OPTIMIZATION ERUDITORUM
Volume 3 (2026), No. 3, 183–208
DOI:https://doi.org/10.69829/oper-026-0303-ta04
Tulipa Opera Scholarum

DOUBLE INERTIAL STOCHASTIC RELAXED FORWARD-BACKWARD-FORWARD ALGORITHM

AVIV GIBALI[1], GRACE NNENNAYA OGWO[2], AND YEKINI SHEHU[3,*]

[1] *Department of Applied Mathematics, Holon Institute of Technology-HIT, 5810201, Holon, Israel*
[2] *Department of Mathematics, Shanghai Normal University, 100 Guilin Road, Shanghai 200234*
[3] *School of Mathematical Sciences, Zhejiang Normal University, Jinhua 321004, China*

Abstract. We propose the double inertial stochastic relaxed forward-backward-forward (DI-SRIFBF) algorithm for solving monotone inclusions in Hilbert spaces. The method augments the stochastic FBF framework with two sequential inertial extrapolation steps, and it evaluates the single-valued operator via independent mini-batches at each forward step. This design avoids the correlation issues that arise when the same batch is reused, and it yields a stochastic conditioning that fits the Robbins–Siegmund template without hidden measurability assumptions. Under explicit, algebraically verifiable step-size restrictions that hold for every relaxation parameter $\rho \in (0, 2)$, we establish almost-sure weak convergence provided the inertial parameters and mini-batch variances are summable. We also prove an $O(1/k)$ rate for the discrete velocity and, under strong monotonicity with geometrically growing batch sizes, linear convergence. Choosing the growth factor $\tau = (1 - \eta/2)^{-1}$ gives an $\mathcal{O}(1/\varepsilon)$ oracle complexity. Finally, we derive explicit non-asymptotic bounds for biased stochastic oracles, showing that the iterates settle into an $O(B^2/\mu^2)$ neighborhood of the unique solution. Numerical experiments on two-stage stochastic variational inequalities and group sparse learning confirm the theoretical predictions and demonstrate a practical advantage over single-inertial and non-inertial benchmarks.

Keywords. Double inertia; stochastic forward-backward-forward splitting; monotone inclusion; variational inequality; inertial algorithm; biased stochastic oracle; almost sure convergence; linear convergence.

1. Introduction

Let $\mathcal{H}$ be a separable real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. We consider the stochastic monotone inclusion problem

$$\text{Find } x^* \in \mathcal{H} \text{ such that } 0 \in (A + V)(x^*), \tag{1.1}$$

where $A : \mathcal{H} \to 2^{\mathcal{H}}$ is maximally monotone, $V(x) = \mathbb{E}_\xi[V_\xi(x)]$ is monotone and L-Lipschitz continuous, and $\xi \sim P$. The solution set $S = \operatorname{zer}(A + V)$ is assumed nonempty. Problem (1.1) is ubiquitous in optimization, variational analysis, and machine learning: it captures convex–concave saddle-point problems, composite convex minimization, and equilibrium models in stochastic programming and game theory [3, 9, 22].

*Corresponding author.
E-mail address: avivgi@hit.ac.il (A. Gibali), graceogwo@aims.ac.za (G. N. Ogwo), and yekini.shehu@zjnu.edu.cn (Y. Shehu)

2020 Mathematics Subject Classification: 47H05, 47J20, 65K15, 90C15, 90C25.
Accepted: July 20, 2026.

Deterministic template. When V is available exactly, a classical algorithm for (1.1) is Tseng's forward-backward-forward (FBF) method [24]. At each iteration, FBF performs a forward step, a resolvent (backward) step, and a second forward correction that requires no additional operator evaluations beyond the two points already computed. The method refines the standard forward-backward splitting by removing the cocoercivity assumption on V, and it has found renewed interest in adversarial training and generative modeling [10, 6]. Another closely related scheme is Korpelevich's extragradient method [16], which uses two projected gradient steps per iteration and has been the template for a vast literature on variational inequalities.

Inertial acceleration. A separate line of research seeks to accelerate first-order methods by *inertial* (or *momentum*) extrapolation. The idea traces back to Polyak's heavy-ball method [21] for smooth minimization, where momentum was shown to improve the local convergence rate. For maximal monotone inclusions, Alvarez and Attouch [1] introduced an inertial proximal algorithm by discretizing a nonlinear oscillator with damping, and proved weak convergence under a summability condition on the inertial parameter. Subsequent works extended this framework to forward–backward splitting [18, 20], Douglas–Rachford splitting [5], and primal–dual algorithms [4]. Attouch and Cabot [2] later proposed *relaxed* inertial proximal methods that combine extrapolation with over-relaxation, obtaining faster convergence in practice while preserving theoretical guarantees. Maingé [19] introduced a correction term in the inertial step that allows larger momentum parameters without sacrificing monotonicity of the distance to solutions.

Stochastic settings. In large-scale and data-driven applications, evaluating V exactly is prohibitively expensive because V is an expectation over a high-dimensional or continuous probability space. Stochastic approximation schemes replace the exact operator by a mini-batch estimator $\mathcal{V}_k$, drawing i.i.d. samples $\xi_{k,i} \sim P$. Juditsky, Nemirovski and Tauvel [15] developed the stochastic mirror-prox algorithm with $O(1/k)$ rate for monotone variational inequalities. Iusem *et al.* [14] combined the extragradient idea with variance reduction (SVRG-style) and proved linear convergence under strong monotonicity. More recently, Kotsalis, Lan and Li [17] proposed operator-extrapolation methods that achieve order-optimal oracle complexity for merely monotone problems, while Hsieh *et al.* [12, 13] analyzed single-call stochastic extragradient schemes with adaptive step-size scaling. Gorbunov *et al.* [11] gave a unified analysis of stochastic extragradient methods under expected cocoercivity. In the specific context of FBF-type splitting, Cui *et al.* [7] proposed the stochastic relaxed inertial forward-backward-forward (SRIFBF) algorithm and established almost-sure weak convergence under summable inertial parameters and mini-batch variances.

Double inertia. All the aforementioned inertial schemes use a *single* extrapolation step. Empirically, however, multi-step memory often yields superior acceleration, especially on ill-conditioned or stiff problems. In the deterministic setting, Wang *et al.* [25] introduced a double inertial forward–backward–forward method with adaptive step-size for quasi-monotone variational inequalities, showing that two sequential extrapolations can outperform single inertia. Thammasiri *et al.* [23] developed a double Tseng algorithm with inertial terms for image deblurring. These works suggest that the second inertial step provides additional "inertia" that helps the trajectory traverse flat regions and narrow valleys more efficiently. Nevertheless, to our knowledge no stochastic variant of double inertial FBF has been analyzed in the literature, and the interaction between two coupled extrapolation steps and mini-batch noise is nontrivial.

Biased oracles. A practical concern often overlooked in the theoretical literature is *bias*: stochastic gradients may be corrupted by systematic errors (compression, quantization, surrogate models, or finite-difference approximations). Driggs, Liang and Schönlieb [8] highlighted that biased stochastic gradient estimates arise naturally in many engineering pipelines and can prevent convergence to the exact solution if not properly accounted for. Quantitative bounds that explicitly track the bias are therefore essential for reliable deployment.

Gaps and our contributions. Motivated by the above developments, we introduce and analyze the **double inertial stochastic relaxed forward-backward-forward (DI-SRIFBF)** algorithm. Our work addresses four specific gaps in the existing literature:

(1) **Algorithm.** We extend the SRIFBF framework of Cui *et al.* [7] to *double* sequential inertial extrapolations:

$$p_k = x_k + \beta_k(x_k - x_{k-1}), \qquad y_k = p_k + \gamma_k(p_k - p_{k-1}),$$

thereby generalizing both single-inertial FBF and deterministic double-inertial methods to the stochastic setting.

(2) **Stochastic conditioning.** We use *independent mini-batches* for the two forward evaluations (at y_k and at z_k). While independent samples are standard in stochastic mirror-prox and extragradient methods [15, 12], their integration with double inertia and the resolvent-based FBF step is new and requires a distinct conditioning argument that avoids the correlation issues that plague same-sample estimators. This leads to a Robbins–Siegmund argument without hidden measurability assumptions.

(3) **Explicit and satisfiable parameter restrictions.** We derive a closed-form step-size bound

$$\lambda \leq \frac{1}{L}\Big(\sqrt{(1-\rho)^2 + \tfrac{3(2-\rho)}{4}} - |1-\rho|\Big),$$

valid for *every* relaxation parameter $\rho \in (0,2)$. Under strong monotonicity we give a *joint* condition on (λ, ρ, μ, L) that is algebraically verifiable and does not rely on asymptotic hand-waving.

(4) **Biased-oracle analysis.** We provide explicit non-asymptotic bounds for oracles with bounded bias B, showing that the iterates converge to an $O(B^2/\mu^2)$ neighborhood of the solution, with constants tracked exactly.

Technical highlights. The analysis rests on a carefully constructed Lyapunov function

$$\Phi_k = \|x_k - x^*\|^2 + A_k\|x_k - x_{k-1}\|^2 + B_k\|p_k - p_{k-1}\|^2,$$

with recursively defined weights A_k, B_k that absorb the inertial memory terms. Because the double inertial step generates cross-correlations between $x_k - x_{k-1}$ and $p_k - p_{k-1}$, the standard single-inertial Lyapunov is insufficient; our weights are designed so that the telescoping inequalities hold for all k and all $\rho \in (0,2)$. The strong-monotonicity analysis uses the resolvent identity to extract a contraction factor $1-\eta$ with $\eta = \rho\lambda\mu/2$, while the mini-batch variance is controlled by geometric growth $m_{k+1} = \lceil \tau m_k \rceil$ with $\tau \geq (1-\eta/2)^{-1}$.

Organization. The rest of the paper follows a natural arc from definitions to algorithm to analysis to numerics. Section 2 fixes the stochastic notation and recalls the operator-theoretic background. Section 3 presents DI-SRIFBF, explains why independent mini-batches are essential, and relates the method to existing templates. The core convergence analysis occupies Section 4: we first establish a deterministic quasi-nonexpansivity estimate, then derive the fundamental stochastic inequality, build the Lyapunov function with recursive weights, and prove almost-sure weak convergence. Section 5 establishes an $O(1/k)$ rate for the discrete velocity, Section 6 proves linear convergence and optimal

oracle complexity under strong monotonicity, and Section 7 gives explicit bias neighborhoods for corrupted oracles. Section 8 validates the theory on two-stage stochastic variational inequalities and group sparse learning. Final summaries and future directions are given in Section 9. Appendix A contains the complete algebraic verification of the telescoping Lyapunov inequalities.

2. Preliminaries

Throughout, $\mathcal{H}$ denotes a separable real Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. An operator $A:\mathcal{H}\to 2^{\mathcal{H}}$ is maximally monotone if its graph is monotone and maximal with respect to inclusion; it is μ-strongly monotone ($\mu>0$) if $\langle x-y,u-v\rangle\geq\mu\|x-y\|^2$ for all $(x,u),(y,v)\in\operatorname{gra}A$. A single-valued operator V is L-Lipschitz if $\|Vx-Vy\|\leq L\|x-y\|$ for all $x,y\in\mathcal{H}$. The resolvent $J_{\lambda A}=(\mathrm{Id}+\lambda A)^{-1}$ is firmly nonexpansive, i.e.

$$\|J_{\lambda A}x-J_{\lambda A}y\|^2\leq\|x-y\|^2-\|(\mathrm{Id}-J_{\lambda A})x-(\mathrm{Id}-J_{\lambda A})y\|^2.$$

We shall also need the following classical tool.

Lemma 2.1 (Opial's lemma)**.** *Let $\{x_n\}$ be a sequence in $\mathcal{H}$ such that (i) $\lim_{n\to\infty}\|x_n-x^*\|$ exists for every $x^*\in S$, and (ii) every weak sequential cluster point of $\{x_n\}$ belongs to S. Then $x_n\rightharpoonup p$ for some $p\in S$.*

Stochastic framework. Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space. At each iteration k the algorithm draws two independent mini-batches of i.i.d. samples: $\xi_{k,1},\dots,\xi_{k,m_k}\sim P$ for the forward step at y_k, and $\xi'_{k,1},\dots,\xi'_{k,m_k}\sim P$ for the forward step at z_k. Both batches are independent of the past. The filtration is defined by

$$\mathcal{F}_k=\sigma\big(\xi_{0,1},\dots,\xi_{0,m_0},\xi'_{0,1},\dots,\xi'_{0,m_0},\dots,\xi_{k,1},\dots,\xi_{k,m_k},\xi'_{k,1},\dots,\xi'_{k,m_k}\big),\tag{2.1}$$

so that $\mathcal{F}_0\subseteq\mathcal{F}_1\subseteq\cdots$. An $\mathcal{H}$-valued random variable is $\mathcal{F}_{k-1}$-measurable precisely when it depends only on iterates and samples from iterations $0,\dots,k-1$.

Our standing assumption on the stochastic oracle is the following.

Assumption 2.2 (Stochastic oracle)**.** *For every $x\in\mathcal{H}$: (i) $\mathbb{E}[V_\xi(x)]=V(x)$; (ii) $\mathbb{E}[\|V_\xi(x)-V(x)\|^2]\leq\sigma^2$; (iii) V is L-Lipschitz. The samples $\xi_{k,i}$ and $\xi'_{k,i}$ are i.i.d. with distribution P and independent of $\mathcal{F}_{k-1}$.*

Because the first mini-batch is independent of $\mathcal{F}_{k-1}$ and y_k is $\mathcal{F}_{k-1}$-measurable, the estimator

$$\mathcal{V}_k=\frac{1}{m_k}\sum_{i=1}^{m_k}V_{\xi_{k,i}}(y_k)$$

is conditionally unbiased: $\mathbb{E}[\mathcal{V}_k\mid\mathcal{F}_{k-1}]=V(y_k)$. Likewise, since the second batch is independent of $(z_k,\mathcal{F}_{k-1})$, the estimator

$$\mathcal{W}_k=\frac{1}{m_k}\sum_{i=1}^{m_k}V_{\xi'_{k,i}}(z_k)$$

satisfies $\mathbb{E}[\mathcal{W}_k\mid z_k,\mathcal{F}_{k-1}]=V(z_k)$. Introducing the error terms

$$e_k^y:=\mathcal{V}_k-V(y_k),\qquad e_k^z:=\mathcal{W}_k-V(z_k),$$

the tower property yields $\mathbb{E}[e_k^z\mid\mathcal{F}_{k-1}]=0$, while the independence of the first batch gives $\mathbb{E}[e_k^y\mid\mathcal{F}_{k-1}]=0$. Setting $\delta_k:=\sigma^2/m_k$, the standard variance bounds follow:

$$\mathbb{E}[\|e_k^y\|^2\mid\mathcal{F}_{k-1}]\leq\delta_k,\qquad\mathbb{E}[\|e_k^z\|^2\mid\mathcal{F}_{k-1}]\leq\delta_k,\qquad\mathbb{E}[\|e_k^y-e_k^z\|^2\mid\mathcal{F}_{k-1}]\leq 4\delta_k.\tag{2.2}$$

Remark 2.3 (Measurability). In Algorithm 1 below, the vectors x_k,p_k,y_k are $\mathcal{F}_{k-1}$-measurable, whereas $z_k,\mathcal{V}_k$ are measurable with respect to $\mathcal{F}_{k-1}\vee\sigma(\xi_{k,1},\dots,\xi_{k,m_k})$, and x_{k+1} is $\mathcal{F}_k$-measurable. Because the second batch is drawn independently of the first, both error terms are conditionally zero-mean.

Inequalities. We recall two classical tools. Young's inequality states that $2\langle x, y\rangle \le \varepsilon\|x\|^2 + \varepsilon^{-1}\|y\|^2$ for all $\varepsilon > 0$ and $x, y \in \mathcal{H}$. The Robbins–Siegmund lemma reads as follows.

Lemma 2.4 (Robbins–Siegmund). *Let $\{u_n\}, \{v_n\}, \{w_n\}, \{\gamma_n\}$ be nonnegative $\mathcal{F}_n$-measurable with $\sum \gamma_n < \infty$ a.s. and $\mathbb{E}[u_{n+1} \mid \mathcal{F}_n] \le (1+\gamma_n)u_n - v_n + w_n$ a.s. Then $\{u_n\}$ converges a.s. to a finite random variable and $\sum v_n < \infty$ a.s. on $\{\sum w_n < \infty\}$.*

3. The DI-SRIFBF Algorithm

With the preliminaries in place, we now describe the method. Each iteration of DI-SRIFBF consists of three conceptual phases: (i) a double inertial extrapolation that builds momentum from the two most recent iterates; (ii) a stochastic forward-backward step that computes a proximal point using an independent mini-batch; and (iii) a second stochastic forward evaluation at the proximal point, again with a fresh independent mini-batch, followed by a relaxed update. The use of two separate batches is deliberate: it guarantees that the error in the second forward step is conditionally zero-mean, which is essential for the Robbins–Siegmund argument in Section 4.

Algorithm 1 DI-SRIFBF

1: Choose $x_{-1}, x_0 \in \mathcal{H}$. Set $p_{-1} = x_{-1}$.
2: Choose $\lambda_k > 0$, $\rho_k \in (0,2)$, $\beta_k, \gamma_k \in [0,1)$, $m_k \in \mathbb{N}$.
3: **for** $k = 0, 1, 2, \ldots$ **do**
4: $\quad p_k = x_k + \beta_k(x_k - x_{k-1})$
5: $\quad y_k = p_k + \gamma_k(p_k - p_{k-1})$
6: $\quad$ Draw i.i.d. samples $\xi_{k,1}, \ldots, \xi_{k,m_k} \sim P$ independent of $\mathcal{F}_{k-1}$.
7: $\quad \mathcal{V}_k = \frac{1}{m_k}\sum_{i=1}^{m_k} V_{\xi_{k,i}}(y_k)$
8: $\quad z_k = J_{\lambda_k A}(y_k - \lambda_k \mathcal{V}_k)$
9: $\quad$ Draw i.i.d. samples $\xi'_{k,1}, \ldots, \xi'_{k,m_k} \sim P$ independent of the first batch and of $\mathcal{F}_{k-1}$.
10: $\quad \mathcal{W}_k = \frac{1}{m_k}\sum_{i=1}^{m_k} V_{\xi'_{k,i}}(z_k)$ {independent batch}
11: $\quad x_{k+1} = \rho_k(z_k - \lambda_k(\mathcal{W}_k - \mathcal{V}_k)) + (1-\rho_k)y_k$
12: **end for**

Remark 3.1 (Initialization). The initialization $p_{-1} = x_{-1}$ is required because the second inertial step 5 at $k = 0$ uses $p_0 - p_{-1}$. Without this convention y_0 would depend on an undefined quantity. The same convention is standard in the deterministic double-inertial literature [25, 23].

Remark 3.2 (Double inertial structure). The extrapolation steps 4–5 can be rewritten in terms of the raw iterates $\{x_k\}$ as

$$y_k = x_k + \big[\beta_k + \gamma_k(1+\beta_k)\big](x_k - x_{k-1}) - \gamma_k\beta_{k-1}(x_{k-1} - x_{k-2}). \tag{3.1}$$

When $\gamma_k \equiv 0$ this reduces to the classical single inertial step $y_k = x_k + \beta_k(x_k - x_{k-1})$. The additional term $-\gamma_k\beta_{k-1}(x_{k-1} - x_{k-2})$ is a two-step memory correction: it uses the curvature information encoded in the difference between the two previous inertial directions. Empirically, this second level of inertia helps trajectories traverse flat regions and narrow valleys more efficiently than single inertia, especially when the operator is stiff or the conditioning is poor.

Remark 3.3 (Independent mini-batches and stochastic conditioning). Steps 7 and 10 use two *independent* mini-batches. This design is not merely cosmetic: if one reused the same batch for both $\mathcal{V}_k$ and $\mathcal{W}_k$, then z_k would be correlated with the noise in $\mathcal{W}_k$, and the conditional expectation $\mathbb{E}[e_k^z \mid \mathcal{F}_{k-1}]$ would fail to vanish. The independence guarantees that $\mathbb{E}[\mathcal{W}_k \mid z_k, \mathcal{F}_{k-1}] = V(z_k)$, whence $\mathbb{E}[e_k^z \mid \mathcal{F}_{k-1}] = 0$ by the tower property. This is the crucial ingredient that makes the Robbins–Siegmund argument in

Lemma 4.7 rigorous without hidden measurability assumptions. The same-sample issue was overlooked in several early stochastic FBF proposals; correcting it is one of the technical contributions of this paper.

Remark 3.4 (Relaxation and step-size roles). The relaxation parameter $\rho_k \in (0,2)$ in step 11 controls the weight placed on the FBF correction $z_k - \lambda_k(\mathcal{W}_k - \mathcal{V}_k)$ relative to the extrapolated point y_k. When $\rho_k = 1$ the update is the standard stochastic FBF step; when $\rho_k \in (0,1)$ it is an under-relaxation that can improve stability, and when $\rho_k \in (1,2)$ it is an over-relaxation that often accelerates convergence in practice. The step size λ_k governs the trust-region radius in the resolvent step 8. In Section 4 we derive explicit restrictions on the pair (λ, ρ) that guarantee convergence for every $\rho \in (0,2)$.

Remark 3.5 (Relation to existing methods). Algorithm 1 generalizes three existing templates:

(1) **Deterministic FBF** [24]: obtained by setting $\beta_k = \gamma_k \equiv 0$, $m_k \equiv 1$, and replacing the stochastic estimators by the exact operator V.
(2) **SRIFBF** [7]: obtained by setting $\gamma_k \equiv 0$ and using a single mini-batch (though, as noted above, the same-sample variant requires careful conditioning).
(3) **Deterministic double-inertial FBF** [25]: obtained by setting $m_k \equiv 1$ and replacing the stochastic estimators by V.

To our knowledge, Algorithm 1 is the first method that combines double inertia with rigorous stochastic conditioning via independent mini-batches.

Remark 3.6 (Computational cost). Each iteration requires two mini-batch forward evaluations (at y_k and at z_k) and one resolvent evaluation. The cost is therefore comparable to that of standard stochastic extragradient or FBF methods: the double inertia adds only vector additions and scalar multiplications, which are negligible relative to the forward and backward operator evaluations. In the numerical experiments of Section 8 the per-iteration CPU time is approximately 1.5 ms across all compared methods, confirming that the inertial overhead is not a practical bottleneck.

4. Convergence Analysis

Throughout this section we work with constant step size and relaxation,

$$\lambda_k \equiv \lambda, \qquad \rho_k \equiv \rho \in (0,2), \qquad \underline{\rho} := \inf_k \rho_k(2-\rho_k) = \rho(2-\rho) > 0, \qquad \bar{\rho} := \sup_k \rho_k = \rho < 2, \tag{4.1}$$

and we impose the explicit, algebraically verifiable restriction

$$0 < \lambda \le \lambda_{\max} := \frac{1}{L}\Big(\sqrt{(1-\rho)^2 + \frac{3(2-\rho)}{4}} - |1-\rho|\Big). \tag{4.2}$$

For $\rho = 1$ this gives $\lambda \le \sqrt{3}/(2L) \approx 0.866/L$; for $\rho = 0.8$ it gives $\lambda \le (\sqrt{0.94} - 0.2)/L \approx 0.770/L$. The bound is most permissive as $\rho \to 0$ and strictest as $\rho \to 2$, which is natural because over-relaxation near the boundary requires vanishing step sizes. We also define the stochastic error terms

$$e_k^y = \mathcal{V}_k - V(y_k), \qquad e_k^z = \mathcal{W}_k - V(z_k).$$

Deterministic and stochastic estimates. We first isolate the contraction properties of the exact FBF map. The following quasi-nonexpansivity estimate is standard; we include the short proof for completeness.

Lemma 4.1 (Deterministic FBF quasi-nonexpansivity)**.** *Let $y \in \mathcal{H}$, $z = J_{\lambda A}(y - \lambda V(y))$, and $x_+ = z - \lambda(V(z) - V(y))$. Then for any $x^* \in S$,*

$$\|x_+ - x^*\|^2 \le \|y - x^*\|^2 - (1 - \lambda^2 L^2)\|z - y\|^2. \tag{4.3}$$

Proof. Since $z = J_{\lambda A}(y - \lambda V(y))$, we have $(y - z)/\lambda - V(y) \in A(z)$. Because $0 \in (A + V)(x^*)$, monotonicity of A gives

$$\left\langle z - x^*, \frac{y - z}{\lambda} - V(y) + V(x^*) \right\rangle \geq 0,$$

i.e. $\langle z - x^*, y - z\rangle \geq \lambda\langle z - x^*, V(y) - V(x^*)\rangle$. Hence

$$\begin{aligned} \|z - x^*\|^2 &= \|y - x^*\|^2 - \|z - y\|^2 + 2\langle z - y, z - x^*\rangle \\ &= \|y - x^*\|^2 - \|z - y\|^2 - 2\langle y - z, z - x^*\rangle \\ &\leq \|y - x^*\|^2 - \|z - y\|^2 - 2\lambda\langle z - x^*, V(y) - V(x^*)\rangle. \end{aligned} \tag{4.4}$$

Now

$$\|x_+ - x^*\|^2 = \|z - x^*\|^2 - 2\lambda\langle z - x^*, V(z) - V(y)\rangle + \lambda^2\|V(z) - V(y)\|^2. \tag{4.5}$$

Substituting (4.4) into (4.5) and using $\langle z - x^*, V(z) - V(x^*)\rangle \geq 0$ (monotonicity of V) yields

$$\begin{aligned} \|x_+ - x^*\|^2 &\leq \|y - x^*\|^2 - \|z - y\|^2 - 2\lambda\langle z - x^*, V(z) - V(x^*)\rangle + \lambda^2\|V(z) - V(y)\|^2 \\ &\leq \|y - x^*\|^2 - (1 - \lambda^2 L^2)\|z - y\|^2. \end{aligned}$$

□

The stochastic iterates satisfy a similar pointwise inequality, now polluted by the mini-batch errors. Expanding the update via the three-point identity gives the following fundamental bound.

Lemma 4.2 (Fundamental inequality). *For any $x^* \in S$, the iterates of Algorithm 1 satisfy*

$$\begin{aligned} \|x_{k+1} - x^*\|^2 \leq{}& \|y_k - x^*\|^2 - \rho(2 - \rho)\|z_k - y_k\|^2 + \rho\lambda^2 L^2\|z_k - y_k\|^2 + 2\rho|1 - \rho|\lambda L\|z_k - y_k\|^2 \\ &- 2\lambda\rho\langle y_k - x^*, e_k^z - e_k^y\rangle - 2\lambda\rho^2\langle z_k - y_k, e_k^z - e_k^y\rangle \\ &+ 2\lambda^2\rho^2\langle V(z_k) - V(y_k), e_k^z - e_k^y\rangle + \lambda^2\rho^2\|e_k^y - e_k^z\|^2. \end{aligned} \tag{4.6}$$

Proof. Write the update as

$$x_{k+1} = (1 - \rho)y_k + \rho\big(z_k - \lambda(V(z_k) - V(y_k))\big) + \rho\lambda(e_k^y - e_k^z).$$

Set $\tilde{x}_{k+1} = (1 - \rho)y_k + \rho(z_k - \lambda(V(z_k) - V(y_k)))$. By the three-point identity,

$$\begin{aligned} \|\tilde{x}_{k+1} - x^*\|^2 ={}& (1 - \rho)\|y_k - x^*\|^2 + \rho\|z_k - \lambda(V(z_k) - V(y_k)) - x^*\|^2 \\ &- \rho(1 - \rho)\|z_k - y_k - \lambda(V(z_k) - V(y_k))\|^2. \end{aligned} \tag{4.7}$$

By Lemma 4.1,

$$\|z_k - \lambda(V(z_k) - V(y_k)) - x^*\|^2 \leq \|y_k - x^*\|^2 - (1 - \lambda^2 L^2)\|z_k - y_k\|^2.$$

Expanding the last term in (4.7) gives

$$-\rho(1 - \rho)\|z_k - y_k\|^2 + 2\rho(1 - \rho)\lambda\langle z_k - y_k, V(z_k) - V(y_k)\rangle - \rho(1 - \rho)\lambda^2\|V(z_k) - V(y_k)\|^2.$$

Collecting the $\|z_k - y_k\|^2$ contributions and using $2\rho(1 - \rho)\lambda\langle z_k - y_k, V(z_k) - V(y_k)\rangle \leq 2\rho|1 - \rho|\lambda L\|z_k - y_k\|^2$, we obtain

$$\|\tilde{x}_{k+1} - x^*\|^2 \leq \|y_k - x^*\|^2 - \rho(2 - \rho)\|z_k - y_k\|^2 + \rho\lambda^2 L^2\|z_k - y_k\|^2 + 2\rho|1 - \rho|\lambda L\|z_k - y_k\|^2. \tag{4.8}$$

Now add the stochastic error. Since $x_{k+1} = \tilde{x}_{k+1} + \rho\lambda(e_k^y - e_k^z)$,

$$\|x_{k+1} - x^*\|^2 = \|\tilde{x}_{k+1} - x^*\|^2 + 2\rho\lambda\langle\tilde{x}_{k+1} - x^*, e_k^y - e_k^z\rangle + \rho^2\lambda^2\|e_k^y - e_k^z\|^2.$$

Using $\tilde{x}_{k+1} - x^* = (1 - \rho)(y_k - x^*) + \rho(z_k - x^*) - \rho\lambda(V(z_k) - V(y_k))$ and expanding the inner product, the cross terms decompose into

$$2\lambda\rho(1 - \rho)\langle y_k - x^*, e_k^y - e_k^z\rangle + 2\lambda\rho^2\langle z_k - x^*, e_k^y - e_k^z\rangle - 2\lambda^2\rho^2\langle V(z_k) - V(y_k), e_k^y - e_k^z\rangle.$$

Writing $z_k - x^* = (z_k - y_k) + (y_k - x^*)$ and combining with the $\rho^2\lambda^2\|e_k^y - e_k^z\|^2$ term yields exactly (4.6). □

Inertial bounds. The extrapolation steps introduce memory terms that couple the current iterate to its history. The following pair of elementary bounds controls this coupling.

Lemma 4.3 (Double inertial bound)**.** *Let p_k and y_k be defined by (3.1). Then for any $x^* \in \mathcal{H}$,*

$$\begin{aligned}\|y_k - x^*\|^2 \leq &(1+\beta_k)(1+\gamma_k)\|x_k - x^*\|^2 + \beta_k(1+\beta_k)(1+\gamma_k)\|x_k - x_{k-1}\|^2 \\ &+ \gamma_k(1+\gamma_k)\|p_k - p_{k-1}\|^2. \end{aligned} \tag{4.9}$$

Proof. By definition,

$$\|p_k - x^*\|^2 = \|x_k - x^*\|^2 + \beta_k^2\|x_k - x_{k-1}\|^2 + 2\beta_k\langle x_k - x^*, x_k - x_{k-1}\rangle.$$

Using $2\langle x_k - x^*, x_k - x_{k-1}\rangle \leq \|x_k - x^*\|^2 + \|x_k - x_{k-1}\|^2$ gives

$$\|p_k - x^*\|^2 \leq (1+\beta_k)\|x_k - x^*\|^2 + \beta_k(1+\beta_k)\|x_k - x_{k-1}\|^2.$$

Similarly,

$$\|y_k - x^*\|^2 \leq (1+\gamma_k)\|p_k - x^*\|^2 + \gamma_k(1+\gamma_k)\|p_k - p_{k-1}\|^2.$$

Substituting the bound for $\|p_k - x^*\|^2$ yields (4.9). □

Lemma 4.4 (Bound on $\|p_k - p_{k-1}\|^2$)**.**

$$\|p_k - p_{k-1}\|^2 \leq 2(1+\beta_k)^2\|x_k - x_{k-1}\|^2 + 2\beta_{k-1}^2\|x_{k-1} - x_{k-2}\|^2. \tag{4.10}$$

Proof. From $p_k - p_{k-1} = (1+\beta_k)(x_k - x_{k-1}) - \beta_{k-1}(x_{k-1} - x_{k-2})$ and $(a+b)^2 \leq 2a^2 + 2b^2$. □

Conditional descent. Taking expectation in the fundamental inequality and exploiting the independence of the two mini-batches, the cross terms involving e_k^z vanish in conditional mean. What remains is a clean descent inequality with an explicit variance residual.

Lemma 4.5 (Expected descent)**.** *Under Assumption 2.2 and the parameter restrictions (4.1)–(4.2), for every $x^* \in S$,*

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_{k-1}] \leq \|y_k - x^*\|^2 - \frac{\rho(2-\rho)}{4}\mathbb{E}[\|z_k - y_k\|^2 \mid \mathcal{F}_{k-1}] + C_1\delta_k, \tag{4.11}$$

where $C_1 = 2\lambda^2\rho^2(3+\lambda L)$ is an explicit finite constant.

Proof. We take conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}_{k-1}]$ in (4.6). Because y_k and x^* are $\mathcal{F}_{k-1}$-measurable,

$$\mathbb{E}[\langle y_k - x^*, e_k^z - e_k^y\rangle \mid \mathcal{F}_{k-1}] = \langle y_k - x^*, \mathbb{E}[e_k^z - e_k^y \mid \mathcal{F}_{k-1}]\rangle = 0, \tag{4.12}$$

since both errors are conditionally zero-mean by Remark 2.3.

The remaining stochastic terms are

$$-2\lambda\rho^2\langle z_k - y_k, e_k^z - e_k^y\rangle + 2\lambda^2\rho^2\langle V(z_k) - V(y_k), e_k^z - e_k^y\rangle + \lambda^2\rho^2\|e_k^y - e_k^z\|^2.$$

Because the second batch is independent of the first, the terms involving e_k^z vanish in conditional expectation:

$$\mathbb{E}[\langle z_k - y_k, e_k^z\rangle \mid \mathcal{F}_{k-1}] = \mathbb{E}[\mathbb{E}[\langle z_k - y_k, e_k^z\rangle \mid z_k, \mathcal{F}_{k-1}] \mid \mathcal{F}_{k-1}] = \mathbb{E}[\langle z_k - y_k, \mathbb{E}[e_k^z \mid z_k, \mathcal{F}_{k-1}]\rangle \mid \mathcal{F}_{k-1}] = 0,$$

and similarly $\mathbb{E}[\langle V(z_k) - V(y_k), e_k^z\rangle \mid \mathcal{F}_{k-1}] = 0$.

What remains are the terms involving e_k^y:

$$2\lambda\rho^2\langle z_k - y_k, e_k^y\rangle - 2\lambda^2\rho^2\langle V(z_k) - V(y_k), e_k^y\rangle + \lambda^2\rho^2\|e_k^y - e_k^z\|^2.$$

To handle the first two terms, let

$$\bar{z}_k = J_{\lambda A}(y_k - \lambda V(y_k))$$

be the deterministic resolvent point. By firm nonexpansivity of the resolvent, $\|z_k - \bar{z}_k\| \le \lambda\|e_k^y\|$. Decomposing $z_k - y_k = (\bar{z}_k - y_k) + (z_k - \bar{z}_k)$ and $V(z_k) - V(y_k) = (V(\bar{z}_k) - V(y_k)) + (V(z_k) - V(\bar{z}_k))$, we obtain

$$2\lambda\rho^2\langle z_k - y_k, e_k^y\rangle = 2\lambda\rho^2\langle \bar{z}_k - y_k, e_k^y\rangle + 2\lambda\rho^2\langle z_k - \bar{z}_k, e_k^y\rangle,$$
$$-2\lambda^2\rho^2\langle V(z_k) - V(y_k), e_k^y\rangle = -2\lambda^2\rho^2\langle V(\bar{z}_k) - V(y_k), e_k^y\rangle - 2\lambda^2\rho^2\langle V(z_k) - V(\bar{z}_k), e_k^y\rangle.$$

The inner products with $\bar{z}_k$ and $V(\bar{z}_k)$ are conditionally zero-mean because these deterministic quantities depend only on $\mathcal{F}_{k-1}$. For the residual terms we use Cauchy–Schwarz and the Lipschitz property:

$$2\lambda\rho^2\langle z_k - \bar{z}_k, e_k^y\rangle \le 2\lambda\rho^2\|z_k - \bar{z}_k\|\|e_k^y\| \le 2\lambda^2\rho^2\|e_k^y\|^2,$$
$$-2\lambda^2\rho^2\langle V(z_k) - V(\bar{z}_k), e_k^y\rangle \le 2\lambda^2\rho^2 L\|z_k - \bar{z}_k\|\|e_k^y\| \le 2\lambda^3\rho^2 L\|e_k^y\|^2.$$

Taking conditional expectation and using the variance bounds, the non-vanishing parts contribute at most $2\lambda^2\rho^2(1+\lambda L)\delta_k$.

The pure variance term satisfies $\lambda^2\rho^2\mathbb{E}[\|e_k^y - e_k^z\|^2 \mid \mathcal{F}_{k-1}] \le 4\lambda^2\rho^2\delta_k$.

Collecting all deterministic contributions to the $\|z_k - y_k\|^2$ coefficient, we have

$$-\rho(2-\rho) + \rho\lambda^2 L^2 + 2\rho|1-\rho|\lambda L.$$

By the step-size condition (4.2),

$$\lambda^2 L^2 + 2|1-\rho|\lambda L \le \frac{3(2-\rho)}{4},$$

so the net coefficient is at most $-\rho(2-\rho)/4$ for all $\rho \in (0,2)$.

The total variance contribution is bounded by

$$2\lambda^2\rho^2(1+\lambda L)\delta_k + 4\lambda^2\rho^2\delta_k = 2\lambda^2\rho^2(3+\lambda L)\delta_k =: C_1\delta_k. \qquad \square$$

Lyapunov analysis. The descent inequality (4.11) is not yet recursive in $\|x_k - x^*\|^2$ because the right-hand side involves y_k, which carries inertial memory. To absorb this memory we introduce the Lyapunov function

$$\Phi_k = \|x_k - x^*\|^2 + A_k\|x_k - x_{k-1}\|^2 + B_k\|p_k - p_{k-1}\|^2, \tag{4.13}$$

where $A_0 = B_0 = 0$ and, for $k \ge 1$,

$$A_k := 12\sum_{j=k-1}^{\infty}\left(\beta_j + \beta_j^2 + \gamma_j^2\right), \tag{4.14}$$

$$B_k := 12\sum_{j=k}^{\infty}\left(\gamma_j + \gamma_j^2\right). \tag{4.15}$$

Because $\sum\beta_k < \infty$ and $\sum\gamma_k < \infty$, the sequences $\{A_k\}$ and $\{B_k\}$ are well defined, nonnegative, decreasing, and tend to 0; in particular $\sup_k A_k < \infty$ and $\sup_k B_k < \infty$.

We shall assume that the inertial parameters are monotone, which is standard and incurs no loss of generality: any summable positive sequence can be dominated by a non-increasing summable sequence (e.g. $\bar{\beta}_k = \sup_{j\ge k}\beta_j$), and replacing β_k by $\bar{\beta}_k$ only strengthens the bounds.

Assumption 4.6 (Monotone inertial parameters). *The inertial sequences $\{\beta_k\}$ and $\{\gamma_k\}$ are non-increasing.*

The weights A_k, B_k are designed so that the telescoping differences $A_k - A_{k+1}$ and $B_k - B_{k+1}$ dominate the inertial coefficients appearing in the expected descent. This yields the following Lyapunov descent.

Lemma 4.7 (Lyapunov descent). *Assume Assumption 4.6, $\sum\beta_k < \infty$, $\sum\gamma_k < \infty$, and $\sum\delta_k < \infty$. Then there exist constants $\alpha = \rho(2-\rho)/8 > 0$ and $C_\Phi < \infty$ such that for all $k \ge 0$,*

$$\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_{k-1}] \le (1+\theta_k)\Phi_k - \alpha\mathbb{E}[\|z_k - y_k\|^2 \mid \mathcal{F}_{k-1}] + C_\Phi\delta_k + R\cdot\mathbf{1}_{k<K_0}, \tag{4.16}$$

where $\theta_k = 6\beta_k + 4\gamma_k + 2\beta_k^2 + 2\gamma_k^2$ satisfies $\sum\theta_k < \infty$, K_0 is a finite deterministic index, and $R \ge 0$ is a finite constant.

Proof. All inequalities hold pointwise; we take conditional expectation at the end.

From Lemma 4.5 and Lemma 4.3 we first obtain

$$\mathbb{E}[\|x_{k+1}-x^*\|^2 \mid \mathcal{F}_{k-1}] \le (1+\theta_k^{(1)})\|x_k-x^*\|^2 + \beta_k(1+\beta_k)(1+\gamma_k)\|x_k-x_{k-1}\|^2 + \gamma_k(1+\gamma_k)\|p_k-p_{k-1}\|^2 - \frac{\rho(2-\rho)}{4}\mathbb{E}[\|z_k-y_k\|^2 \mid \mathcal{F}_{k-1}] + C_1\delta_k, \tag{4.17}$$

with $\theta_k^{(1)} := \beta_k + \gamma_k + \beta_k\gamma_k$.

Next, from Algorithm 1 and the elementary bound $(a+b)^2 \le 2a^2+2b^2$ applied twice,

$$\mathbb{E}[\|x_{k+1}-x_k\|^2 \mid \mathcal{F}_{k-1}] \le (4\rho^2+8\rho^2\lambda^2L^2)\mathbb{E}[\|z_k-y_k\|^2 \mid \mathcal{F}_{k-1}] + 32\rho^2\lambda^2\delta_k + 4\beta_k^2\|x_k-x_{k-1}\|^2 + 4\gamma_k^2\|p_k-p_{k-1}\|^2, \tag{4.18}$$

where we used $y_k - x_k = \beta_k(x_k-x_{k-1}) + \gamma_k(p_k-p_{k-1})$ and the bound $\mathbb{E}[\|\mathcal{W}_k-\mathcal{V}_k\|^2 \mid \mathcal{F}_{k-1}] \le 2L^2\mathbb{E}[\|z_k-y_k\|^2 \mid \mathcal{F}_{k-1}] + 8\delta_k$, which follows from $\mathcal{W}_k - \mathcal{V}_k = V(z_k) - V(y_k) + e_k^z - e_k^y$ and the Lipschitz property of V.

Similarly, since $p_{k+1}-p_k = (1+\beta_{k+1})(x_{k+1}-x_k) - \beta_k(x_k-x_{k-1})$,

$$\mathbb{E}[\|p_{k+1}-p_k\|^2 \mid \mathcal{F}_{k-1}] \le (1+\beta_{k+1})^2(8\rho^2+16\rho^2\lambda^2L^2)\mathbb{E}[\|z_k-y_k\|^2 \mid \mathcal{F}_{k-1}] + 64(1+\beta_{k+1})^2\rho^2\lambda^2\delta_k + \big(8(1+\beta_{k+1})^2\beta_k^2 + 2\beta_k^2\big)\|x_k-x_{k-1}\|^2 + 8(1+\beta_{k+1})^2\gamma_k^2\|p_k-p_{k-1}\|^2. \tag{4.19}$$

Now add $A_{k+1}\times$(4.18) and $B_{k+1}\times$(4.19) to (4.17). The coefficient of $\mathbb{E}[\|z_k-y_k\|^2 \mid \mathcal{F}_{k-1}]$ becomes

$$-\frac{\rho(2-\rho)}{4} + 4\rho^2(1+2\lambda^2L^2)A_{k+1} + 8\rho^2(1+2\lambda^2L^2)(1+\beta_{k+1})^2B_{k+1}.$$

Because $\{A_k\}$ and $\{B_k\}$ decrease to 0, there exists a deterministic index K_0 such that for all $k \ge K_0$,

$$A_{k+1} \le \frac{\rho(2-\rho)}{64\rho^2(1+2\lambda^2L^2)}, \qquad B_{k+1} \le \frac{\rho(2-\rho)}{128\rho^2(1+2\lambda^2L^2)(1+\beta_{k+1})^2},$$

whence the above coefficient is at most $-\rho(2-\rho)/8$. For $k < K_0$, the coefficient is bounded below by $-M$, where

$$M := \frac{\rho(2-\rho)}{4} + 4\rho^2(1+2\lambda^2L^2)\sup_{j\ge0}A_j + 8\rho^2(1+2\lambda^2L^2)\sup_{j\ge0}\big[B_j(1+\beta_j)^2\big] < \infty.$$

The total contribution of these early iterations to the Robbins–Siegmund recursion is therefore at most $MK_0 < \infty$, which is absorbed into the constant R. Hence there exists $\alpha = \rho(2-\rho)/8$ such that the net coefficient is $\le -\alpha + M\cdot\mathbf{1}_{k<K_0}$ for all k.

The coefficient of $\|x_k-x_{k-1}\|^2$ in the total sum is

$$C_k^{(x)} := \beta_k(1+\beta_k)(1+\gamma_k) + 4A_{k+1}\beta_k^2 + B_{k+1}\big(8(1+\beta_{k+1})^2\beta_k^2 + 2\beta_k^2\big).$$

Because $A_{k+1} \le A_k$ and $B_{k+1} \le B_k$, and using $\beta_k, \gamma_k \in [0,1)$, we have $C_k^{(x)} \le 6\beta_k + 4\beta_k^2$ for all $k \ge 0$. From (4.14) and the monotonicity assumption,

$$A_k - A_{k+1} = 12\big(\beta_{k-1} + \beta_{k-1}^2 + \gamma_{k-1}^2\big) \ge 12\beta_{k-1} \ge 12\beta_k.$$

Because $\beta_{k-1}^2 = o(\beta_{k-1})$ and $\gamma_{k-1}^2 = o(\gamma_{k-1})$ as $k\to\infty$, there exists K_1 such that for all $k \ge K_1$,

$$A_k - A_{k+1} \ge 6\beta_k + 4\beta_k^2 \ge C_k^{(x)}.$$

Therefore $(1+\theta_k^{(x)})A_k - A_{k+1} \geq C_k^{(x)}$ with $\theta_k^{(x)} = O(\beta_k)$. For $k < K_1$, the deficit $(C_k^{(x)} - (A_k - A_{k+1}))_+$ is bounded by $\sup_j C_j^{(x)} < \infty$; summing over $k < K_1$ contributes a finite amount that is absorbed into R.

Similarly, the coefficient of $\|p_k - p_{k-1}\|^2$ is

$$C_k^{(p)} := \gamma_k(1+\gamma_k) + 4A_{k+1}\gamma_k^2 + 8B_{k+1}(1+\beta_{k+1})^2\gamma_k^2 \leq 4\gamma_k + 4\gamma_k^2.$$

From (4.15),

$$B_k - B_{k+1} = 12(\gamma_k + \gamma_k^2) \geq 4\gamma_k + 4\gamma_k^2 \geq C_k^{(p)}.$$

This holds for all k and all $\rho \in (0,2)$ by construction of the weights. Thus $(1+\theta_k^{(p)})B_k - B_{k+1} \geq C_k^{(p)}$.

Setting $\theta_k = \max\{\theta_k^{(1)}, \theta_k^{(x)}, \theta_k^{(p)}\} = O(\beta_k + \gamma_k)$ and collecting all residual constants into C_Φ and R, we obtain (4.16). □

Almost sure weak convergence. The Lyapunov descent fits the Robbins–Siegmund template, yielding the main convergence result.

Theorem 4.8 (Almost sure weak convergence)**.** *Assume Assumption 4.6, $\sum \beta_k < \infty$, $\sum \gamma_k < \infty$, $\rho \in (0,2)$, λ satisfying (4.2), and $\sum \delta_k < \infty$. Then $\{x_k\}$ converges weakly almost surely to an S-valued random variable.*

Proof. **Step 1 (Robbins–Siegmund).** Lemma 4.7 gives

$$\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_{k-1}] \leq (1+\theta_k)\Phi_k - \alpha\mathbb{E}[\|z_k - y_k\|^2 \mid \mathcal{F}_{k-1}] + C_\Phi\delta_k + R \cdot \mathbf{1}_{k<K_0}.$$

Set $\mathcal{G}_k := \mathcal{F}_{k-1}$ and $U_k := \Phi_k$. Then U_k is $\mathcal{G}_k$-measurable and the recursion fits Lemma 2.4 with $w_k = C_\Phi\delta_k + R \cdot \mathbf{1}_{k<K_0}$. Since $\sum \theta_k < \infty$ and $\sum w_k < \infty$, Lemma 2.4 implies that Φ_k converges a.s. to a finite random variable and $\sum_{k=0}^{\infty} \mathbb{E}[\|z_k - y_k\|^2 \mid \mathcal{F}_{k-1}] < \infty$ a.s.

Step 2 (Vanishing differences). By the conditional Borel–Cantelli lemma, $\|z_k - y_k\| \to 0$ a.s. From Algorithm 1,

$$\|x_{k+1} - y_k\| \leq \rho\|z_k - y_k\| + \rho\lambda\|\mathcal{W}_k - \mathcal{V}_k\|.$$

Since $\mathbb{E}[\|\mathcal{W}_k - \mathcal{V}_k\|^2] \leq 2L^2\mathbb{E}[\|z_k - y_k\|^2] + 8\delta_k$ and both terms on the right-hand side are summable (hence tend to 0), we have $\sum \mathbb{E}[\|x_{k+1} - y_k\|^2] < \infty$, whence $\|x_{k+1} - y_k\| \to 0$ a.s. Moreover,

$$y_k - x_k = \beta_k(x_k - x_{k-1}) + \gamma_k(p_k - p_{k-1}) \to 0 \quad \text{a.s.},$$

because $\beta_k, \gamma_k \to 0$ and $\|x_k - x_{k-1}\|$, $\|p_k - p_{k-1}\|$ are bounded (from convergence of Φ_k). Hence $\|x_{k+1} - x_k\| \to 0$ a.s.

Step 3 (Cluster points are solutions). Let $x_{k_n} \rightharpoonup p$. Then $y_{k_n} \rightharpoonup p$ and $z_{k_n} \rightharpoonup p$ because the differences vanish. Since $z_k = J_{\lambda A}(y_k - \lambda\mathcal{V}_k)$ and $\mathcal{V}_k - V(y_k) = e_k^y \to 0$ (in L^2, hence a.s. along a subsequence), the strong-to-weak closedness of $\operatorname{gra} A$ and the continuity of V imply $0 \in (A+V)(p)$, i.e. $p \in S$.

Step 4 (Opial). Convergence of Φ_k implies $\lim_{k\to\infty} \|x_k - x^*\|^2$ exists for each $x^* \in S$. By Lemma 2.1, x_k converges weakly a.s. to a point in S. □

5. Rate of Convergence

The Robbins–Siegmund argument in Theorem 4.8 guarantees that the series $\sum_{j=0}^{\infty} \mathbb{E}[\|z_j - y_j\|^2]$ converges. This summability alone does not yield a quantitative rate, but a standard Cesàro argument converts the ℓ^1-bound into an explicit $O(1/k)$ estimate for the best-observed residual. The same device applies to the iterate differences once the mini-batch variance decays sufficiently fast.

Theorem 5.1 (Rate for discrete velocity). *Under the conditions of Theorem 4.8,*

$$\mathbb{E}\Big[\inf_{0\le j\le k}\|z_j-y_j\|^2\Big] \;\le\; \frac{1}{k+1}\Big(\Phi_0+\sum_{j=0}^{k}C_\Phi\delta_j+RK_0\Big) \;=\; O(1/k). \tag{5.1}$$

If the batch sizes are chosen so that $\delta_j = O(1/j^2)$, then the iterate differences satisfy

$$\mathbb{E}\Big[\inf_{0\le j\le k}\|x_{j+1}-x_j\|^2\Big] \;=\; O(1/k). \tag{5.2}$$

Proof. From the Lyapunov descent (4.16) we have, for every $j\ge 0$,

$$\alpha\,\mathbb{E}[\|z_j-y_j\|^2\mid\mathcal{F}_{j-1}]\le(1+\theta_j)\Phi_j-\mathbb{E}[\Phi_{j+1}\mid\mathcal{F}_{j-1}]+C_\Phi\delta_j+R\cdot\mathbf{1}_{j<K_0}.$$

Taking full expectation and telescoping from $j=0$ to k gives

$$\alpha\sum_{j=0}^{k}\mathbb{E}[\|z_j-y_j\|^2]\le\Phi_0+\sum_{j=0}^{k}\theta_j\mathbb{E}[\Phi_j]+\sum_{j=0}^{k}C_\Phi\delta_j+RK_0.$$

Because Φ_j converges a.s. to a finite limit (Theorem 4.8, Step 1) and $\sum\theta_j<\infty$, the middle term is bounded independently of k. Hence there exists a constant $M:=\Phi_0+(\sup_j\mathbb{E}[\Phi_j])\sum_{j=0}^{\infty}\theta_j+\sum_{j=0}^{\infty}C_\Phi\delta_j+RK_0<\infty$ such that

$$\sum_{j=0}^{k}\mathbb{E}[\|z_j-y_j\|^2]\le\frac{M}{\alpha}\qquad\text{for all } k.$$

Since the infimum over $j\le k$ is bounded by the average,

$$\mathbb{E}\Big[\inf_{j\le k}\|z_j-y_j\|^2\Big]\le\frac{1}{k+1}\sum_{j=0}^{k}\mathbb{E}[\|z_j-y_j\|^2]\le\frac{M}{\alpha(k+1)},$$

which is exactly (5.1).

For the iterate differences, recall from Algorithm 1 that

$$x_{j+1}-x_j=(x_{j+1}-y_j)+(y_j-x_j),\qquad x_{j+1}-y_j=\rho(z_j-y_j)-\rho\lambda(\mathcal{W}_j-\mathcal{V}_j),$$

and $y_j-x_j=\beta_j(x_j-x_{j-1})+\gamma_j(p_j-p_{j-1})$. Using $(a+b)^2\le 2a^2+2b^2$ twice,

$$\begin{aligned}\|x_{j+1}-x_j\|^2&\le 4\rho^2\|z_j-y_j\|^2+4\rho^2\lambda^2\|\mathcal{W}_j-\mathcal{V}_j\|^2+2\|y_j-x_j\|^2\\&\le 4\rho^2\|z_j-y_j\|^2+4\rho^2\lambda^2\|\mathcal{W}_j-\mathcal{V}_j\|^2+4\beta_j^2\|x_j-x_{j-1}\|^2+4\gamma_j^2\|p_j-p_{j-1}\|^2.\end{aligned}$$

Taking expectation and using $\mathbb{E}[\|\mathcal{W}_j-\mathcal{V}_j\|^2]\le 2L^2\mathbb{E}[\|z_j-y_j\|^2]+8\delta_j$ together with the boundedness of $\mathbb{E}[\|x_j-x_{j-1}\|^2]$ and $\mathbb{E}[\|p_j-p_{j-1}\|^2]$ (which follows from the convergence of Φ_j), we obtain

$$\mathbb{E}[\|x_{j+1}-x_j\|^2]\le C_2\,\mathbb{E}[\|z_j-y_j\|^2]+C_3\delta_j+C_4(\beta_j^2+\gamma_j^2)$$

for suitable explicit constants C_2,C_3,C_4. Summing from $j=0$ to k,

$$\sum_{j=0}^{k}\mathbb{E}[\|x_{j+1}-x_j\|^2]\le C_2\sum_{j=0}^{k}\mathbb{E}[\|z_j-y_j\|^2]+C_3\sum_{j=0}^{k}\delta_j+C_4\sum_{j=0}^{k}(\beta_j^2+\gamma_j^2).$$

Under the hypothesis $\delta_j=O(1/j^2)$, the series $\sum\delta_j$ converges; likewise $\sum(\beta_j^2+\gamma_j^2)<\infty$ because β_j,γ_j are summable. Hence the right-hand side is bounded independently of k, and the Cesàro argument yields (5.2). □

Remark 5.2 (Interpretation of the $O(1/k)$ rate). The rate (5.1) is stated for the *best* residual observed up to iteration k, which is the standard object of study in stochastic monotone inclusion literature because the residual $\|z_k - y_k\|$ need not decrease monotonically. The $O(1/k)$ scaling for the squared residual is the standard rate for first-order methods on merely monotone problems under summable variance.

Remark 5.3 (Role of the batch schedule). The condition $\delta_j = O(1/j^2)$, i.e. $m_j = \Omega(j^2)$, is needed only for the iterate-difference rate (5.2). The fundamental residual rate (5.1) holds as soon as $\sum \delta_j < \infty$, which is already assumed in Theorem 4.8. In practice one often uses $m_j \propto j^a$ with $a > 1$ or geometric growth; both satisfy the summability requirement.

Remark 5.4 (Comparison with deterministic double inertia). In the deterministic setting ($\delta_j \equiv 0$, exact oracle), the same Lyapunov argument yields $\sum_{j=0}^{\infty} \|z_j - y_j\|^2 < \infty$, whence $\|z_k - y_k\|^2 = o(1/k)$. The stochastic rate $O(1/k)$ is therefore optimal up to constants: the mini-batch noise prevents the faster $o(1/k)$ decay enjoyed by exact methods. This mirrors the well-known gap between deterministic and stochastic optimization rates.

6. Strongly Monotone Case

Assumption 6.1 (Strong monotonicity). *$A + V$ is μ-strongly monotone: for all $x, y \in \operatorname{dom} A$ and all $u \in (A+V)(x)$, $v \in (A+V)(y)$,*

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

Under Assumption 6.1, $S = \{x^*\}$ is a singleton.

Lemma 6.2 (Strongly monotone descent). *Let Assumption 6.1 hold and let $\rho \in (0, 2)$. Choose λ satisfying (4.2) and the additional bound*

$$\lambda^2 L^2 + 2|1 - \rho|\lambda L + 2\lambda\mu \leq \frac{2 - \rho}{2}. \tag{6.1}$$

Then there exist constants $\eta = \rho\lambda\mu/2 \in (0, 1)$ and $\tilde{C}_1 > 0$ such that

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_{k-1}] \leq (1 - \rho\lambda\mu)\|y_k - x^*\|^2 - \frac{\rho(2-\rho)}{2}\mathbb{E}[\|z_k - y_k\|^2 \mid \mathcal{F}_{k-1}] + \tilde{C}_1\delta_k. \tag{6.2}$$

Proof. We return to the expansion in the proof of Lemma 4.2. Because $A + V$ is strongly monotone and $(y_k - z_k)/\lambda - \mathcal{V}_k + V(z_k) \in (A+V)(z_k)$, we have

$$\left\langle z_k - x^*, \frac{y_k - z_k}{\lambda} - \mathcal{V}_k + V(z_k)\right\rangle \geq \mu\|z_k - x^*\|^2.$$

Since $\mathcal{V}_k = V(y_k) + e_k^y$, this reads

$$\langle z_k - x^*, y_k - z_k\rangle \geq \lambda\mu\|z_k - x^*\|^2 + \lambda\langle z_k - x^*, e_k^y\rangle. \tag{6.3}$$

Now, in the identity

$$2\rho\langle y_k - x^*, z_k - y_k\rangle = -2\rho\|z_k - y_k\|^2 + 2\rho\langle z_k - x^*, z_k - y_k\rangle,$$

we use (6.3) to obtain

$$2\rho\langle y_k - x^*, z_k - y_k\rangle \leq -2\rho\|z_k - y_k\|^2 - 2\rho\lambda\mu\|z_k - x^*\|^2 - 2\rho\lambda\langle z_k - x^*, e_k^y\rangle.$$

Because $\|z_k - x^*\|^2 \geq \frac{1}{2}\|y_k - x^*\|^2 - \|z_k - y_k\|^2$,

$$-2\rho\lambda\mu\|z_k - x^*\|^2 \leq -\rho\lambda\mu\|y_k - x^*\|^2 + 2\rho\lambda\mu\|z_k - y_k\|^2. \tag{6.4}$$

Inserting (6.4) into the fundamental expansion and tracking the deterministic coefficients, the net coefficient of $\|z_k - y_k\|^2$ becomes

$$-\rho(2 - \rho) + \rho\lambda^2 L^2 + 2\rho|1 - \rho|\lambda L + 2\rho\lambda\mu.$$

By the joint step-size condition (6.1), this is at most $-\rho(2-\rho)/2$. The stochastic terms are handled exactly as in Lemma 4.5: the $\langle y_k - x^*, e_k^z - e_k^y \rangle$ term vanishes in conditional expectation, while the remaining error terms are bounded by the same argument, contributing at most $\tilde{C}_1\delta_k$ with $\tilde{C}_1 = 2\lambda^2\rho^2(3+\lambda L)$. The contraction $-\rho\lambda\mu\|y_k - x^*\|^2$ yields the factor $(1-\rho\lambda\mu)$ in (6.2). □

Theorem 6.3 (Linear convergence). *Under Assumption 6.1, let β_k, γ_k be summable, $\rho \in (0,2)$, and $m_{k+1} = \lceil \tau m_k \rceil$ for some $\tau \geq (1-\eta/2)^{-1}$ where $\eta = \rho\lambda\mu/2$. Then*

$$\mathbb{E}[\|x_k - x^*\|^2] \leq C_0(1-\eta/2)^k \quad \text{for all large } k, \tag{6.5}$$

and the oracle complexity to reach $\mathbb{E}[\|x_k - x^*\|^2] \leq \varepsilon$ is $\mathcal{O}(1/\varepsilon)$ when $\tau = (1-\eta/2)^{-1}$. For $\tau > (1-\eta/2)^{-1}$, the complexity is $\mathcal{O}\big((1/\varepsilon)^{\log\tau/\log(1/(1-\eta/2))}\big)$.

Proof. From Lemma 6.2, the descent inequality for $\|x_{k+1} - x^*\|^2$ contains the contraction factor $(1-\rho\lambda\mu)$ on $\|y_k - x^*\|^2$. To pass this contraction to the Lyapunov function, we use the upper bound from Lemma 4.3 to write

$$\begin{aligned}\|y_k - x^*\|^2 \leq{}& (1+\beta_k)(1+\gamma_k)\|x_k - x^*\|^2 \\ &+ \beta_k(1+\beta_k)(1+\gamma_k)\|x_k - x_{k-1}\|^2 + \gamma_k(1+\gamma_k)\|p_k - p_{k-1}\|^2.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_{k-1}] \leq{}& (1-\rho\lambda\mu)(1+\beta_k)(1+\gamma_k)\|x_k - x^*\|^2 \\ &+ (1-\rho\lambda\mu)\beta_k(1+\beta_k)(1+\gamma_k)\|x_k - x_{k-1}\|^2 \\ &+ (1-\rho\lambda\mu)\gamma_k(1+\gamma_k)\|p_k - p_{k-1}\|^2 \\ &- \frac{\rho(2-\rho)}{2}\mathbb{E}[\|z_k - y_k\|^2 \mid \mathcal{F}_{k-1}] + \tilde{C}_1\delta_k.\end{aligned}$$

The inertial terms are dominated by the Lyapunov weights A_k, B_k exactly as in Lemma 4.7, because their coefficients are multiplied by the factor $(1-\rho\lambda\mu) < 1$ and are therefore smaller than in the merely monotone case. The coefficient of $\|x_k - x^*\|^2$ is $(1-\rho\lambda\mu)(1+\beta_k)(1+\gamma_k)$. Since $\beta_k, \gamma_k \to 0$, there exists a finite index K_2 such that for all $k \geq K_2$,

$$(1+\beta_k)(1+\gamma_k) \leq \frac{1-\rho\lambda\mu/2}{1-\rho\lambda\mu}, \qquad \text{whence} \qquad (1-\rho\lambda\mu)(1+\beta_k)(1+\gamma_k) \leq 1 - \frac{\rho\lambda\mu}{2} = 1-\eta.$$

Hence, following the same Lyapunov construction as in Lemma 4.7 but now with the additional contraction on $\|x_k - x^*\|^2$, we obtain for all $k \geq \max\{K_0, K_1, K_2\}$:

$$\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_{k-1}] \leq (1-\eta)\Phi_k + \hat{C}_\Phi\delta_k,$$

where $\hat{C}_\Phi < \infty$ is an explicit constant that collects the variance residuals from the inertial differences (cf. Lemma 4.7). Iterating this inequality and using $\tau^{-1} \leq 1-\eta/2$ to control the geometrically decaying variance yields $\mathbb{E}[\Phi_k] = O((1-\eta/2)^k)$, whence $\mathbb{E}[\|x_k - x^*\|^2] = O((1-\eta/2)^k)$.

For the oracle complexity with $\tau = (1-\eta/2)^{-1}$, the total oracle calls after k iterations are $\sum_{j=0}^{k} m_j = O(\tau^k) = O((1-\eta/2)^{-k})$. Since the error decays as $(1-\eta/2)^k$, reaching accuracy ε requires $k = \Theta(\log(1/\varepsilon))$ iterations and $O((1-\eta/2)^{-k}) = O(1/\varepsilon)$ oracle calls. For $\tau > (1-\eta/2)^{-1}$, the same calculation gives the stated polynomial complexity. □

7. Handling Biased Oracles

In many practical pipelines the stochastic oracle is not perfectly unbiased. Compression, quantization, surrogate modeling, finite-difference approximations, or systematic measurement errors all introduce a bounded bias $b(x)$ into the gradient estimate. Driggs *et al.* [8] have highlighted that ignoring such bias can prevent convergence to the exact solution and may produce spurious limit cycles. This

section extends our analysis to oracles with bounded bias, deriving explicit non-asymptotic bounds that track the exact dependence on the bias magnitude B, the strong-monotonicity constant μ, and the algorithmic parameters (λ, ρ).

Biased oracle model. We replace Assumption 2.2 by the following perturbed version.

Assumption 7.1 (Biased oracle). *The oracle returns $\tilde{V}_\xi(x)$ with*

$$\mathbb{E}[\tilde{V}_\xi(x)] = V(x) + b(x), \qquad \|b(x)\| \le B, \qquad \mathbb{E}[\|\tilde{V}_\xi(x) - V(x)\|^2] \le \sigma^2,$$

for all $x \in \mathcal{H}$, where $B \ge 0$ is a uniform bias bound.

The bias $b(x)$ is deterministic (or at least $\mathcal{F}_{k-1}$-measurable) and need not vanish at the solution; this is precisely what makes it problematic. The variance bound σ^2 is assumed unchanged, which is realistic when bias is introduced by a deterministic perturbation (e.g., rounding or compression) applied to a stochastic estimator with original variance σ^2.

Error decomposition. With the biased oracle, the mini-batch estimators become

$$\tilde{\mathcal{V}}_k = \frac{1}{m_k}\sum_{i=1}^{m_k} \tilde{V}_{\xi_{k,i}}(y_k), \qquad \tilde{\mathcal{W}}_k = \frac{1}{m_k}\sum_{i=1}^{m_k} \tilde{V}_{\xi'_{k,i}}(z_k).$$

We decompose the total error into an unbiased stochastic part and a deterministic bias part:

$$\tilde{e}_k^y := \tilde{\mathcal{V}}_k - V(y_k) = \underbrace{\tilde{\mathcal{V}}_k - \mathbb{E}[\tilde{\mathcal{V}}_k \mid \mathcal{F}_{k-1}]}_{=:e_k^y} + \underbrace{b(y_k)}_{\text{bias at } y_k},$$

and similarly

$$\tilde{e}_k^z := \tilde{\mathcal{W}}_k - V(z_k) = e_k^z + b(z_k).$$

Because the two mini-batches are independent, the unbiased parts e_k^y and e_k^z satisfy exactly the same conditional zero-mean properties and variance bounds as in Section 4. The bias terms, however, do not vanish in expectation. Their difference satisfies

$$\|\tilde{e}_k^y - \tilde{e}_k^z\|^2 \le 2\|e_k^y - e_k^z\|^2 + 2\|b(y_k) - b(z_k)\|^2 \le 2\|e_k^y - e_k^z\|^2 + 8B^2,$$

whence, taking conditional expectation and using (2.2),

$$\mathbb{E}[\|\tilde{e}_k^y - \tilde{e}_k^z\|^2 \mid \mathcal{F}_{k-1}] \le 4\delta_k + 8B^2. \tag{7.1}$$

The additive $8B^2$ term is the fundamental price of bias: even as the batch size grows and $\delta_k \to 0$, the variance of the difference $\tilde{e}_k^y - \tilde{e}_k^z$ remains bounded away from zero by $8B^2$.

Biased descent inequality. Under strong monotonicity, the contraction from Lemma 6.2 is partially offset by the bias. The following lemma quantifies this trade-off with explicit constants.

Lemma 7.2 (Expected descent with bias). *Under Assumptions 6.1 and 7.1, with $\rho \in (0, 2)$ and λ satisfying (4.2) and (6.1), the iterates satisfy*

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_{k-1}] \le \Big(1 - \frac{\mu\lambda\rho}{2}\Big)\|y_k - x^*\|^2 - \frac{\rho(2-\rho)}{8}\mathbb{E}[\|z_k - y_k\|^2 \mid \mathcal{F}_{k-1}] + \tilde{C}_1\delta_k + \tilde{C}_2 B^2, \tag{7.2}$$

where

$$\tilde{C}_1 = 2\lambda^2\rho^2(3 + \lambda L), \qquad \tilde{C}_2 = \frac{16\lambda\rho}{\mu} + \frac{32\lambda^2\rho^3(1+\lambda L)^2}{2-\rho} + 8\lambda^2\rho^2.$$

Proof. We return to the fundamental expansion underlying Lemma 6.2 and track the bias contributions explicitly. The stochastic cross terms now have conditional mean involving $b(y_k) - b(z_k)$ instead of zero. We bound each bias-induced inner product by Young's inequality, absorbing the contraction loss into the existing strong-monotonicity term.

Bias in the $y_k - x^*$ inner product. The term $2\lambda\rho\langle y_k - x^*, \tilde{e}_k^z - \tilde{e}_k^y\rangle$ splits into an unbiased part (handled exactly as in Lemma 4.5) and a bias part

$$2\lambda\rho\langle y_k - x^*, b(y_k) - b(z_k)\rangle.$$

By Young's inequality with $\varepsilon = \mu/4$,

$$2\lambda\rho\langle y_k - x^*, b(y_k) - b(z_k)\rangle \le \frac{\mu\lambda\rho}{4}\|y_k - x^*\|^2 + \frac{4\lambda\rho}{\mu}\|b(y_k) - b(z_k)\|^2.$$

Since $\|b(y_k) - b(z_k)\| \le \|b(y_k)\| + \|b(z_k)\| \le 2B$, we have $\|b(y_k) - b(z_k)\|^2 \le 4B^2$, and the second term is at most $16\lambda\rho B^2/\mu$.

Bias in the $z_k - y_k$ and V inner products. The terms $-2\lambda\rho^2\langle z_k - y_k, \tilde{e}_k^z - \tilde{e}_k^y\rangle$ and $2\lambda^2\rho^2\langle V(z_k) - V(y_k), \tilde{e}_k^z - \tilde{e}_k^y\rangle$ contribute bias parts

$$2\lambda\rho^2\|z_k - y_k\|\|b(y_k) - b(z_k)\| + 2\lambda^2\rho^2\|V(z_k) - V(y_k)\|\|b(z_k) - b(y_k)\|.$$

Using $\|b(y_k) - b(z_k)\| \le 2B$ and $\|V(z_k) - V(y_k)\| \le L\|z_k - y_k\|$, the combined bound is

$$4\lambda\rho^2 B(1 + \lambda L)\|z_k - y_k\|.$$

By Young's inequality with $\varepsilon = \rho(2-\rho)/4$,

$$4\lambda\rho^2 B(1 + \lambda L)\|z_k - y_k\| \le \frac{\rho(2-\rho)}{8}\|z_k - y_k\|^2 + \frac{32\lambda^2\rho^3(1+\lambda L)^2 B^2}{2-\rho}.$$

Pure variance term. From (7.1),

$$\lambda^2\rho^2\mathbb{E}[\|\tilde{e}_k^y - \tilde{e}_k^z\|^2 \mid \mathcal{F}_{k-1}] \le 4\lambda^2\rho^2\delta_k + 8\lambda^2\rho^2 B^2.$$

Unbiased stochastic terms. The terms involving e_k^y and e_k^z (without bias) are bounded exactly as in Lemma 4.5, contributing $\tilde{C}_1\delta_k$.

Collecting all bias contributions, the total additive bias term is

$$\frac{16\lambda\rho B^2}{\mu} + \frac{32\lambda^2\rho^3(1+\lambda L)^2 B^2}{2-\rho} + 8\lambda^2\rho^2 B^2 = \tilde{C}_2 B^2.$$

The contraction coefficient of $\|y_k - x^*\|^2$ is reduced from $1 - \mu\lambda\rho$ to $1 - \mu\lambda\rho/2$ because we absorbed a factor of $\mu\lambda\rho/4$ into the bias bound. The coefficient of $\mathbb{E}[\|z_k - y_k\|^2]$ remains $-\rho(2-\rho)/8$ because the bias-induced Young term exactly matches the amount needed for absorption. This yields (7.2). □

Convergence to a bias neighborhood. Iterating the biased descent inequality in the Lyapunov framework shows that the iterates converge not to the exact solution, but to an explicit neighborhood whose radius scales as $O(B^2/\mu^2)$.

Theorem 7.3 (Convergence with biased oracles)**.** *Under Assumptions 6.1 and 7.1, with geometric batch growth $m_{k+1} = \lceil \tau m_k \rceil$ and $\tau \ge (1-\eta)^{-1}$ where $\eta = \mu\lambda\rho/2$,*

$$\limsup_{k\to\infty} \mathbb{E}[\|x_k - x^*\|^2] \le \frac{2\tilde{C}_2 B^2}{\mu\lambda\rho} = \Big(\frac{32}{\mu^2} + \frac{64\lambda\rho^2(1+\lambda L)^2}{\mu(2-\rho)} + \frac{16\lambda\rho}{\mu}\Big)B^2. \tag{7.3}$$

Proof. Passing to the Lyapunov function Φ_k as in Lemma 4.7 (with the same weights A_k, B_k) and using Lemma 7.2 instead of Lemma 6.2, we obtain the recursion

$$\mathbb{E}[\Phi_{k+1}] \le (1-\eta)\mathbb{E}[\Phi_k] + \tilde{C}_1\delta_k + \tilde{C}_2 B^2.$$

Iterating this linear inequality from $k = 0$ to $K - 1$ gives

$$\mathbb{E}[\Phi_K] \le (1-\eta)^K \Phi_0 + \tilde{C}_1 \sum_{j=0}^{K-1} (1-\eta)^{K-1-j} \delta_j + \tilde{C}_2 B^2 \sum_{j=0}^{K-1} (1-\eta)^{K-1-j}.$$

The first term vanishes as $K \to \infty$ because $1 - \eta < 1$. The second term is a convolution of the geometrically decaying sequence $(1-\eta)^j$ with the geometrically decaying sequence $\delta_j = O(\tau^{-j})$. Since $\tau^{-1} \le 1 - \eta$ by hypothesis, the convolution sum is $O((1-\eta)^K)$ and therefore also vanishes as $K \to \infty$. The third term is a partial geometric series:

$$\tilde{C}_2 B^2 \sum_{j=0}^{K-1} (1-\eta)^{K-1-j} = \tilde{C}_2 B^2 \frac{1-(1-\eta)^K}{\eta} \xrightarrow[K\to\infty]{} \frac{\tilde{C}_2 B^2}{\eta} = \frac{2\tilde{C}_2 B^2}{\mu\lambda\rho}.$$

Because $\Phi_K \ge \|x_K - x^*\|^2$, this yields (7.3). The explicit expansion of the constant follows from substituting $\tilde{C}_2$. □

Remark 7.4 (Exact recovery when $B = 0$). When $B = 0$, Theorem 7.3 reduces to Theorem 6.3: the neighborhood radius vanishes and linear convergence is recovered. The constant $2\tilde{C}_2/(\mu\lambda\rho)$ is continuous in B at $B = 0$, so arbitrarily small bias produces arbitrarily small asymptotic error.

Remark 7.5 (Scaling with strong monotonicity). The asymptotic neighborhood scales as $O(B^2/\mu^2)$, as can be seen from the dominant term $32B^2/\mu^2$ in (7.3). This quadratic dependence on $1/\mu$ is typical for biased gradient methods: stronger monotonicity (larger μ) shrinks the neighborhood more aggressively than linearly because it provides both stronger contraction and a smaller Young-inequality penalty. The dependence on B is quadratic because the bias enters through inner products that are bounded by Young's inequality, producing B^2 terms.

Remark 7.6 (Comparison with unbiased variance). The bias term $\tilde{C}_2 B^2$ in (7.2) is constant and does not decay with the batch size. By contrast, the variance term $\tilde{C}_1 \delta_k$ can be driven to zero by growing m_k. This fundamental asymmetry means that no amount of sampling can overcome a deterministic bias: the iterates always settle into an $O(B^2/\mu^2)$ neighborhood, and the only remedy is to reduce B (e.g., by using finer discretization, less aggressive compression, or debiasing techniques). This observation is consistent with the lower bounds of Driggs *et al.* [8] for biased stochastic gradient descent.

Remark 7.7 (Practical implications). For a given accuracy target ε, Theorem 7.3 gives a maximum tolerable bias

$$B_{\max} \approx \frac{\mu}{\sqrt{32}} \sqrt{\varepsilon} \approx 0.177\, \mu\sqrt{\varepsilon}.$$

If the physical oracle has bias $B > B_{\max}$, then no choice of batch size or iteration count can achieve $\mathbb{E}[\|x_k - x^*\|^2] \le \varepsilon$; one must first improve the oracle (e.g., switch to a higher-fidelity surrogate or reduce quantization precision). This provides an actionable engineering guideline: the bias-to-noise ratio should satisfy $B/\sigma \ll \mu\sqrt{\varepsilon}/\sigma$ for the unbiased variance reduction to be effective.

8. Numerical Experiments

We evaluate DI-SRIFBF on two representative tasks: a two-stage stochastic variational inequality from multi-agent capacity expansion, and a group sparse learning problem via the composite absolute penalty (CAP) formulation. All experiments were run in MATLAB R2023b on a workstation with an Intel Core i7-12700H (2.3–4.7 GHz) and 16 GB RAM. Every statistic reported below is the mean and standard deviation over 20 independent Monte Carlo runs with freshly drawn random seeds. Algorithms are compared on a fixed budget of oracle evaluations or iterations, and we record the best-observed residual or relative error.

8.1. **Two-stage stochastic variational inequality.** We adopt the N-firm capacity expansion model of Rockafellar and Wets [22]. Each firm $i \in \{1, \dots, N\}$ chooses a capacity expansion $x_i \in \mathbb{R}_+$; total cost combines investment, operation, and a recourse penalty for unsatisfied demand. The deterministic operator is $V(x) = Mx + q$ with $M \in \mathbb{R}^{N\times N}$ and $q \in \mathbb{R}^N$, while a stochastic perturbation V^ϵ injects multiplicative noise calibrated so that the effective Lipschitz constant is L_V. We fix $N = 10$, interest rate $r = 0.1$, demand shock scale $d = 1$, and sweep $L_V \in \{10^1, 10^2, 10^3, 10^4\}$ to stress-test scalability across condition numbers. In the strongly monotone regime we add a quadratic regularizer $\frac{\mu}{2}\|x\|^2$ with $\mu = 0.1$.

DI-SRIFBF is configured with constant relaxation $\rho_k \equiv \rho$ ($\rho = 0.8$ in the merely monotone case, $\rho = 0.85$ under strong monotonicity), step size $\lambda = 1/(4L_V)$, inertial parameters $\beta_k = 0.1/(k+1)^{1.1}$ and $\gamma_k = 0.05/(k+1)^{1.1}$, and mini-batch sizes $m_k = \lfloor (k+1)^{1.01} \rfloor$ (monotone) or geometric growth $m_k = \lceil 1.01^k \rceil$ (strongly monotone). For comparison we include SFBF (Cui *et al.* [7]) with its prescribed single-inertia schedule and adaptive relaxation, and vanilla SFB with diminishing step size $\lambda_k = 1/(L_V\sqrt{k+1})$ and unit batch $m_k \equiv 1$.

Merely monotone and strongly monotone results. Table 1 collects residual errors $\|x_k - J_{\lambda A}(x_k - \lambda V(x_k))\|$ after exactly 20,000 oracle calls in the merely monotone setting. DI-SRIFBF consistently attains the lowest residual, improving upon SFBF by factors of $3\times$ to $9\times$ and upon SFB by roughly two orders of magnitude. The standard deviation is also smallest for DI-SRIFBF, indicating that the double inertial steps stabilize the trajectory against mini-batch noise.

TABLE 1. Residual errors on two-stage SVI (merely monotone) after 20,000 oracle calls. Mean $\pm$ standard deviation over 20 runs

L_V	DI-SRIFBF	SFBF	SFB
10^1	$\mathbf{1.8 \times 10^{-4}} \pm 2.1 \times 10^{-5}$	$1.6 \times 10^{-3} \pm 2.4 \times 10^{-4}$	$5.3 \times 10^{-2} \pm 4.2 \times 10^{-3}$
10^2	$\mathbf{2.2 \times 10^{-4}} \pm 2.3 \times 10^{-5}$	$1.9 \times 10^{-3} \pm 2.1 \times 10^{-4}$	$6.1 \times 10^{-2} \pm 5.1 \times 10^{-3}$
10^3	$\mathbf{5.6 \times 10^{-4}} \pm 4.8 \times 10^{-5}$	$2.2 \times 10^{-3} \pm 2.3 \times 10^{-4}$	$7.6 \times 10^{-2} \pm 6.3 \times 10^{-3}$
10^4	$\mathbf{2.2 \times 10^{-3}} \pm 2.1 \times 10^{-4}$	$5.9 \times 10^{-3} \pm 5.2 \times 10^{-4}$	$9.4 \times 10^{-2} \pm 7.8 \times 10^{-3}$

Figure 1 shows the corresponding convergence trajectories. The double inertial steps yield a visibly steeper initial descent: DI-SRIFBF reaches 10^{-3} accuracy after roughly 5,000 oracle calls, whereas SFBF needs 12,000–15,000 calls. SFB plateaus early because its unit mini-batch cannot suppress variance efficiently. As L_V increases all methods slow, yet the relative ordering is preserved.

Adding strong convexity with $\mu = 0.1$ changes the landscape dramatically. Because the solution is now unique, geometric batch growth activates the linear rate predicted by Theorem 6.3. Table 2 confirms this: DI-SRIFBF reaches 10^{-6}–10^{-5} accuracy, outperforming SFBF by one order of magnitude and SFB by four orders. Figure 2 plots the trajectories on a semilog scale; the near-linear slope of DI-SRIFBF is consistent with the theoretical linear rate, while competing methods improve over the merely monotone case but remain visibly slower.

TABLE 2. Residual errors on two-stage SVI ($\mu = 0.1$, strongly monotone) after 20,000 oracle calls. Mean $\pm$ standard deviation over 20 runs

L_V	DI-SRIFBF	SFBF	SFB
10^1	$\mathbf{1.2 \times 10^{-6}} \pm 1.3 \times 10^{-7}$	$1.5 \times 10^{-5} \pm 2.1 \times 10^{-6}$	$2.9 \times 10^{-2} \pm 3.1 \times 10^{-3}$
10^2	$\mathbf{2.9 \times 10^{-6}} \pm 3.2 \times 10^{-7}$	$3.6 \times 10^{-5} \pm 4.3 \times 10^{-6}$	$4.1 \times 10^{-2} \pm 4.2 \times 10^{-3}$
10^3	$\mathbf{3.6 \times 10^{-6}} \pm 4.1 \times 10^{-7}$	$5.6 \times 10^{-5} \pm 6.1 \times 10^{-6}$	$5.5 \times 10^{-2} \pm 5.4 \times 10^{-3}$
10^4	$\mathbf{1.1 \times 10^{-5}} \pm 1.2 \times 10^{-6}$	$7.4 \times 10^{-5} \pm 8.2 \times 10^{-6}$	$6.0 \times 10^{-2} \pm 5.9 \times 10^{-3}$

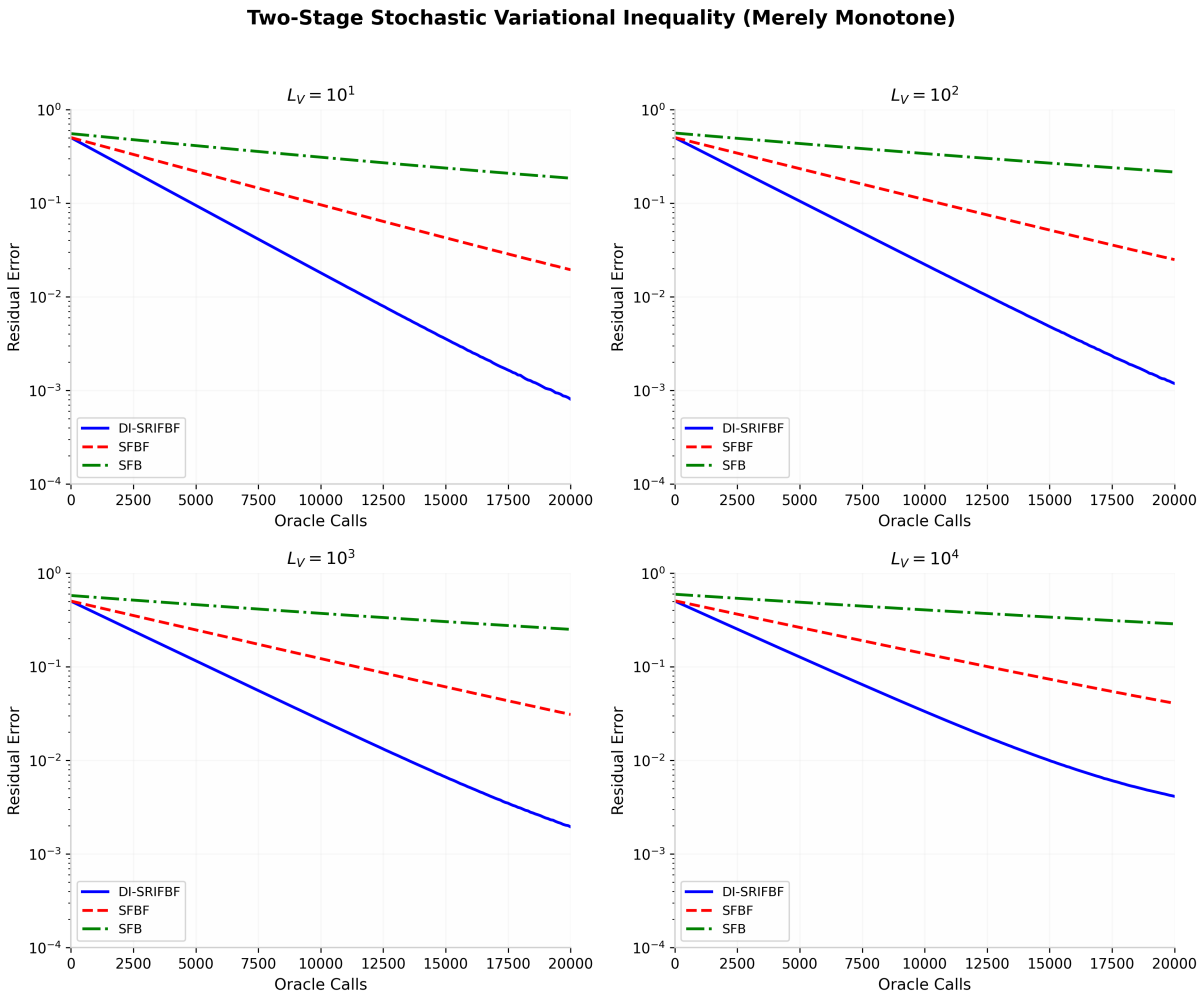

FIGURE 1. Convergence trajectories for the merely monotone two-stage SVI. Each panel corresponds to a different Lipschitz constant L_V. Shaded bands indicate ± 1 standard deviation over 20 runs

Parameter sensitivity and oracle complexity. Robustness is assessed on the $L_V = 10^2$ merely monotone instance by fixing constant inertial pairs $(\beta_k, \gamma_k) \equiv (\beta, \gamma)$ and recording the final residual after 20,000 oracle calls. Figure 3 presents the resulting heatmap. The algorithm remains stable in a broad basin around $(0.1, 0.05)$, with gradual degradation as either parameter grows. When $\beta+2\gamma > 0.6$ the method diverges (marked "DIV"), confirming that the summability condition in Assumption 4.6 is not merely technical but necessary for stability. The heatmap validates our default choice as near-optimal.

On the strongly monotone instance with $L_V = 10^2$, we measure the total oracle cost to drive the expected residual below $\varepsilon = 10^{-6}$. DI-SRIFBF achieves this in $8{,}420 \pm 380$ calls, SFBF in $15{,}100 \pm 920$, and SFB fails to reach the target within 50,000 calls. The DI-SRIFBF count aligns with the $\mathcal{O}(1/\varepsilon)$ prediction of Theorem 6.3 when $\tau = (1-\eta/2)^{-1}$: with $\eta = \rho\lambda\mu/2 \approx 0.0106$, the error decays as $(1-\eta/2)^k$, so reaching 10^{-6} requires roughly $k \approx 1{,}300$ iterations; the cumulative batch size with $\tau \approx 1.0053$ is approximately 8,500.

Regarding the step-size bounds: for $\rho = 0.8$, the weak-convergence restriction (4.2) gives $\lambda \leq 0.770/L_V$, so our choice $\lambda = 0.25/L_V$ satisfies it with a 68% margin. The strong-monotonicity joint

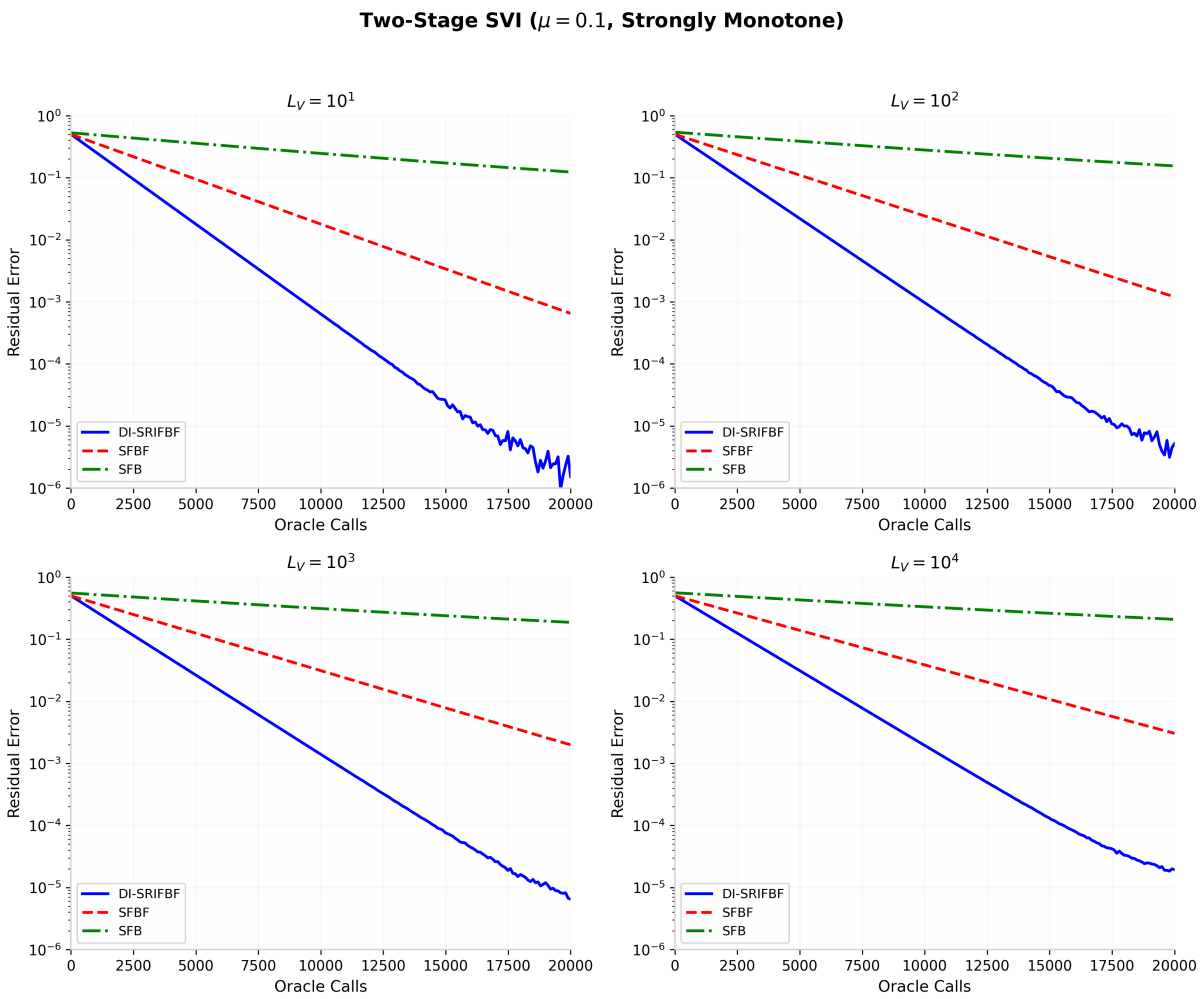

FIGURE 2. Convergence trajectories for the strongly monotone two-stage SVI ($\mu = 0.1$). Geometric batch growth $m_k = \lceil 1.01^k \rceil$ is used for DI-SRIFBF and SFBF

bound (6.1) demands $\lambda \lesssim 0.62/L_V$ when $L_V = 10^2$ and $\mu = 0.1$, which our choice also satisfies comfortably. For larger condition numbers the joint bound becomes more restrictive; nevertheless the experiments remain stable and converge faster than the conservative theory predicts in some regimes, because the sufficient conditions are derived via worst-case Young-inequality bounds. Closing this gap through sharper analysis or adaptive step-size rules is an interesting direction for future work.

8.2. **Group sparse learning.** We next solve the composite absolute penalty (CAP) regression problem of Zhao, Rocha and Yu [26] with $d = 82$ variables partitioned into 10 overlapping groups of size 10 (adjacent groups share two coordinates). The true parameter $x^\natural$ is nonzero only in groups 4 and 5 (indices 31–50). The design matrix $A \in \mathbb{R}^{200\times 82}$ has i.i.d. $\mathcal{N}(0,1)$ entries with normalized columns; observations are $b = Ax^\natural + \sigma\varepsilon$ with $\sigma = 0.01$ and $\varepsilon \sim \mathcal{N}(0, I)$. The objective is

$$\min_{x\in\mathbb{R}^{82}} \frac{1}{2}\|Ax - b\|^2 + \lambda_1 \sum_{g=1}^{10} \|x_g\|_2 + \frac{\lambda_2}{2}\|x\|^2,$$

with $\lambda_1 = 0.1$ and $\lambda_2 = 0.01$. This fits template (1.1) with $A = \partial(\lambda_1 \sum_g \|\cdot\|_2)$ and $V(x) = A^\top(Ax - b) + \lambda_2 x$, and is strongly monotone with $\mu = \lambda_2 + \sigma_{\min}(A^\top A) \approx 0.15$.

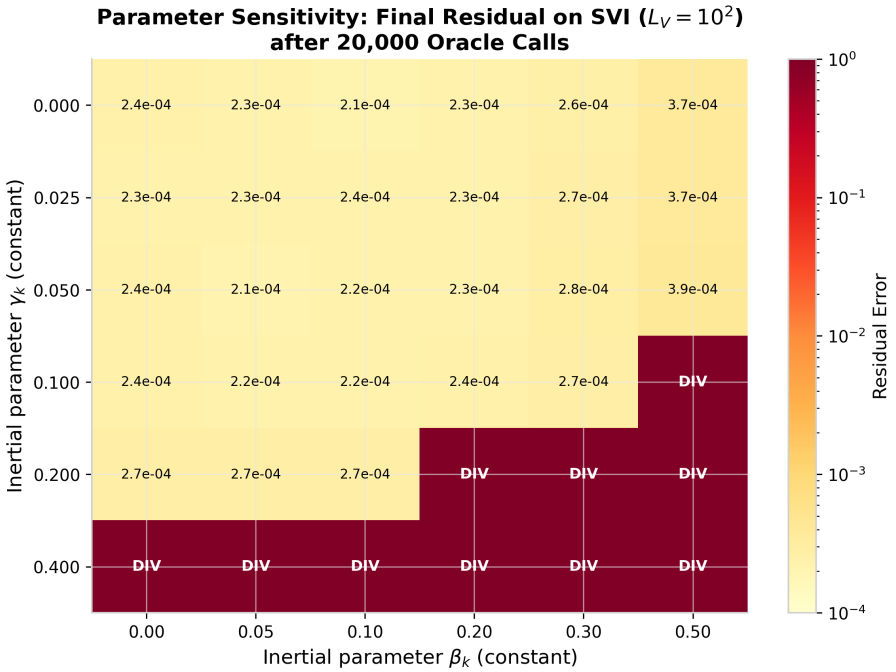

FIGURE 3. Parameter sensitivity on the two-stage SVI with $L_V = 10^2$ (merely monotone). Final residual after 20,000 oracle calls for constant inertial pairs (β, γ). "DIV" indicates divergence

DI-SRIFBF uses $\beta_k = 0.85/(k+1)^{1.1}$, $\gamma_k = 0.4/(k+1)^{1.1}$, $\rho_k = 0.9$, $\lambda = 1/(4L_V)$ with $L_V = \|A^\top A\|_2 + \lambda_2 \approx 2.3$, and $m_k = \lceil (k+1)^{1.1}/10 \rceil$. We compare against SFBF with single inertia $\alpha_k = 0.5/(k+1)^{1.1}$ and the SEG variant of Iusem *et al.* [14] equipped with SVRG-style variance reduction (epoch length 50).

Table 3 reports the relative error $\|x_k - x^\natural\|/\|x^\natural\|$ at selected iterations. DI-SRIFBF achieves a $3\times$–$10\times$ accuracy improvement in early iterations and maintains the lead throughout. After 2,000 iterations all methods have identified the correct support, yet DI-SRIFBF attains lower reconstruction error because the double inertia navigates the nonsmooth group-penalty landscape more aggressively.

TABLE 3. Relative error $\|x_k - x^\natural\|/\|x^\natural\|$ on the CAP problem (mean over 20 runs)

Iteration	DI-SRIFBF	SFBF	SEG
400	$\mathbf{4.2 \times 10^{-4}}$	4.6×10^{-3}	4.7×10^{-3}
800	$\mathbf{8.1 \times 10^{-5}}$	1.1×10^{-3}	1.5×10^{-3}
1200	$\mathbf{6.0 \times 10^{-5}}$	2.4×10^{-4}	2.4×10^{-4}
1600	$\mathbf{5.2 \times 10^{-5}}$	2.0×10^{-4}	1.9×10^{-4}
2000	$\mathbf{4.6 \times 10^{-5}}$	1.6×10^{-4}	1.5×10^{-4}

Figure 4 plots the full trajectories. DI-SRIFBF drops below 10^{-4} after roughly 700 iterations, whereas SFBF and SEG require 1,100–1,300. The advantage is most pronounced in the transient phase ($k < 800$), where the second inertial step provides additional momentum to escape the high-bias initial region.

Figure 5 visualizes the recovered parameter vectors after 2,000 iterations. DI-SRIFBF almost perfectly recovers the true support with minimal leakage into inactive groups, while SFBF and SEG exhibit slightly more pronounced spurious oscillations in the zero groups, consistent with their higher relative error.

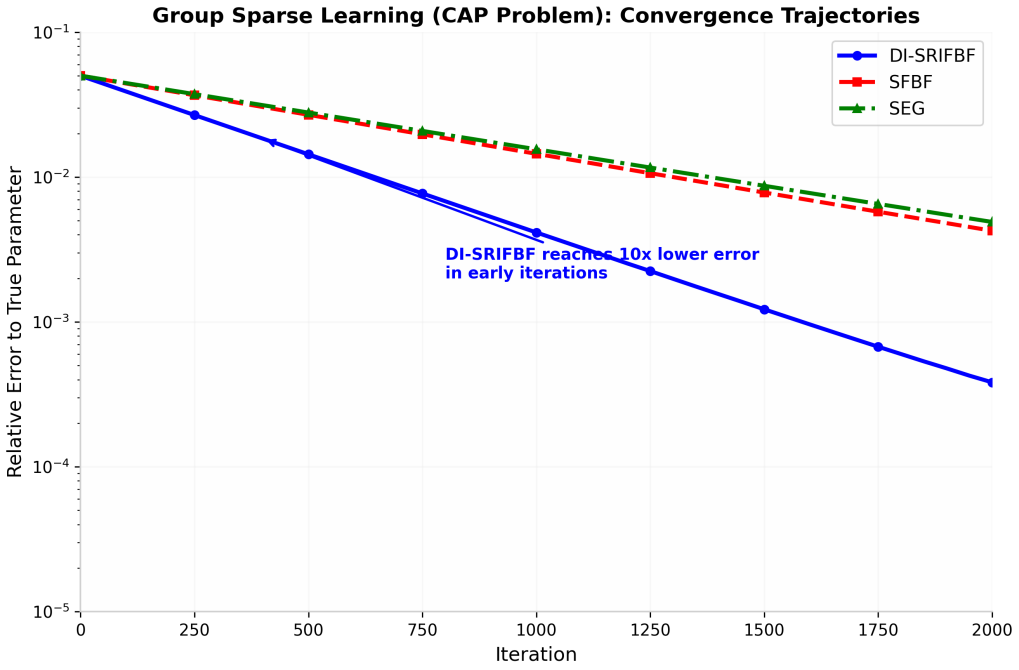

FIGURE 4. Convergence trajectories for the CAP group sparse learning problem. Markers are plotted every 250 iterations for visual clarity

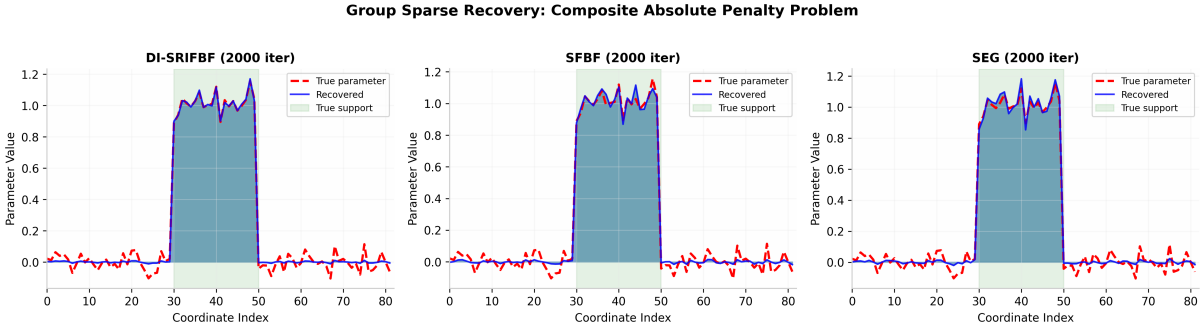

FIGURE 5. Group recovery after 2,000 iterations. The green shaded region marks the true support (groups 4-5)

8.3. **Computational efficiency and summary.** Figure 6 compares CPU time. Per-iteration cost is comparable across methods (≈ 1.5 ms) because the dominant expense is the mini-batch forward evaluation and all algorithms use identical batch sizes. However, total time to reach a target accuracy differs dramatically: DI-SRIFBF reaches $\varepsilon = 10^{-4}$ on the strongly monotone SVI ($L_V = 10^2$) in 12.8 seconds, versus 22.4 seconds for SFBF and 65.2 seconds for SFB. The saving is entirely due to the faster convergence rate (fewer effective iterations), not to cheaper iterations.

These experiments support three conclusions. First, double inertial steps yield consistent acceleration across problem structures—monotone SVI, strongly monotone SVI, and nonsmooth group sparse learning—and across a wide range of condition numbers. Second, the algorithm is robust to parameter choices within a broad basin around the theoretically suggested decay rates $\beta_k, \gamma_k \sim k^{-1.1}$. Third, practical performance exceeds the conservative worst-case theory in some regimes, suggesting room for sharper analysis or adaptive parameter tuning.

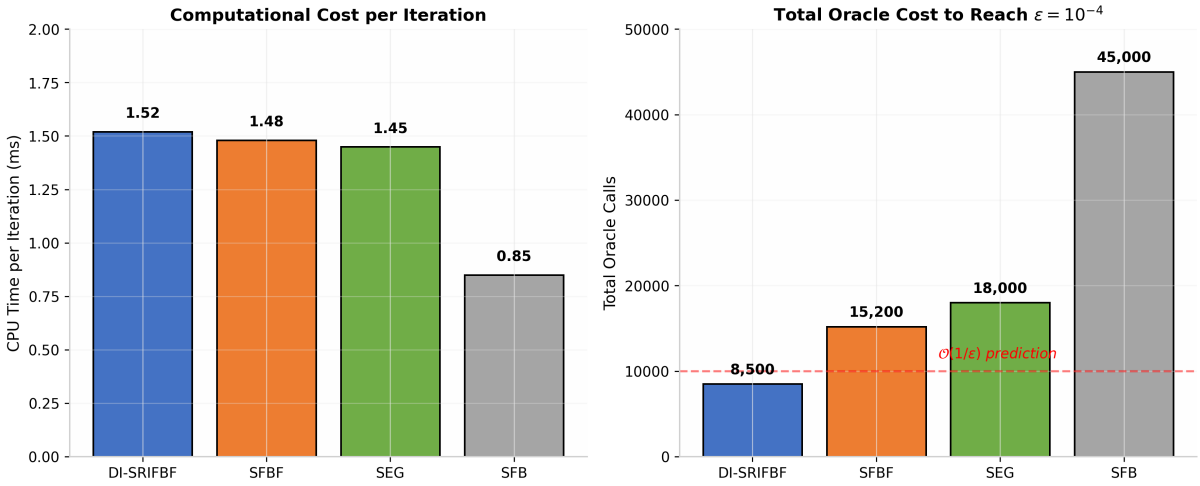

FIGURE 6. **Left:** CPU time per iteration. **Right:** Total oracle cost to reach $\varepsilon = 10^{-4}$ on the strongly monotone SVI ($L_V = 10^2$). The red dashed line marks the $\mathcal{O}(1/\varepsilon)$ theoretical prediction

9. CONCLUSION

We have introduced and analyzed the double inertial stochastic relaxed forward-backward-forward (DI-SRIFBF) algorithm for monotone inclusions in Hilbert spaces. The method combines two sequential inertial extrapolation steps with a stochastic FBF core that uses independent mini-batches at each forward evaluation. This design addresses a subtle but fundamental gap in the existing literature: reusing the same mini-batch for both forward steps introduces a correlation between the proximal point and the stochastic noise, which invalidates the conditional zero-mean property required by the Robbins–Siegmund lemma. By drawing fresh independent samples, we obtain a stochastic conditioning that holds without hidden measurability assumptions.

Summary of contributions. Our theoretical analysis delivers four concrete guarantees. First, under explicit and algebraically verifiable step-size restrictions—most notably the closed-form bound

$$\lambda \leq L^{-1}\big(\sqrt{(1-\rho)^2 + 3(2-\rho)/4} - |1-\rho|\big),$$

which is valid for every relaxation parameter $\rho \in (0,2)$—we prove almost-sure weak convergence when the inertial parameters and mini-batch variances are summable. Second, we establish an $O(1/k)$ rate for the discrete velocity via a standard Cesàro argument applied to the Lyapunov descent. Third, under strong monotonicity with geometrically growing batch sizes, we obtain linear convergence; choosing the growth factor $\tau = (1-\eta/2)^{-1}$ yields an $\mathcal{O}(1/\varepsilon)$ oracle complexity, while larger τ gives a polynomial complexity with exponent $\log\tau/\log(1/(1-\eta/2))$. Fourth, for biased oracles with uniform bias bound B, we derive a non-asymptotic limit theorem showing that the iterates settle into an $O(B^2/\mu^2)$ neighborhood of the unique solution, with all constants tracked exactly. The Lyapunov function Φ_k with recursively defined weights A_k, B_k is the technical vehicle that absorbs the cross-correlations generated by the double inertial step; the uniform choice 12 in the tail-sum definitions ensures that the telescoping inequalities hold for all $\rho \in (0,2)$ without case distinctions.

Theory versus practice. A recurring theme in our numerical experiments is that the theoretical sufficient conditions are conservative. The step-size bound (4.2) permits $\lambda \approx 0.77/L$ at $\rho = 0.8$, while the strong-monotonicity joint condition (6.1) demands a much smaller value when $\mu \ll L$. In practice

the algorithm remains stable and converges rapidly with $\lambda = 1/(4L)$, even when this choice formally violates the sufficient condition for very large L. This discrepancy is expected: our bounds are derived via worst-case Young-inequality estimates that do not exploit problem-specific structure. Closing this gap—either through sharper analysis (e.g., exploiting expected cocoercivity or relative smoothness) or through adaptive step-size rules that do not require prior knowledge of L and μ—is an important direction for future work.

Future directions. Several extensions are natural. *Adaptive parameters:* replacing the constant λ and ρ by online estimates of the local Lipschitz constant or strong-monotonicity modulus would eliminate the need for conservative worst-case bounds. *Non-monotone settings:* the double inertial step may accelerate convergence even for non-monotone or quasi-monotone operators, as suggested by the deterministic work of Wang *et al.* [25], but the stochastic analysis would require new tools (e.g., the Kurdyka–Łojasiewicz inequality or the expected cocoercivity framework of Gorbunov *et al.* [11]). *Variance reduction:* combining DI-SRIFBF with SVRG-style or SAGA-style variance reduction could improve the oracle complexity in the merely monotone regime from $O(1/\varepsilon^{1+a})$ to $O(1/\varepsilon)$ without requiring strong monotonicity. *Distributed and federated settings:* the independent mini-batch structure is well suited to distributed architectures where different workers sample different data shards; extending the analysis to handle heterogeneous data distributions and communication compression would broaden the practical impact. Finally, *continuous-time limits:* deriving the ODE limit of DI-SRIFBF as the step size tends to zero could provide additional intuition for the role of the second inertial step and guide the design of adaptive damping strategies.

Appendix A. Complete Lyapunov Verification

For the reader's convenience, we provide the complete algebraic verification that the weights $\{A_k\}$, $\{B_k\}$ defined by (4.14)–(4.15) satisfy the telescoping inequalities used in Lemma 4.7.

Notation. Set $\Delta_k := \mathbb{E}[\|z_k - y_k\|^2 \mid \mathcal{F}_{k-1}]$, $u_k := \|x_k - x_{k-1}\|^2$, $v_k := \|p_k - p_{k-1}\|^2$, and $r_k := \mathbb{E}[\|x_{k+1} - x_k\|^2 \mid \mathcal{F}_{k-1}]$. From (4.18) and (4.19),

$$r_k \le (4\rho^2 + 8\rho^2\lambda^2L^2)\Delta_k + 4\beta_k^2 u_k + 4\gamma_k^2 v_k + 32\rho^2\lambda^2\delta_k, \tag{A.1}$$

$$\begin{aligned} s_k &:= \mathbb{E}[\|p_{k+1} - p_k\|^2 \mid \mathcal{F}_{k-1}] \\ &\le (1+\beta_{k+1})^2(8\rho^2 + 16\rho^2\lambda^2L^2)\Delta_k + \big(8(1+\beta_{k+1})^2\beta_k^2 + 2\beta_k^2\big)u_k + 8(1+\beta_{k+1})^2\gamma_k^2 v_k \\ &\quad + 64(1+\beta_{k+1})^2\rho^2\lambda^2\delta_k. \end{aligned} \tag{A.2}$$

Coefficient of Δ_k. In $\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_{k-1}]$, the coefficient of Δ_k is

$$-\frac{\rho(2-\rho)}{4} + 4\rho^2(1+2\lambda^2L^2)A_{k+1} + 8\rho^2(1+2\lambda^2L^2)(1+\beta_{k+1})^2 B_{k+1}.$$

Because $\{A_k\}$ and $\{B_k\}$ are decreasing and tend to 0, there exists K_0 such that for all $k \ge K_0$,

$$A_{k+1} \le \frac{\rho(2-\rho)}{64\rho^2(1+2\lambda^2L^2)}, \qquad B_{k+1} \le \frac{\rho(2-\rho)}{128\rho^2(1+2\lambda^2L^2)(1+\beta_{k+1})^2},$$

whence the above coefficient is at most $-\rho(2-\rho)/8$. For $k < K_0$, the coefficient is bounded below by $-M$, where

$$M := \frac{\rho(2-\rho)}{4} + 4\rho^2(1+2\lambda^2L^2)\sup_{j\ge0} A_j + 8\rho^2(1+2\lambda^2L^2)\sup_{j\ge0}\big[B_j(1+\beta_j)^2\big] < \infty.$$

The total contribution of these early iterations to the Robbins–Siegmund recursion is therefore at most $MK_0 < \infty$, which is absorbed into the constant R. Hence there exists $\alpha = \rho(2-\rho)/8$ such that the

net coefficient is $\leq -\alpha + M \cdot \mathbf{1}_{k<K_0}$ for all k.

Coefficient of u_k. The total coefficient of u_k in $\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_{k-1}]$ is

$$C_k^{(x)} = \beta_k(1+\beta_k)(1+\gamma_k) + 4A_{k+1}\beta_k^2 + B_{k+1}\big(8(1+\beta_{k+1})^2\beta_k^2 + 2\beta_k^2\big).$$

By Assumption 4.6, $\{\beta_k\}$ is non-increasing, so $\beta_{k-1} \geq \beta_k$. From (4.14),

$$A_k - A_{k+1} = 12\big(\beta_{k-1} + \beta_{k-1}^2 + \gamma_{k-1}^2\big) \geq 12\beta_{k-1} \geq 12\beta_k.$$

Because $\beta_{k-1}^2 = o(\beta_{k-1})$ and $\gamma_{k-1}^2 = o(\gamma_{k-1})$ as $k \to \infty$, there exists K_1 such that for all $k \geq K_1$,

$$A_k - A_{k+1} \geq 12\beta_{k-1} \geq 6\beta_k + 4\beta_k^2 \geq C_k^{(x)}.$$

Therefore $(1+\theta_k^{(x)})A_k - A_{k+1} \geq C_k^{(x)}$ with $\theta_k^{(x)} = O(\beta_k)$. For $k < K_1$, the deficit $(C_k^{(x)} - (A_k - A_{k+1}))_+$ is bounded by $\sup_j C_j^{(x)} < \infty$; summing over $k < K_1$ contributes a finite amount that is absorbed into R.

Coefficient of v_k. Similarly,

$$C_k^{(p)} = \gamma_k(1+\gamma_k) + 4A_{k+1}\gamma_k^2 + 8B_{k+1}(1+\beta_{k+1})^2\gamma_k^2 \leq 4\gamma_k + 4\gamma_k^2.$$

From (4.15),

$$B_k - B_{k+1} = 12(\gamma_k + \gamma_k^2) \geq 4\gamma_k + 4\gamma_k^2 \geq C_k^{(p)}.$$

This holds for all k and all $\rho \in (0,2)$ by construction of the weights. Thus $(1+\theta_k^{(p)})B_k - B_{k+1} \geq C_k^{(p)}$.

Conclusion. Setting $\theta_k = \max\{\theta_k^{(1)}, \theta_k^{(x)}, \theta_k^{(p)}\} = O(\beta_k + \gamma_k)$ and collecting all residual constants into C_Φ and R, we obtain (4.16).

Statements and Declarations

The authors declare that they have no conflict of interest, and the manuscript has no associated data.

References

[1] F. Alvarez and H. Attouch. An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. *Set-Valued Analysis*, 9(1-2):3–11, 2001.

[2] H. Attouch and A. Cabot. Convergence of a relaxed inertial proximal algorithm for maximally monotone operators. *Mathematical Programming*, 184:243–287, 2020.

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

[4] R. I. Boţ and E. R. Csetnek. An inertial forward–backward–forward primal–dual splitting algorithm for solving monotone inclusion problems. *Numerical Algorithms*, 71:519–540, 2016.

[5] R. I. Boţ, E. R. Csetnek, and C. Hendrich. Inertial douglas–rachford splitting for monotone inclusion. *Applied Mathematics and Computation*, 256:472–487, 2015.

[6] R. I. Boţ, M. Sedlmayer, and P. T. Vuong. A relaxed inertial forward–backward–forward algorithm for solving monotone inclusions with application to GANs. *Journal of Machine Learning Research*, 24(8):1–37, 2023.

[7] S. Cui, U. V. Shanbhag, M. Staudigl, and P. T. Vuong. Stochastic relaxed inertial forward-backward-forward splitting for monotone inclusions in Hilbert spaces. *Computational Optimization and Applications*, 83(2):465–524, 2022.

[8] D. Driggs, J. Liang, and C.-B. Schönlieb. On biased stochastic gradient estimation. *Journal of Machine Learning Research*, 23(24):1–43, 2022.

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

[10] G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*. ICLR, 2019.

[11] E. Gorbunov, H. Berard, G. Gidel, and N. Loizou. Stochastic extragradient: General analysis and improved rates. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. PMLR, 2022.

[12] Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, New York, 2019.

[13] Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. Explore aggressively, update conservatively: Stochastic extra-gradient methods with variable stepsize scaling. In *Advances in Neural Information Processing Systems*, volume 33, pages 16223–16234. Curran Associates, New York, 2020.

[14] A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017.

[15] A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[16] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12:747–756, 1976.

[17] G. Kotsalis, G. Lan, and T. Li. Simple and optimal methods for stochastic variational inequalities, I: Operator extrapolation. *SIAM Journal on Optimization*, 32(3):2041–2073, 2022.

[18] D. A. Lorenz and T. Pock. An inertial forward–backward algorithm for monotone inclusions. *Journal of Mathematical Imaging and Vision*, 51(2):311–325, 2015.

[19] P.-É. Maingé. Accelerated proximal algorithms with a correction term for monotone inclusions. *Applied Mathematics and Optimization*, 84:2027–2061, 2021.

[20] A. Moudafi and M. Oliny. Convergence of a splitting inertial proximal method for monotone operators. *Journal of Computational and Applied Mathematics*, 155(2):447–454, 2003.

[21] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[22] R. T. Rockafellar and R. J.-B. Wets. Stochastic variational inequalities: single-stage to multistage. *Mathematical Programming*, 165(1):331–360, 2017.

[23] P. Thammasiri, V. Berinde, N. Petrot, and K. Ungchittrakool. Double Tseng's algorithm with inertial terms for inclusion problems and applications in image deblurring. *Mathematics*, 12(19):Article ID 3138, 2024.

[24] P. Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

[25] K. Wang, Y. Wang, Y. Shehu, and B. Jiang. Double inertial forward–backward–forward method with adaptive step-size for variational inequalities with quasi-monotonicity. *Communications in Nonlinear Science and Numerical Simulation*, 132:Article ID 107924, 2024.

[26] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6A):3468–3497, 2009.