operators $A_i : x \mapsto (\nabla_y^\top f_i(y,z), -\nabla_z^\top f_i(y,z))^\top$; that is,

$$A_i(x) = \left(\begin{array}{c|c} M_i & Q_i^\top \\ \hline -Q_i & N_i \end{array}\right) x + \begin{pmatrix} b_i \\ c_i \end{pmatrix} = \mathbb{B}_i x + r_i.$$

We take $n = 200$, $d_y = 3$, $d_z = 4$. Each operator A_i is 1-strongly monotone and L-Lipschitz-continuous with $L = 1000$.

We compute the similarity constant δ as follows. By Assumption 2, we have

$$\frac{1}{n}\sum_{i=1}^{n} \|A_i(x) - A(x) - A_i(x^\star) + A(x^\star)\|^2 \leq \delta^2 \|x - x^\star\|^2,$$

Plugging in the expression for $A_i(x) = \mathbb{B}_i x + r_i$, we obtain

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \|A_i(x) - A(x) - A_i(x^\star) + A(x^\star)\|^2 &= \frac{1}{n}\sum_{i=1}^{n} \left\| \mathbb{B}_i x - \frac{1}{n}\sum_{j=1}^{n} \mathbb{B}_j x - \mathbb{B}_i x^\star + \frac{1}{n}\sum_{j=1}^{n} \mathbb{B}_j x^\star \right\|^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \left\| \mathbb{B}_i - \frac{1}{n}\sum_{j=1}^{n} \mathbb{B}_j \right\|^2 \|x - x^\star\|^2.\end{aligned}$$

Thus we have the simple and easy to compute upper bound

$$\delta \leq \frac{1}{n}\sum_{i=1}^{n} \left\| \mathbb{B}_i - \frac{1}{n}\sum_{j=1}^{n} \mathbb{B}_j \right\|^2 \approx 26.5. \tag{11}$$

As we can see, $\delta \ll L$.

In Figure 1, we compare SPPM with 3 different values of γ, SPPM-OC with the theoretically optimal value $\gamma = \frac{\mu}{\delta^2} \approx 10^{-3}$, L-SVRP with $p = 0.05$ and the theoretically optimal value of γ in (17). As predicted by the theory, SPPM converges only to a neighborhood of the solution, whose size is larger if γ is larger. SPPM-OC is faster than L-SVRP, but its per-iteration cost is much higher, as we detail in Figure 2.

In Figure 2, we compare the variance-reduced algorithms SPPM-OC, L-SVRP with different values of p, and Point-SAGA. SPPM-OC and L-SVRP with $p = 1$ are identical. We show convergence with respect to the number of operator calls, counting 1 for a call to an A_ξ or its resolvent, and n for a call to A in L-SVRP. As a result, L-SVRP with $p = 0.1$ and Point-SAGA perform best. We should keep in mind that L-SVRP does a full pass over the n operators with a small probability, whereas Point-SAGA requires memory storage of size n times the dimension of x. Thus, the best algorithm depends on the problem at hand.

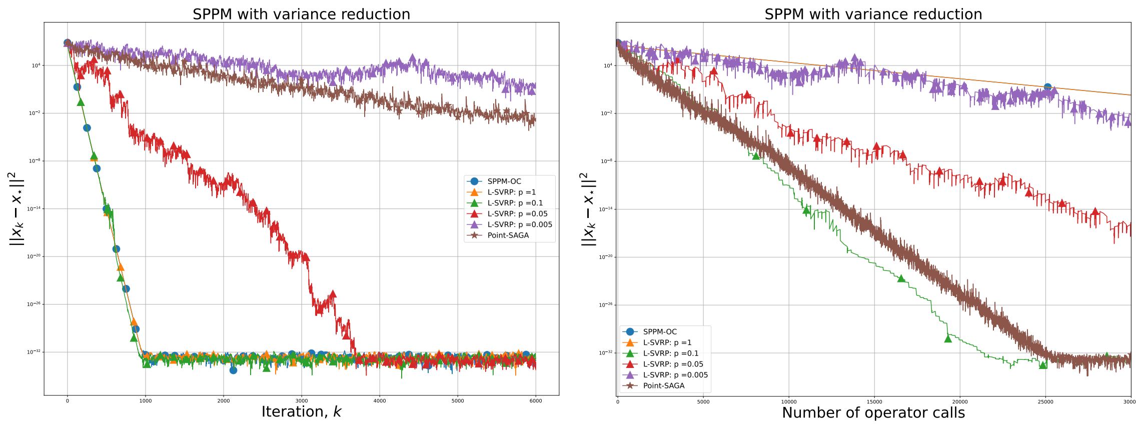

Figure 2: Performance comparison of SPPM-OC, L-SVRP with different values of $p = 1, 0.1, 0.05, 1/n = 0.005$, and Point-SAGA. The theoretically optimal value of γ is chosen in all cases. The error is shown with respect to the number of iterations on the left, and the number of operator calls on the right.

Appendix C. Proof of Lemma 1

Let $(x, y) \in \mathcal{X}^2$. From the definition of the resolvent, we have $x - x^+ \in B(x^+)$ and $y - y^+ \in B(y^+)$. Then it follows from (4) that

$$\mu\left\|x^+ - y^+\right\|^2 \leq \left\langle (x - x^+) - (y - y^+), x^+ - y^+ \right\rangle = \left\langle x - y, x^+ - y^+ \right\rangle - \left\|x^+ - y^+\right\|^2.$$

Therefore

$$(1+\mu)\left\|x^+ - y^+\right\|^2 \leq \left\langle x - y, x^+ - y^+ \right\rangle \leq \|x - y\| \|x^+ - y^+\|,$$

so that

$$(1+\mu)\|x^+ - y^+\| \leq \|x - y\|.$$

Appendix D. SPPM: Convergence Analysis

Algorithm 1 Stochastic Proximal Point Method (SPPM)

Parameters: stepsize $\gamma > 0$, initial estimate $x^0 \in \mathcal{X}$
for $k = 0, 1, \ldots$ **do**
  Sample $\xi^k \sim \mathcal{D}$
  $x^{k+1} := \left(I + \gamma A_{\xi^k}\right)^{-1}\left(x^k\right)$
end for

D.1. Proof of Theorem 1

Let $k \geq 0$. We have

$$x^\star = \left(I + \gamma A_{\xi^k}\right)^{-1}\left(x^\star + \gamma a^\star_{\xi^k}\right), \tag{12}$$

so that

$$\begin{aligned}
\left\|x^{k+1}-x^{\star}\right\|^{2} &= \left\|\left(I+\gamma A_{\xi^k}\right)^{-1}(x^k)-\left(I+\gamma A_{\xi^k}\right)^{-1}\left(x^{\star}+\gamma a^{\star}_{\xi^k}\right)\right\|^{2} \\
&\overset{(5)}{\leq} \frac{1}{(1+\gamma\mu)^2}\left\|x^k-x^{\star}-\gamma a^{\star}_{\xi^k}\right\|^{2} \\
&= \frac{1}{(1+\gamma\mu)^2}\left(\left\|x^k-x^{\star}\right\|^{2}-2\gamma\left\langle a^{\star}_{\xi^k},x^k-x^{\star}\right\rangle+\gamma^2\left\|a^{\star}_{\xi^k}\right\|^{2}\right).
\end{aligned}$$

We denote by $\mathcal{F}^k$ the σ-algebra generated by the collection of random variables $(x^0,\ldots,x^k)$. Taking the expectation conditionally on $\mathcal{F}^k$, we have, using the fact that $\mathrm{E}_{\xi\sim\mathcal{D}}\left[a^{\star}_{\xi}\right]=0$,

$$\begin{aligned}
\mathrm{E}\left[\left\|x^{k+1}-x^{\star}\right\|^{2}\mid\mathcal{F}^k\right] &\leq \frac{1}{(1+\gamma\mu)^2}\left(\left\|x^k-x^{\star}\right\|^{2}-2\gamma\left\langle\underbrace{\mathrm{E}\left[a^{\star}_{\xi^k}\mid\mathcal{F}^k\right]}_{0},x^k-x^{\star}\right\rangle\right) \\
&\quad+\frac{\gamma^2}{(1+\gamma\mu)^2}\mathrm{E}\left[\left\|a^{\star}_{\xi^k}\right\|^{2}\mid\mathcal{F}^k\right] \\
&\overset{(8)}{=} \frac{1}{(1+\gamma\mu)^2}\left\|x^k-x^{\star}\right\|^{2}+\frac{\gamma^2\sigma_{\star}^2}{(1+\gamma\mu)^2}.
\end{aligned}$$

By unrolling the recursion, we obtain

$$\begin{aligned}
\mathrm{E}\left[\left\|x^{k}-x^{\star}\right\|^{2}\right] &\leq \left(\frac{1}{1+\gamma\mu}\right)^{2k}\left\|x^0-x^{\star}\right\|^2+\sum_{l=0}^{k-1}\left(\frac{1}{1+\gamma\mu}\right)^{2l}\frac{\gamma^2\sigma_{\star}^2}{(1+\gamma\mu)^2} \\
&= \left(\frac{1}{1+\gamma\mu}\right)^{2k}\left\|x^0-x^{\star}\right\|^2+\frac{(1+\gamma\mu)^2-(1+\gamma\mu)^{2(1-k)}}{(1+\gamma\mu)^2-1}\frac{\gamma^2\sigma_{\star}^2}{(1+\gamma\mu)^2} \\
&= \left(\frac{1}{1+\gamma\mu}\right)^{2k}\left\|x^0-x^{\star}\right\|^2+\frac{1-(1+\gamma\mu)^{-2k}}{(1+\gamma\mu)^2-1}\gamma^2\sigma_{\star}^2 \\
&\leq \left(\frac{1}{1+\gamma\mu}\right)^{2k}\left\|x^0-x^{\star}\right\|^2+\frac{1}{(1+\gamma\mu)^2-1}\gamma^2\sigma_{\star}^2 \\
&= \left(\frac{1}{1+\gamma\mu}\right)^{2k}\left\|x^0-x^{\star}\right\|^2+\frac{\gamma\sigma_{\star}^2}{2\mu+\gamma\mu^2}.
\end{aligned}$$

Appendix E. SPPM-OC: Convergence Analysis

Algorithm 2 Stochastic Proximal Point Method with Operator Correction (SPPM-OC)

Parameters: stepsize $\gamma>0$, initial estimate $x^0\in\mathcal{X}$
for $k=0,1,\ldots$ **do**
  Sample $\xi^k\sim\mathcal{D}$
  Choose $a^k_{\xi^k}\in A_{\xi^k}(x^k)$ and $a^k\in A(x^k)$ so that $a^k=\mathrm{E}_{\xi\sim\mathcal{D}}\left[a^k_{\xi}\right]$
  $h^k:=a^k_{\xi^k}-a^k$
  $x^{k+1}:=\left(I+\gamma A_{\xi^k}\right)^{-1}\left(x^k+\gamma h^k\right)$
end for

Theorem 5 *In Problem (1), let Assumptions 1 and 2 hold. Then in SPPM-OC with any stepsize $\gamma > 0$ and initial estimate $x^0 \in \mathcal{X}$, we have, for every $k \geq 0$,*

$$\mathrm{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq \left(\frac{1+\gamma^2\delta^2}{(1+\gamma\mu)^2}\right)^k \|x^0 - x^\star\|^2. \tag{13}$$

Moreover, x^k converges to $x^\star$, almost surely.

The contraction factor in (13) can always be made less than 1 with γ small enough. It is minimized when $\gamma = \frac{\mu}{\delta^2}$, for which

$$\frac{1+\gamma^2\delta^2}{(1+\gamma\mu)^2} = \frac{\delta^2}{\delta^2+\mu^2} < 1.$$

With this value of γ, the iteration complexity of SPPM-OC to achieve ϵ-accuracy for any $\epsilon > 0$ is

$$\mathcal{O}\left(\left(\frac{\delta^2}{\mu^2}+1\right)\log\left(\frac{\left\|x^0 - x^\star\right\|^2}{\epsilon}\right)\right).$$

E.1. Proof of Theorem 5

For every $\xi \sim \mathcal{D}$, let $a_\xi^\star \in A_\xi(x^\star)$, such that $\mathrm{E}_{\xi\sim\mathcal{D}}\left[a_\xi^\star\right] = 0$ and Assumption 2 holds at x^k with these elements. Let $k \geq 0$. Using (12), we have

$$\begin{aligned}
\left\|x^{k+1} - x^\star\right\|^2 &= \left\|\left(I+\gamma A_{\xi^k}\right)^{-1}(x^k+\gamma h^k) - \left(I+\gamma A_{\xi^k}\right)^{-1}\left(x^\star + \gamma a_{\xi^k}^\star\right)\right\|^2 \\
&\overset{\text{Lemma 1}}{\leq} \frac{1}{(1+\gamma\mu)^2}\left\|x^k - x^\star + \gamma h^k - \gamma a_{\xi^k}^\star\right\|^2 \\
&= \frac{1}{(1+\gamma\mu)^2}\left(\left\|x^k - x^\star\right\|^2 + 2\gamma\left\langle h^k - a_{\xi^k}^\star, x^k - x^\star\right\rangle + \gamma^2\left\|h^k - a_{\xi^k}^\star\right\|^2\right).
\end{aligned}$$

We denote by $\mathcal{F}^k$ the σ-algebra generated by the collection of random variables $(x^l, a^l, a_{\xi^l}^l)_{l=0}^k$. Taking the expectation conditionally on $\mathcal{F}^k$, we have, using the fact that $\mathrm{E}_{\xi\sim\mathcal{D}}\left[a_\xi^\star\right] = 0$,

$$\begin{aligned}
\mathrm{E}\left[\left\|x^{k+1} - x^\star\right\|^2 \mid \mathcal{F}^k\right] &\leq \frac{1}{(1+\gamma\mu)^2}\left\|x^k - x^\star\right\|^2 \\
&\quad + \frac{2\gamma}{(1+\gamma\mu)^2}\left\langle \underbrace{\mathrm{E}\left[a_{\xi^k}^k - a^k - a_{\xi^k}^\star \mid \mathcal{F}^k\right]}_{0}, x^k - x^\star\right\rangle \\
&\quad + \frac{\gamma^2}{(1+\gamma\mu)^2}\mathrm{E}\left[\left\|a_{\xi^k}^k - a^k - a_{\xi^k}^\star\right\|^2 \mid \mathcal{F}^k\right] \\
&\overset{(6)}{\leq} \frac{1}{(1+\gamma\mu)^2}\left\|x^k - x^\star\right\|^2 + \frac{\gamma^2\delta^2}{(1+\gamma\mu)^2}\left\|x^k - x^\star\right\|^2 \\
&= \frac{1+\gamma^2\delta^2}{(1+\gamma\mu)^2}\left\|x^k - x^\star\right\|^2.
\end{aligned} \tag{14}$$

By unrolling the recursion, we obtain the desired result. Moreover, using classical results on supermartingale convergence [8, Proposition A.4.5], it follows from (14) that $\left\|x^k - x^\star\right\|^2 \to 0$ almost surely.

Appendix F. L-SVRP: Convergence Analysis

Algorithm 3 Loopless Stochastic Variance-Reduced Proximal Point Method (L-SVRP)

Parameters: stepsize $\gamma > 0$, initial estimates $x^0, w^0 \in \mathcal{X}$, probability $p \in (0, 1]$, $a^0 \in A(x^0)$.
for $k = 0, 1, \ldots$ **do**
  Sample $\xi^k \sim \mathcal{D}$
  Choose $a^k_{\xi^k} \in A_{\xi^k}(w^k)$ so that $\mathrm{E}_{\xi\sim\mathcal{D}}\left[a^k_\xi\right] = a^k$
  $h^k := a^k_{\xi^k} - a^k$
  $x^{k+1} := \left(I + \gamma A_{\xi^k}\right)^{-1}\left(x^k + \gamma h^k\right)$
  Flip a coin $\theta^k \in \{0, 1\}$ with $\mathrm{Prob}(\theta^k = 1) = p$.
  $w^{k+1} := \begin{cases} x^{k+1} & \text{if } \theta^k = 1 \\ w^k & \text{if } \theta^k = 0 \end{cases}$
  $a^{k+1} := \begin{cases} \text{any element in } A(x^{k+1}) & \text{if } \theta^k = 1 \\ a^k & \text{if } \theta^k = 0 \end{cases}$
end for

Theorem 6 *In Problem* (1)*, let Assumptions 1 and 2 hold. Then in* L-SVRP *with any stepsize* $\gamma > 0$*, probability* $p \in (0, 1]$*, and initial estimates* $x^0, w^0 \in \mathcal{X}$*, we have, for every* $k \geq 0$*,*

$$\mathrm{E}\left[V^k\right] \leq \max\left\{\frac{1}{1+\gamma\mu}, 1 - p + \frac{\gamma\delta^2 p}{\mu\left(1+\gamma\mu\right)}\right\}^k V^0, \tag{15}$$

where the Lyapunov function is

$$V^k := \left\|x^k - x^\star\right\|^2 + \frac{\gamma\mu}{p}\left\|w^k - x^\star\right\|^2. \tag{16}$$

Moreover, x^k *and* w^k *converge to* $x^\star$*, almost surely.*

The contraction factor in (15) can always be made less than 1 with γ small enough. It is minimized when $\frac{1}{1+\gamma\mu} = 1 - p + \frac{\gamma\delta^2 p}{\mu(1+\gamma\mu)}$. This is the case for

$$\gamma = \frac{\mu}{\delta^2 + \frac{1-p}{p}\mu^2}, \tag{17}$$

for which

$$\frac{1}{1+\gamma\mu} = 1 - p + \frac{\gamma\delta^2 p}{\mu\left(1+\gamma\mu\right)} = \frac{p\delta^2 + (1-p)\mu^2}{p\delta^2 + \mu^2} < 1.$$

With this value of γ, the iteration complexity of L-SVRP to achieve ϵ-accuracy for any $\epsilon > 0$ is

$$\mathcal{O}\left(\left(\frac{\delta^2}{\mu^2} + \frac{1}{p}\right)\log\left(\frac{V^0}{\epsilon}\right)\right).$$

F.1. Proof of Theorem 6

For every $\xi \sim \mathcal{D}$, let $a_\xi^\star \in A_\xi(x^\star)$, such that $\mathrm{E}_{\xi\sim\mathcal{D}}\left[a_\xi^\star\right] = 0$ and Assumption 2 holds at x^k with these elements. Let $k \geq 0$. Using (12), we have

$$\begin{aligned}
\left\|x^{k+1}-x^\star\right\|^2 &= \left\|\left(I+\gamma A_{\xi^k}\right)^{-1}(x^k+\gamma h^k) - \left(I+\gamma A_{\xi^k}\right)^{-1}\left(x^\star+\gamma a_{\xi^k}^\star\right)\right\|^2 \\
&\overset{\text{Lemma 1}}{\leq} \frac{1}{(1+\gamma\mu)^2}\left\|x^k-x^\star+\gamma h^k-\gamma a_{\xi^k}^\star\right\|^2 \\
&= \frac{1}{(1+\gamma\mu)^2}\left(\left\|x^k-x^\star\right\|^2 + 2\gamma\left\langle h^k-a_{\xi^k}^\star, x^k-x^\star\right\rangle + \gamma^2\left\|h^k-a_{\xi^k}^\star\right\|^2\right).
\end{aligned}$$

We denote by $\mathcal{F}^k$ the σ-algebra generated by the collection of random variables $(x^l, w^l, a^l, a_{\xi^l}^l)_{l=0}^k$. Taking the expectation conditionally on $\mathcal{F}^k$, we have, using the fact that $\mathrm{E}_{\xi\sim\mathcal{D}}\left[a_\xi^\star\right]=0$,

$$\begin{aligned}
\mathrm{E}\left[\left\|x^{k+1}-x^\star\right\|^2 \mid \mathcal{F}^k\right] &\leq \frac{1}{(1+\gamma\mu)^2}\left\|x^k-x^\star\right\|^2 \\
&\quad + \frac{2\gamma}{(1+\gamma\mu)^2}\left\langle \underbrace{\mathrm{E}\left[a_{\xi^k}^k - a^k - a_{\xi^k}^\star \mid \mathcal{F}^k\right]}_{0}, x^k-x^\star\right\rangle \\
&\quad + \frac{\gamma^2}{(1+\gamma\mu)^2}\mathrm{E}\left[\left\|a_{\xi^k}^k - a^k - a_{\xi^k}^\star\right\|^2 \mid \mathcal{F}^k\right] \\
&\overset{(6)}{\leq} \frac{1}{(1+\gamma\mu)^2}\left\|x^k-x^\star\right\|^2 + \frac{\gamma^2\delta^2}{(1+\gamma\mu)^2}\left\|w^k-x^\star\right\|^2. \qquad (18)
\end{aligned}$$

Moreover,

$$\mathrm{E}\left[\left\|w^{k+1}-x^\star\right\|^2 \mid \mathcal{F}^k\right] = (1-p)\|w^k-x^\star\|^2 + p\mathrm{E}\left[\left\|x^{k+1}-x^\star\right\|^2 \mid \mathcal{F}^k\right].$$

Let $\alpha \coloneqq \frac{\gamma\mu}{p}$. Combining the two previous inequalities and using the Lyapunov function $V^{k+1} \coloneqq \left\|x^{k+1}-x^\star\right\|^2 + \alpha\left\|w^{k+1}-x^\star\right\|^2$, we obtain

$$\begin{aligned}
\mathrm{E}\left[V^{k+1} \mid \mathcal{F}^k\right] &\leq \mathrm{E}\left[\left\|x^{k+1}-x^\star\right\|^2 \mid \mathcal{F}^k\right] + \alpha p\mathrm{E}\left[\left\|x^{k+1}-x^\star\right\|^2 \mid \mathcal{F}^k\right] \\
&\quad + (1-p)\alpha\left\|w^k-x^\star\right\|^2 \\
&= (1+\alpha p)\mathrm{E}\left[\left\|x^{k+1}-x^\star\right\|^2 \mid \mathcal{F}^k\right] + (1-p)\alpha\left\|w^k-x^\star\right\|^2 \\
&\overset{(18)}{\leq} \frac{1+\alpha p}{(1+\gamma\mu)^2}\left\|x^k-x^\star\right\|^2 + \frac{(1+\alpha p)\gamma^2\delta^2}{(1+\gamma\mu)^2}\left\|w^k-x^\star\right\|^2 + (1-p)\alpha\left\|w^k-x^\star\right\|^2 \\
&= \frac{1+\alpha p}{(1+\gamma\mu)^2}\left\|x^k-x^\star\right\|^2 + \left(1-p+\frac{(1+\alpha p)\gamma^2\delta^2}{\alpha(1+\gamma\mu)^2}\right)\alpha\left\|w^k-x^\star\right\|^2 \\
&\overset{\alpha=\frac{\gamma\mu}{p}}{\leq} \max\left\{\frac{1}{1+\gamma\mu}, 1-p+\frac{\gamma\delta^2 p}{\mu(1+\gamma\mu)}\right\} V^k. \qquad (19)
\end{aligned}$$

By unrolling the recursion, we obtain the desired result. Moreover, using classical results on supermartingale convergence [8, Proposition A.4.5], it follows from (19) that $V^k \to 0$ almost surely.

Appendix G. Point-SAGA: Convergence Analysis

Algorithm 4 Point-SAGA

Parameters: stepsize $\gamma > 0$, initial estimates x^0, $(w_i^0)_{i=1}^n \in \mathcal{X}^n$, initial elements $a_i^0 \in A_i(w_i^0)$ for every $i \in [n]$, $a^0 := \frac{1}{n}\sum_{i=1}^n a_i^0$

for $k = 0, 1, \dots$ **do**

  Sample $i^k \in [n]$ uniformly at random

  $h^k := a_{i^k}^k - a^k$

  $x^{k+1} := (I + \gamma A_{i^k})^{-1}\left(x^k + \gamma h^k\right)$

  $w_j^{k+1} := \begin{cases} x^{k+1} & \text{for } j = i^k \\ w_j^k & \text{for every } j \in [n] \backslash \{i^k\} \end{cases}$ // not stored, defined only for the analysis

  $a_j^{k+1} := \begin{cases} \text{any element in } A_{i^k}(x^{k+1}) & \text{for } j = i^k \\ a_j^k & \text{for every } j \in [n] \backslash \{i^k\} \end{cases}$ // e.g. $a_{i^k}^{k+1} := \frac{1}{\gamma}(x^k - x^{k+1}) + h^k$

  $a^{k+1} := a^k + \frac{1}{n}(a_{i^k}^{k+1} - a_{i^k}^k)$ // $= \frac{1}{n}\sum_{j=1}^n a_j^{k+1}$

end for

Theorem 7 *In Problem (2), let Assumptions 1 and 3 hold. Then in Point-SAGA with any stepsize $\gamma > 0$, initial estimates x^0, $(w_i^0)_{i=1}^n \in \mathcal{X}^n$ and elements $a_i^0 \in A_i(w_i^0)$, we have, for every $k \geq 0$,*

$$\mathrm{E}\left[V^k\right] \leq \max\left\{\frac{1}{1+\gamma\mu}, 1 - \frac{1}{n} + \frac{\gamma\tilde{\delta}^2}{n\mu(1+\gamma\mu)}\right\}^k V^0, \tag{20}$$

where the Lyapunov function is

$$V^k := \left\|x^k - x^\star\right\|^2 + \gamma\mu\sum_{i=1}^n \left\|w_i^k - x^\star\right\|^2. \tag{21}$$

Moreover, x^k and all w_i^k converge to $x^\star$, almost surely.

The contraction factor in (20) can always be made less than 1 with γ small enough. It is minimized when $\frac{1}{1+\gamma\mu} = 1 - \frac{1}{n} + \frac{\gamma\tilde{\delta}^2}{n\mu(1+\gamma\mu)}$. This is the case for

$$\gamma = \frac{\mu}{\tilde{\delta}^2 + (n-1)\mu^2},$$

for which

$$\frac{1}{1+\gamma\mu} = 1 - \frac{1}{n} + \frac{\gamma\tilde{\delta}^2}{n\mu(1+\gamma\mu)} = \frac{\tilde{\delta}^2 + (n-1)\mu^2}{\tilde{\delta}^2 + n\mu^2} < 1.$$

With this value of γ, the iteration complexity of Point-SAGA to achieve ϵ-accuracy for any $\epsilon > 0$ is

$$\mathcal{O}\left(\left(\frac{\delta^2}{\mu^2} + n\right)\log\left(\frac{V^0}{\epsilon}\right)\right).$$

G.1. Proof of Theorem 7

For every $i \in [n]$, let $a_i^\star \in A_i(x^\star)$, such that $\frac{1}{n}\sum_{i=1}^n a_i^\star = 0$ and Assumption 3 holds at the $(w_i^k)_{i=1}^n$ with these elements. Let $k \geq 0$. We have

$$x^\star = (I + \gamma A_{i^k})^{-1}\left(x^\star + \gamma a_{i^k}^\star\right),$$

so that

$$\begin{aligned}
\left\|x^{k+1} - x^\star\right\|^2 &= \left\|(I + \gamma A_{i^k})^{-1}(x^k + \gamma h^k) - (I + \gamma A_{i^k})^{-1}\left(x^\star + \gamma a_{i^k}^\star\right)\right\|^2 \\
&\overset{\text{Lemma 1}}{\leq} \frac{1}{(1+\gamma\mu)^2}\left\|x^k - x^\star + \gamma h^k - \gamma a_{i^k}^\star\right\|^2 \\
&= \frac{1}{(1+\gamma\mu)^2}\left(\left\|x^k - x^\star\right\|^2 + 2\gamma\left\langle h^k - a_{i^k}^\star, x^k - x^\star\right\rangle + \gamma^2\left\|h^k - a_{i^k}^\star\right\|^2\right).
\end{aligned}$$

We denote by $\mathcal{F}^k$ the σ-algebra generated by the collection of random variables $\left(x^l, (w_i^l)_{i=1}^n, (a_i^l)_{i=1}^n\right)_{l=0}^k$. Taking the expectation conditionally on $\mathcal{F}^k$, we have

$$\begin{aligned}
\mathrm{E}\left[\left\|x^{k+1} - x^\star\right\|^2 \mid \mathcal{F}^k\right] &\leq \frac{1}{(1+\gamma\mu)^2}\left\|x^k - x^\star\right\|^2 \\
&\quad + \frac{\gamma^2}{(1+\gamma\mu)^2}\mathrm{E}\left[\left\|a_{i^k}^k - a^k - a_{i^k}^\star\right\|^2 \mid \mathcal{F}^k\right] \\
&\quad + \frac{2\gamma}{(1+\gamma\mu)^2}\left\langle \underbrace{\mathrm{E}\left[a_{i^k}^k - a^k - a_{i^k}^\star \mid \mathcal{F}^k\right]}_{0}, x^k - x^\star\right\rangle \\
&= \frac{1}{(1+\gamma\mu)^2}\left\|x^k - x^\star\right\|^2 + \frac{\gamma^2}{n(1+\gamma\mu)^2}\sum_{i=1}^n\left\|a_i^k - a^k - a_i^\star\right\|^2 \\
&\overset{(7)}{\leq} \frac{1}{(1+\gamma\mu)^2}\left\|x^k - x^\star\right\|^2 + \frac{\gamma^2\tilde{\delta}^2}{n(1+\gamma\mu)^2}\sum_{i=1}^n\left\|w_i^k - x^\star\right\|^2. \qquad (22)
\end{aligned}$$

Moreover,

$$\frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[\left\|w_i^{k+1} - x^\star\right\|^2 \mid \mathcal{F}^k\right] = \left(1 - \frac{1}{n}\right)\frac{1}{n}\sum_{i=1}^n\left\|w_i^k - x^\star\right\|^2 + \frac{1}{n}\mathrm{E}\left[\left\|x^{k+1} - x^\star\right\|^2 \mid \mathcal{F}^k\right].$$

Let $\alpha \coloneqq n\gamma\mu$. Combining the two previous inequalities and using the Lyapunov function $V^{k+1} \coloneqq \left\|x^{k+1} - x^\star\right\|^2 + \frac{\alpha}{n}\sum_{i=1}^n \left\|w_i^{k+1} - x^\star\right\|^2$, we obtain

$$
\begin{aligned}
\mathrm{E}\left[V^{k+1} \mid \mathcal{F}^k\right] &\leq \mathrm{E}\left[\left\|x^{k+1} - x^\star\right\|^2 \mid \mathcal{F}^k\right] + \frac{\alpha}{n}\mathrm{E}\left[\left\|x^{k+1} - x^\star\right\|^2 \mid \mathcal{F}^k\right] \\
&\quad + \left(1 - \frac{1}{n}\right)\frac{\alpha}{n}\sum_{i=1}^n \left\|w_i^k - x^\star\right\|^2 \\
&= \left(1 + \frac{\alpha}{n}\right)\mathrm{E}\left[\left\|x^{k+1} - x^\star\right\|^2 \mid \mathcal{F}^k\right] + \left(1 - \frac{1}{n}\right)\frac{\alpha}{n}\sum_{i=1}^n \left\|w_i^k - x^\star\right\|^2 \\
&\overset{(22)}{\leq} \frac{1 + \alpha/n}{(1+\gamma\mu)^2}\left\|x^k - x^\star\right\|^2 + \frac{(1+\alpha/n)\gamma^2\tilde{\delta}^2}{(1+\gamma\mu)^2}\frac{1}{n}\sum_{i=1}^n \left\|w_i^k - x^\star\right\|^2 \\
&\quad + \left(1 - \frac{1}{n}\right)\frac{\alpha}{n}\sum_{i=1}^n \left\|w_i^k - x^\star\right\|^2 \\
&= \frac{1 + \alpha/n}{(1+\gamma\mu)^2}\left\|x^k - x^\star\right\|^2 + \left(1 - \frac{1}{n} + \frac{(1+\alpha/n)\gamma^2\tilde{\delta}^2}{\alpha(1+\gamma\mu)^2}\right)\frac{\alpha}{n}\sum_{i=1}^n \left\|w_i^k - x^\star\right\|^2 \\
&\overset{\alpha = n\gamma\mu}{\leq} \max\left\{\frac{1}{1+\gamma\mu}, 1 - \frac{1}{n} + \frac{\gamma\tilde{\delta}^2}{n\mu(1+\gamma\mu)}\right\} V^k. \qquad (23)
\end{aligned}
$$

By unrolling the recursion, we obtain the desired result. Moreover, using classical results on supermartingale convergence [8, Proposition A.4.5], it follows from (23) that $V^k \to 0$ almost surely.