Direct Acceleration of Stochastic Root-Finding Without Variance Reduction and Regularization

TaeHo Yoon **Nicolas Loizou**
Department of Applied Mathematics & Statistics
Mathematical Institute for Data Science
Johns Hopkins University
{tyoon7,nloizou}@jhu.edu

Abstract

Acceleration for deterministic root-finding problems has been extensively studied in recent years; specifically, the anchor-based, or Halpern-type methods achieve optimal convergence rates with respect to the operator norm. However, acceleration via these methods does not directly carry over to stochastic setting due to accumulation of errors, unless one enforces diminishing variance via increasing batch sizes or variance reduction techniques. In this work, we show that another class of acceleration, namely the *dual-anchor* mechanism, extends to the stochastic setting without such error accumulation, in contrast to anchor-based algorithms. Consequently, we cleanly achieve $\mathcal{O}(\epsilon^{-3})$ complexity with iteration-independent batch size, without any variance reduction or double-loop recursive regularization, for stochastic root-finding (*resp.* fixed-point) problems with cocoercivity (*resp.* square-nonexpansivity) in expectation. For strongly monotone operators, the same algorithm attains a sharper $\widetilde{\mathcal{O}}(\epsilon^{-2})$ complexity, nearly matching the lower bound in terms of ϵ-dependence.

1 Introduction

Root-finding problems for an operator $\mathbb{F}\colon \mathbb{R}^d \to \mathbb{R}^d$, stated as

$$\operatorname*{find}_{x\in\mathbb{R}^d} \quad \mathbb{F}(x) = 0, \tag{1}$$

generalize minimization problems and have been studied under distinct formulations including monotone inclusion [36, 42, 3, 43], minimax optimization [38, 19, 33, 13, 35, 31], equilibrium search [47, 17, 46], or fixed-point problems [24, 34, 21, 52, 45]. In the past few years, there have been significant developments in accelerated algorithms for reducing the squared operator/residual norm with the optimal $\|\mathbb{F}(x_k)\|^2 = \mathcal{O}(1/k^2)$ rate [29, 16, 22, 40, 54, 27]. A recurring message from the optimization literature, however, is that deterministic acceleration is often *brittle* under oracle errors; in the stochastic setting where $\mathbb{F}(x) = \mathbb{E}\left[\widehat{\mathbb{F}}(x;\xi)\right]$ and the algorithm uses $\widehat{\mathbb{F}}(x;\xi)$, error may accumulate quickly with these accelerated methods and break their fast convergence from the noiseless setting. Consequently, for stochastic optimization problems, taking advantage of the study of acceleration was not straightforward—it required decaying oracle inexactness, variance reduction, restart mechanisms, or stronger assumptions on the problem [15, 18, 28, 9]. Specifically for stochastic monotone inclusion problems, stochastic Halpern iteration and accelerated (anchored) extragradient type analyses suffered precisely from these issues [27, 9, 10].

We take a different viewpoint toward handling stochastic instability. Recent work [56] showed that the acceleration mechanism in fixed-point and minimax problems is not unique: besides anchor-based algorithms [29, 22, 54, 27, 51], there exist their *H-dual* algorithms, which display materially different

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Comparison of stochastic algorithms for finding an ϵ-approximate solution with $\mathbb{E}[\|\mathbb{F}(x)\|] \leq \epsilon$ for $1/L$-cocoercive operator $\mathbb{F}$.

Algorithms	Complexity	Last-iterate	Variance-reduction free	Single-loop	Algorithm parameters	Theory requirements
SGDA/SEG [20]	$\mathcal{O}(\epsilon^{-4})^{\dagger}$	✗	✓	✓	L, B	D, σ
S-HALPERN-PAGE [9]	$\mathcal{O}(\epsilon^{-3})$	✓	✗	✓	$L, p_k, S_1^{(k)}, S_2^{(k)}$	D, σ
S-HALPERN-VR-FINITE [10]	$\widetilde{\mathcal{O}}(n + \sqrt{n}\,\epsilon^{-1})^{\ddagger}$	✓	✗	✓	L, B, λ_k; p_k or M_k	D, n
RAIN [12]	$\widetilde{\mathcal{O}}(\epsilon^{-2})$	✓	✓	✗*	$L, \lambda, \gamma, N_s, K_s$	D, σ
S-DUAL-OHM (This work)	$\mathcal{O}(\epsilon^{-3})$	✓	✓	✓	L, B, N	D, σ

SGDA denotes Stochastic Gradient Descent-Ascent $x_{k+1} = x_k - \alpha \mathbb{F}(x_k; \xi_k)$ and SEG denotes stochastic version of Extragradient [23]. "Last-iterate" indicates whether the stated guarantee is for the final iterate and not the averaged, best, or uniform random iterate. "Variance-reduction free" indicates whether the algorithm avoids using variance reduction techniques. "Single-loop" indicates whether the algorithm avoids using outer–inner loop structure (e.g., regularization techniques). "Algorithm parameters" lists the hyperparemeters whose values should be specified for the run, where the cocoercivity/Lipschitzness parameter L is used for determining the step-sizes of the algorithm. "Theory requirements" lists the information of problem-related parameters that are needed for selecting the hyperparameters that theoretically guarantee $\mathbb{E}[\|\mathbb{F}(x)\|] \leq \epsilon$. Here N is the total number of algorithm iterations, $D = \|x_0 - x_\star\|$, σ is a bound on the operator variance, and n is the number of component operators in the case where the problem has a finite-sum structure. The $\widetilde{\mathcal{O}}(\cdot)$ notation hides logarithmic factors.

†[20, Corollary E.4] shows this for uniform random iterate of SEG, and a similar proof for SGDA can be derived.

‡Works only for finite-sum problems where $\mathbb{F}(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{F}(\cdot; \xi_i)$.

*[12] presents a single-loop version of RAIN for numerical experiments (which we use in our experiments for comparison), but their theoretical convergence analysis still requires a nontrivial double-loop structure.

update rules but share exactly the same deterministic last-iterate rate with the optimal anchor-based acceleration. Hence, the deterministic worst-case complexity alone does not identify a single preferred accelerated method. In this paper, we demonstrate that the stochastic extension can break this tie: the *Stochastic Dual Optimal Halpern Method (S-Dual-OHM)* attains $\mathbb{E}\left[\|\mathbb{F}(x_{N-1})\|\right] \leq \epsilon$ with $\mathcal{O}(\epsilon^{-3})$ oracle complexity, with iteration complexity $N = \mathcal{O}(\epsilon^{-1})$ and with constant mini-batch size $\mathcal{O}(\epsilon^{-2})$. This is done cleanly and more directly without assuming diminishing variance or using variance reduction techniques, which is not what stochastic anchoring methods were capable of [27, 9, 10].

In our view, this result has a conceptual implication beyond simply providing another efficient stochastic algorithm. It challenges the folklore view that acceleration, as studied in deterministic optimization, is brittle and incompatible with stochastic noise. Equivalently, optimal deterministic methods can behave very differently under stochastic perturbations, and selecting the right representation of acceleration may enable the design of efficient stochastic algorithms.

Contributions. We summarize our main contributions as follows.

- ◇ We show that under cocoercivity of $\mathbb{F}$ in expectation, our proposed S-Dual-OHM with mini-batching over $\widehat{\mathbb{F}}(\cdot; \xi)$ with *constant step-size* $\alpha \in \left(0, \frac{2}{L}\right]$ and *constant batch size* $B = \mathcal{O}(\epsilon^{-2})$ achieves the bound $\mathbb{E}\left[\|\mathbb{F}(x_{N-1})\|\right] \leq \epsilon$ within $N = \mathcal{O}(\epsilon^{-1})$ iterations. The resulting total oracle complexity $\mathcal{O}(\epsilon^{-3})$, to the best of our knowledge, has not been achieved in prior work without variance reduction or double-loop regularization techniques (Table 1).
- ◇ We show that when $\mathbb{F}$ is additionally strongly monotone, the same S-Dual-OHM algorithm can be early-stopped in $k = \widetilde{\mathcal{O}}(L/\mu \log \epsilon^{-1})$ iterations to attain $\mathbb{E}\left[\|\mathbb{F}(x_k)\|\right] \leq \epsilon$. This yields the oracle complexity $\widetilde{\mathcal{O}}(\epsilon^{-2})$, which has near-optimal dependence on ϵ.
- ◇ We provide numerical experiments demonstrating that Dual-OHM is indeed more robust to stochastic noise than Halpern iteration, and that S-Dual-OHM can effectively solve stochastic root-finding problems.

2 Related Work

Fixed-point problems, Halpern iteration and acceleration. For fixed-point problems

$$\operatorname*{find}_{x \in \mathbb{R}^d} \quad x = \mathbb{T}(x), \tag{2}$$

with a nonexpansive operator $\mathbb{T}\colon \mathbb{R}^d \to \mathbb{R}^d$, the Halpern iteration $x_{k+1} = \beta_k x_0 + (1 - \beta_k)\mathbb{T}x_k$ has been studied over a long history in the literature [21, 52], and has been found to provide accelerated worst-case convergence rate $\|x_k - \mathbb{T}(x_k)\| = \mathcal{O}(1/k)$ with $\beta_k = \mathcal{O}(1/k)$ [45, 16, 29, 22, 25]. The

best choice $\beta_k = \frac{1}{k+2}$, due to [29, 22], was later shown to yield an optimal rate $\|x_k - \mathbb{T}(x_k)\|^2 \leq \frac{4\|x_0 - x_\star\|^2}{(k+1)^2}$, exactly matching the lower bound [40]. More recently, it was discovered that the same exact optimal rate for a fixed finite horizon N can be achieved by distinct types of acceleration forming an infinite family of algorithms [58] including Dual-OHM [56], the basis of this work.

Acceleration of minimax optimization and monotone inclusion. Minimax optimization with convex-concave objective can be recast into monotone inclusion, i.e., root-finding problem (1) with saddle gradient operator $\mathbb{F}$, which is a monotone operator [42, 17]. Some early works including [44, 16] first connected Halpern acceleration to these problems, and [16] achieved near-optimal complexity for Lipschitz continuous monotone inclusions. The remaining logarithmic gap was removed by the Extra Anchored Gradient algorithm [54], and the so-called anchor acceleration for minimax optimization and monotone inclusion has been studied extensively since then [27, 51, 55, 48, 11, 5, 1, 26]. Several works proposed its extension via distinct interpretation in connection with Nesterov momentum [6, 49]. On the other hand, the dual-anchor acceleration analogous to Dual-OHM from fixed-point problems is achievable in this setting as well [56].

Residual-reducing algorithms for stochastic fixed-point and monotone inclusion problems. The Fast Extragradient algorithm [27], an anchor-based accelerated minimax algorithm, was proposed with its stochastic extension under the assumption that the operator's variance at k-th iteration diminishes at the order of $\mathcal{O}(1/k)$. Then [9] showed that stochastic Halpern iteration and its minimax variants, combined with PAGE variance reduction [28] can reduce the complexity of achieving $\mathbb{E}\left[\|\mathbb{F}(x_k)\|\right] \leq \epsilon$ to $\mathcal{O}(\epsilon^{-3})$, which could be further improved to $\widetilde{\mathcal{O}}(n + \sqrt{n}\epsilon^{-1})$ [10, 50] in the finite-sum setting where $\mathbb{F}$ is the average of $n = o(\epsilon^{-3})$ sample operators. Near-optimal complexity of $\widetilde{\mathcal{O}}(\epsilon^{-2})$ for stochastic monotone inclusion has been achieved in [12] with the matching $\Omega(\epsilon^{-2})$ lower bound via more sophisticated recursive anchoring (regularization) algorithm. On another line of work, [7, 41, 8] studied stochastic Krasnosel'skiĭ–Mann or Halpern iterations in general normed spaces, where the lower bound of $\Omega(\epsilon^{-3})$ was established [8].

Broader view of extending acceleration to stochastic settings. In smooth convex minimization, it is now standard that accelerated methods are fragile to persistent oracle error: preserving acceleration under stochasticity typically requires either variance reduction or catalyst-type outer–inner regularization [15, 30, 2]. A similar viewpoint has been adopted for monotone root-finding problems in prior works [9, 10, 12]. Our approach departs from this line of work. Rather than stabilizing Halpern-type acceleration by additional variance-reduction or regularization, we leverage the fact that deterministic acceleration is not unique, and there exist distinct optimal accelerated algorithms with the same worst-case deterministic rate. We deliver the new observation that among them, Dual-OHM is intrinsically less sensitive to stochastic perturbations than Halpern-type algorithms, enabling direct stochastic extension.

3 Preliminaries and Assumptions

In this section, we establish some necessary preliminary concepts for subsequent analysis.

3.1 Monotonicity, cocoercivity, and nonexpansiveness of operators

We say that an operator $\mathbb{F}\colon \mathbb{R}^d \to \mathbb{R}^d$ is

- *Monotone* if $\langle \mathbb{F}(x) - \mathbb{F}(y), x - y\rangle \geq 0$ for all $x, y \in \mathbb{R}^d$
- *μ-strongly monotone* if $\langle \mathbb{F}(x) - \mathbb{F}(y), x - y\rangle \geq \mu \|x - y\|^2$ for all $x, y \in \mathbb{R}^d$
- *M-Lipschitz* if $\|\mathbb{F}(x) - \mathbb{F}(y)\| \leq M\|x - y\|$ for all $x, y \in \mathbb{R}^d$
- *$1/L$-Cocoercive* if $\langle \mathbb{F}(x) - \mathbb{F}(y), x - y\rangle \geq \frac{1}{L}\|\mathbb{F}(x) - \mathbb{F}(y)\|^2$ for all $x, y \in \mathbb{R}^d$

where $\mu, M, L > 0$. In particular, $1/L$-cocoercivity implies monotonicity and L-Lipschitzness, but not vice versa. If $\mathbb{F}$ is M-Lipschitz and μ-strongly monotone, then it is $1/L$-cocoercive with $L = M^2/\mu$ [17]. When $\mathbb{F}$ is $1/L$-cocoercive, the problem (1) can be recast into a fixed-point problem (2) where

$\mathbb{T} = \mathbb{I} - \alpha\mathbb{F}\colon \mathbb{R}^d \to \mathbb{R}^d$ is *nonexpansive*, i.e., 1-Lipschitz for $0 < \alpha \leq \frac{2}{L}$ [3, Proposition 4.39], where $\mathbb{I}\colon \mathbb{R}^d \to \mathbb{R}^d$ is the identity operator. This is because

$$\operatorname{Zer}(\mathbb{F}) := \{x \in \mathbb{R}^d \,|\, \mathbb{F}(x) = 0\} = \{x \in \mathbb{R}^d \,|\, \mathbb{T}(x) = x\} := \operatorname{Fix}(\mathbb{T}).$$

Given an unconstrained minimax problem

$$\underset{u\in\mathbb{R}^m}{\text{minimize}} \ \underset{v\in\mathbb{R}^n}{\text{maximize}} \ \Phi(u, v) \tag{3}$$

with a convex-concave (*resp.* μ-strongly-convex-strongly-concave) and M-smooth $\Phi\colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, its saddle operator

$$\mathbb{F}(u, v) = (\nabla_u \Phi(u, v), -\nabla_v \Phi(u, v))$$

is monotone (*resp.* μ-strongly monotone) and M-Lipschitz [17].

3.2 Stochastic oracle models

We consider an unbiased stochastic operator oracle with bounded variance.

Assumption 3.1 (Stochastic operator oracle)**.** Given $x \in \mathbb{R}^d$, one can query $\widehat{\mathbb{F}}(x; \xi)$ such that

$$\mathbb{E}\left[\widehat{\mathbb{F}}(x; \xi)\right] = \mathbb{F}(x), \qquad \mathbb{E}\left[\left\|\widehat{\mathbb{F}}(x; \xi) - \mathbb{F}(x)\right\|^2\right] \leq \sigma^2.$$

For our main result (Theorem 4.1), we additionally assume the following condition.

Assumption 3.2 (Cocoercivity in expectation)**.** For any $x, y \in \mathbb{R}^d$, we have

$$\langle \mathbb{F}(x) - \mathbb{F}(y), x - y\rangle \geq \mathbb{E}\left[\frac{1}{L}\left\|\widehat{\mathbb{F}}(x; \xi) - \widehat{\mathbb{F}}(y; \xi)\right\|^2\right].$$

This condition holds for some natural cases, including: **(i)** stochastic oracle with additive noise $\widehat{\mathbb{F}}(x; \xi) = \mathbb{F}(x) + \zeta(\xi)$, where the noise $\zeta(\xi)$ is independent of x, **(ii)** finite sum or expectation models where each sample operator is $1/L$-cocoercive, and as its special case, **(iii)** sample operators that are each μ-strongly monotone ($\mu > 0$) and M-Lipschitz, which implies sample-wise cocoercivity [17]. Note that Assumption 3.2 does not capture general monotone and Lipschitz sample operators or the cases where the noise $\zeta(\xi)$ depends on x. Similar assumptions or even stronger samplewise cocoercivity have been commonly used in stochastic optimization [39, 2, 32, 37, 4, 59, 57, 14, 10].

Given batch size $B \geq 1$, we denote $\mathbb{F}_{\mathcal{B}}(x) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\mathbb{F}}(x; \xi_b)$, where ξ_b are conditionally independent samples given x and $\mathcal{B} = \{\xi_1, \ldots, \xi_B\}$. Then $\mathbb{F}_{\mathcal{B}}(x)$ is an unbiased estimator of $\mathbb{F}(x)$ with $\mathbb{E}\left[\|\mathbb{F}_{\mathcal{B}}(x) - \mathbb{F}(x)\|^2\right] \leq \frac{\sigma^2}{B}$ and $\langle \mathbb{F}(x) - \mathbb{F}(y), x - y\rangle \geq \mathbb{E}\left[\frac{1}{L}\|\mathbb{F}_{\mathcal{B}}(x) - \mathbb{F}_{\mathcal{B}}(y)\|^2\right]$ by Jensen's inequality. For $0 < \alpha \leq \frac{2}{L}$, if we let $\widehat{\mathbb{T}}(\cdot; \xi) = \mathbb{I} - \alpha\widehat{\mathbb{F}}(\cdot; \xi)$ and $\mathbb{T}_{\mathcal{B}}(\cdot) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\mathbb{T}}(\cdot; \xi_b)$, then these are unbiased estimators of the nonexpansive $\mathbb{T} = \mathbb{I} - \alpha\mathbb{F}$ with variance bounds $\alpha^2\sigma^2$ and $\frac{\alpha^2\sigma^2}{B}$, respectively. Furthermore, we have: $\mathbb{E}\left[\|\mathbb{T}_{\mathcal{B}}(x) - \mathbb{T}_{\mathcal{B}}(y)\|^2\right] \leq \|x - y\|^2$ (square-nonexpansivity in expectation).

3.3 Known acceleration in fixed-point problems and its extension

Our general strategy is to reformulate (1) with cocoercive $\mathbb{F}$ into (2) with $\mathbb{T} = \mathbb{I} - \alpha\mathbb{F}$ and build upon deterministic acceleration results. Consider the following algorithms: given $y_0 \in \mathbb{R}^d$,

$$y_{k+1} = \frac{1}{k+2}y_0 + \frac{k+1}{k+2}\mathbb{T}y_k \tag{OHM}$$

and, with the convention $\mathbb{T}y_{-1} = y_0$,

$$y_{k+1} = y_k + \frac{N-k-1}{N-k}\left(\mathbb{T}y_k - \mathbb{T}y_{k-1}\right). \tag{Dual-OHM}$$

Here, OHM is an anytime algorithm whose update is defined for all $k = 0, 1, \ldots$, while for Dual-OHM, the terminal iteration number N is fixed, and the update is defined only for $k = 0, \ldots, N-2$. This is a substantive distinction that makes the two algorithms behave differently under stochastic extension (see the discussion in Section 4.1.1 for further details).

Proposition 3.3 (Deterministic acceleration [29, 22, 56])**.** For a nonexpansive operator $\mathbb{T}\colon \mathbb{R}^d \to \mathbb{R}^d$ with a fixed point $y_\star \in \operatorname{Fix}\mathbb{T}$ and $N \geq 2$, both OHM and Dual-OHM exhibit the rate

$$\|y_{N-1} - \mathbb{T}(y_{N-1})\|^2 \leq \frac{4\|y_0 - y_\star\|^2}{N^2}.$$

With $\mathbb{T} = \mathbb{I} - \alpha\mathbb{F}$ and $\alpha = \frac{1}{L}$, Proposition 3.3 shows that $\|\mathbb{F}(y_{N-1})\| = \frac{\|y_{N-1} - \mathbb{T}(y_{N-1})\|}{\alpha} \leq \epsilon$ is achieved in $\mathcal{O}\left(\frac{LD}{\epsilon}\right)$ iterations, where $D = \|y_0 - y_\star\|$.

Proposition 3.4 (Cai et al. [9], Informal)**.** Stochastic OHM (S-OHM), which has the update rule $x_{k+1} = \frac{1}{k+2}x_0 + \frac{k+1}{k+2}\mathbb{T}_{\mathcal{B}_k}(x_k)$, maintains the deterministic iteration bound $N = \mathcal{O}(L\|x_0 - x_\star\|/\epsilon)$ if $\mathbb{E}\|\mathbb{F}_{\mathcal{B}_k}(x_k) - \mathbb{F}(x_k)\|^2 \lesssim \frac{\epsilon^2}{k}$. Simple mini-batching with $|\mathcal{B}_k| = \Theta(k\epsilon^{-2})$ yields $\mathcal{O}(\epsilon^{-4})$ stochastic oracle complexity, and using PAGE variance reduction, this can be reduced to $\mathcal{O}(\epsilon^{-3})$.

4 Stochastic Dual-Anchor Acceleration

In this section, we analyze *Stochastic Dual-OHM (S-Dual-OHM)*: with $\mathbb{T}_{\mathcal{B}_{-1}}(x_{-1}) = x_0$,

$$x_{k+1} = x_k + \frac{N-k-1}{N-k}\left(\mathbb{T}_{\mathcal{B}_k}(x_k) - \mathbb{T}_{\mathcal{B}_{k-1}}(x_{k-1})\right) \qquad \text{(S-Dual-OHM)}$$

for $k = 0, \ldots, N-2$, and present the following main result of this paper.

Theorem 4.1. Consider the root-finding problem (1) with a solution $x_\star$, satisfying Assumptions 3.1 and 3.2. Then for $0 < \alpha \leq \frac{2}{L}$, S-Dual-OHM with $\mathbb{T}_{\mathcal{B}_k} = \mathbb{I} - \alpha\mathbb{F}_{\mathcal{B}_k}$ with $|\mathcal{B}_k| \equiv B$ satisfies

$$\mathbb{E}\left[\|\mathbb{F}(x_{N-1})\|\right]^2 \leq \mathbb{E}\left[\|\mathbb{F}(x_{N-1})\|^2\right] \leq \frac{4\|x_0 - x_\star\|^2}{\alpha^2 N^2} + \frac{6\sigma^2}{B}.$$

This shows that we will have $\mathbb{E}\left[\|\mathbb{F}(x_{N-1})\|\right] \leq \epsilon$ with $N = \Theta(1/\epsilon)$ and $B = \Theta(1/\epsilon^2)$, so in total, $\mathcal{O}(1/\epsilon^3)$ oracle complexity. This also proves an equivalent bound on the expected fixed-point residual $\mathbb{E}\left[\|x_{N-1} - \mathbb{T}(x_{N-1})\|\right]^2 \leq \frac{4\|x_0 - x_\star\|^2}{N^2} + \frac{6\alpha^2\sigma^2}{B}$. While we defer technical details to Appendix A, we provide the outline of key arguments below. Then, we show that with the additional assumption of strong monotonicity of $\mathbb{F}$, the total complexity can be reduced to $\widetilde{\mathcal{O}}(1/\epsilon^2)$ with early stopping.

4.1 Proof outline for Theorem 4.1

We start with showing the identity characterizing the S-Dual-OHM algorithm.

Lemma 4.2. For S-Dual-OHM with $\mathbb{T}_{\mathcal{B}_k} = \mathbb{I} - \alpha\mathbb{F}_{\mathcal{B}_k}$, the following identity holds:

$$0 = \frac{(N-1)\alpha^2}{4}\|\mathbb{F}(x_{N-1})\|^2 + \frac{\alpha}{2}\langle\mathbb{F}(x_{N-1}), x_{N-1} - x_0\rangle + \sum_{j=1}^{N-1}\lambda_{N,j}\mathcal{Q}_{N,j}$$

where for $j = 1, \ldots, N-1$, $\lambda_{N,j} = \frac{N\alpha}{2(N-j)(N-j+1)}$ and

$$\mathcal{Q}_{N,j} = \left\langle x_{j-1} - x_{N-1}, \mathbb{F}_{\mathcal{B}_{j-1}}(x_{j-1}) - \mathbb{F}(x_{N-1})\right\rangle - \frac{\alpha}{2}\left\|\mathbb{F}_{\mathcal{B}_{j-1}}(x_{j-1}) - \mathbb{F}(x_{N-1})\right\|^2.$$

Here, when the problem is deterministic so that $\mathbb{F}_{\mathcal{B}_k} = \mathbb{F}$ for all $k = 0, \ldots, N-1$ and $0 < \alpha \leq \frac{2}{L}$, then by cocoercivity $\mathcal{Q}_{N,j} \geq 0$ for all $j = 1, \ldots, N-1$, so $\sum_{j=1}^{N-1}\lambda_{N,j}\mathcal{Q}_{N,j} \geq 0$. The

resulting inequality, together with Young's inequality and one more application of cocoercivity: $\langle \mathbb{F}(x_{N-1}), x_{N-1} - x_\star \rangle \geq \frac{1}{L} \|\mathbb{F}(x_{N-1})\|^2 \geq \frac{\alpha}{2} \|\mathbb{F}(x_{N-1})\|^2$, yields

$$0 \geq \frac{(N-1)\alpha^2}{4} \|\mathbb{F}(x_{N-1})\|^2 + \frac{\alpha}{2} \langle \mathbb{F}(x_{N-1}), x_{N-1} - x_0 \rangle \tag{4}$$

$$\geq \frac{N\alpha^2}{4} \|\mathbb{F}(x_{N-1})\|^2 - \frac{N\alpha^2}{8} \|\mathbb{F}(x_{N-1})\|^2 - \frac{1}{2N} \|x_\star - x_0\|^2 \tag{5}$$

which implies $\|\mathbb{F}(x_{N-1})\|^2 \leq \frac{4\|x_0 - x_\star\|^2}{\alpha^2 N^2}$.

On the other hand, in the stochastic setting, we do not have $\mathcal{Q}_{N,j} \geq 0$ due to noise, and these quantities are not local as in typical stochastic analyses, making the convergence proof seemingly much harder. However, we can isolate the dependence on stochastic errors at each iteration and control them through the following two lemmas. These key results allow us to follow the reasoning similar to the deterministic case above.

Lemma 4.3. Under the setting of Theorem 4.1, let $e_k = \mathbb{F}_{\mathcal{B}_k}(x_k) - \mathbb{F}(x_k)$ for $k = 0, \ldots, N-2$. Then

$$\begin{aligned} 0 \geq \mathbb{E}\left[\frac{(N-1)\alpha^2}{4} \|\mathbb{F}(x_{N-1})\|^2 + \frac{\alpha}{2} \langle \mathbb{F}(x_{N-1}), x_{N-1} - x_0 \rangle \right] \\ - \sum_{j=1}^{N-1} \lambda_{N,j} \left(\frac{\alpha\sigma^2}{2B} + \mathbb{E}\left[\langle e_{j-1}, \mathbb{T} x_{N-1} \rangle \right] \right). \end{aligned} \tag{6}$$

Lemma 4.4. Under the setting of Theorem 4.1, $\mathbb{E}\left[\langle e_{j-1}, \mathbb{T} x_{N-1} \rangle\right] \leq \frac{\alpha\sigma^2}{B}$ for $j = 1, \ldots, N-1$.

Lemma 4.4 states that the amount by which the error e_{j-1} propagates to the last iterate and affects the analysis is uniformly bounded by a constant independent of N. Combining Lemmas 4.3 and 4.4, and using $\sum_{j=1}^{N-1} \lambda_{N,j} = \frac{\alpha(N-1)}{2}$, we can proceed as in (4) to obtain

$$\begin{aligned} 0 &\geq \mathbb{E}\left[\frac{N\alpha^2}{4} \|\mathbb{F}(x_{N-1})\|^2 + \frac{\alpha}{2} \langle \mathbb{F}(x_{N-1}), x_\star - x_0 \rangle \right] - \frac{3(N-1)\alpha^2\sigma^2}{4B} \\ &\geq \mathbb{E}\left[\frac{N\alpha^2}{4} \|\mathbb{F}(x_{N-1})\|^2 - \frac{N\alpha^2}{8} \|\mathbb{F}(x_{N-1})\|^2 - \frac{1}{2N} \|x_0 - x_\star\|^2 \right] - \frac{3(N-1)\alpha^2\sigma^2}{4B}. \end{aligned}$$

Rearranging, we have the desired bound of Theorem 4.1.

4.1.1 Intuition for why S-Dual-OHM is stable under noise

The previous analysis shows that the net accumulation of error in S-Dual-OHM is proportional to $\sum_{j=1}^{N-1} \lambda_{N,j} = \frac{\alpha(N-1)}{2} = \mathcal{O}(N)$. On the other hand, stochastic OHM satisfies the similar identity

$$0 = \frac{(N-1)\alpha^2}{4} \|\mathbb{F}(x_{N-1})\|^2 + \frac{\alpha}{2} \langle \mathbb{F}(x_{N-1}), x_{N-1} - x_0 \rangle + \sum_{j=1}^{N-1} \nu_{j+1,j} \mathcal{Q}_{j+1,j}$$

where $\mathcal{Q}_{j+1,j} = \langle x_{j-1} - x_j, \mathbb{F}_{\mathcal{B}_{j-1}}(x_{j-1}) - \mathbb{F}_{\mathcal{B}_j}(x_j) \rangle - \frac{\alpha}{2} \left\| \mathbb{F}_{\mathcal{B}_{j-1}}(x_{j-1}) - \mathbb{F}_{\mathcal{B}_j}(x_j) \right\|^2$ but with $\nu_{j+1,j} = \frac{\alpha j(j+1)}{2N}$. Here each $\mathcal{Q}_{j+1,j}$ is not nonnegative, but contributes a negative lower bound proportional to σ^2 in expectation, which then gets multiplied with $\sum_{j=1}^{N-1} \nu_{j+1,j} = \Theta(N^2)$ and the resulting error term is no longer controllable as in Theorem 4.1. This is a central difference that makes the dual-anchor mechanism more amenable to stochastic extension than anchoring. A concurrent work [53] has also provided a related theory that the sum of weights associated with inequalities used in the convergence proof measures a fixed-point algorithm's robustness to oracle noise, albeit in deterministic setting.

A complementary interpretation is that OHM is an anytime optimal algorithm, achieving the rate $\|\mathbb{F}(x_{k-1})\|^2 \leq \frac{4\|x_0 - x_\star\|^2}{\alpha^2 k^2}$ for all $k = 1, 2, \ldots$. This means that OHM must maintain the optimal guarantee at every iteration, forcing it to exploit the geometric property of $\mathbb{F}$, i.e. cocoercivity,

throughout the trajectory. Its analysis is therefore more tightly tuned to the exact problem class and is more easily disrupted by oracle noise. Dual-OHM, by contrast, is optimized only for the prescribed last iterate and does not suffer from the same accumulation of errors.

4.2 Faster convergence under strong monotonicity

In this section, we show that if $\mathbb{F}$ is additionally strongly monotone, then S-Dual-OHM can achieve the desired accuracy $\mathbb{E}\left[\|\mathbb{F}(x_k)\|\right] \le \epsilon$ much earlier than the prescribed horizon N. Suppose that $\mathbb{F}$ is μ-strongly monotone and $1/L$-cocoercive. Then, with $0 < \alpha < \frac{2}{L}$, $\mathbb{T} = \mathbb{I} - \alpha\mathbb{F}$ is γ-*contractive* (i.e., γ-Lipschitz) with $\gamma = \sqrt{1 - \alpha\mu(2 - \alpha L)}$ because

$$\begin{aligned}\|\mathbb{T}(x) - \mathbb{T}(y)\|^2 &= \|x - y\|^2 - 2\alpha \langle x - y, \mathbb{F}(x) - \mathbb{F}(y)\rangle + \alpha^2 \|\mathbb{F}(x) - \mathbb{F}(y)\|^2 \\ &\le \|x - y\|^2 - \alpha(2 - \alpha L) \langle x - y, \mathbb{F}(x) - \mathbb{F}(y)\rangle \le (1 - \alpha\mu(2 - \alpha L)) \|x - y\|^2.\end{aligned}$$

Theorem 4.5. Consider the root-finding problem (1) with $1/L$-cocoercive and μ-strongly monotone $\mathbb{F}$ and a solution $x_\star$, satisfying Assumption 3.1. Let $0 < \alpha < \frac{2}{L}$ and $\gamma = \sqrt{1 - \alpha\mu(2 - \alpha L)} \in [0, 1)$. Let y_k be the iterates from deterministic Dual-OHM using true evaluations of $\mathbb{T}$, with $x_0 = y_0$. Then, for $k = 0, \dots, N-1$, S-Dual-OHM with $\mathbb{T}_{\mathcal{B}_k} = \mathbb{I} - \alpha\mathbb{F}_{\mathcal{B}_k}$ satisfies

$$\mathbb{E}\left[\|x_k - y_{N-1}\|^2\right] \le 2\left(\frac{1+\gamma}{1-\gamma}\right)^2 \gamma^{2k} \|x_0 - x_\star\|^2 + \frac{2\alpha^2\sigma^2}{B(1-\gamma^2)}. \tag{7}$$

In particular, when $\alpha = \frac{1}{L}$, we have

$$\mathbb{E}\left[\|\mathbb{F}(x_k)\|^2\right] \le \frac{8L^2 \|x_0 - x_\star\|^2}{N^2} + \frac{64L^4}{\mu^2} e^{-k\frac{\mu}{L}} \|x_0 - x_\star\|^2 + \frac{L}{\mu}\frac{4\sigma^2}{B}. \tag{8}$$

Notably, Theorem 4.5 uses only Assumption 3.1 and cocoercivity of $\mathbb{F}$, but not Assumption 3.2.

The reason why this result holds is two-fold. First, when $\mathbb{T}$ is contractive, S-Dual-OHM does not deviate far from deterministic Dual-OHM (Lemma 4.6). Second, deterministic Dual-OHM achieves the designated terminal accuracy quickly as it linearly converges to its final iterate y_{N-1} (Lemma 4.7). Because Dual-OHM has a small residual norm guarantee at y_{N-1}, this implies that x_k also attains the comparably small residual quickly.

Lemma 4.6. Under the conditions of Theorem 4.5, $\mathbb{E}\left[\|x_k - y_k\|^2\right] \le \frac{\alpha^2\sigma^2}{B(1-\gamma^2)}$ for $k = 0, \dots, N-1$.

Lemma 4.7. Under the conditions of Lemma 4.6, we have $\|y_k - y_{N-1}\| \le \left(\frac{1+\gamma}{1-\gamma}\right)\gamma^k \|x_0 - x_\star\|$ for $k = 0, \dots, N-1$ and $x_\star \in \operatorname{Fix}\mathbb{T}$.

Proof of Theorem 4.5. Note that $\mathbb{E}\left[\|x_k - y_{N-1}\|^2\right] \le 2\mathbb{E}\left[\|x_k - y_k\|^2\right] + 2\mathbb{E}\left[\|y_k - y_{N-1}\|^2\right]$. Then, by Lemmas 4.6 and 4.7, $\mathbb{E}\left[\|x_k - y_{N-1}\|^2\right] \le \frac{2\alpha^2\sigma^2}{B(1-\gamma^2)} + 2\left(\frac{1+\gamma}{1-\gamma}\right)^2 \gamma^{2k} \|x_0 - x_\star\|^2$, which proves (7). With $\alpha = \frac{1}{L}$, we have $1 - \gamma^2 = \frac{\mu}{L}$, so combining (7) with Proposition 3.3 we obtain

$$\begin{aligned}\mathbb{E}\left[\|\mathbb{F}(x_k)\|^2\right] &\le 2\mathbb{E}\left[\|\mathbb{F}(y_{N-1})\|^2\right] + 2\mathbb{E}\left[\|\mathbb{F}(x_k) - \mathbb{F}(y_{N-1})\|^2\right] \\ &\le 2\mathbb{E}\left[\|\mathbb{F}(y_{N-1})\|^2\right] + 2L^2\mathbb{E}\left[\|x_k - y_{N-1}\|^2\right] \\ &\le \frac{8L^2 \|x_0 - x_\star\|^2}{N^2} + 2L^2\left[\frac{2\sigma^2}{\mu L B} + 2\left(\frac{1+\gamma}{1-\gamma}\right)^2 \gamma^{2k} \|x_0 - x_\star\|^2\right].\end{aligned}$$

where the second inequality uses L-Lipschitzness of $\mathbb{F}$. Rearranging and using $\gamma < 1$, $(1-\gamma)^2 \ge \frac{(1-\gamma)^2(1+\gamma)^2}{4} = \frac{\mu^2}{4L^2}$ and $\gamma^{2k} = \left(1 - \frac{\mu}{L}\right)^k \le e^{-k\frac{\mu}{L}}$, we obtain (8). □

Let $\|x_0 - x_\star\| = D$. The bound (8) shows that if one chooses, e.g., $N \geq \frac{4LD}{\epsilon}$, $k \geq \frac{L}{\mu} \log \frac{256L^4D^2}{\mu^2\epsilon^2}$ and $B \geq \frac{16L\sigma^2}{\mu\epsilon^2}$, then $\mathbb{E}\left[\|\mathbb{F}(x_k)\|\right]^2 \leq \mathbb{E}\left[\|\mathbb{F}(x_k)\|^2\right] \leq \epsilon^2$. If we run S-Dual-OHM with these choices of N and B, we will achieve the accuracy ϵ in $\mathcal{O}\left(\frac{L}{\mu}\log\epsilon^{-1}\right)$ iterations without the need to run all N iterations. With this early stopping, S-Dual-OHM has improved $\mathcal{O}\left(\frac{L^2\sigma^2}{\mu^2\epsilon^2}\log\frac{1}{\epsilon}\right)$ oracle complexity for strongly monotone $\mathbb{F}$, which has near-optimal dependence on ϵ [12]. The dependence on condition number, however, is still suboptimal, and improving it based on the Dual-OHM mechanism remains an open question.

5 Numerical Experiments

We provide numerical evaluations to demonstrate that the dual-anchor mechanism is indeed preferable to anchoring (Halpern algorithms) under stochasticity, and compare the empirical performance of S-Dual-OHM with prior algorithms for stochastic monotone inclusions/fixed-point problems. We consider three distinct problems: the worst-case nonexpansive fixed-point operator designed by [40], a finite-sum cocoercive operator using uniform random vectors from the unit sphere, and a strongly-convex–strongly-concave minimax problem with Huber-type regularizers following [54, 12]. All experiments were run on a MacBook Air with Apple M3 Chip and 24GB Memory.

Baseline algorithms. We consider the simple algorithm $x_{k+1} = x_k - \alpha\mathbb{F}_{\mathcal{B}_k}(x_k)$, which we call Stochastic Gradient Descent-Ascent (SGDA) following the nomenclature of the minimax optimization literature. We consider the Stochastic Extragradient (SEG) with the update rule

$$x_{k+1/2} = x_k - \alpha\mathbb{F}_{\mathcal{B}_k}(x_k), \quad x_{k+1} = x_k - \alpha\mathbb{F}_{\mathcal{B}_{k+1/2}}(x_{k+1/2}).$$

Next, we consider the naive stochastic extension of OHM: $x_{k+1} = \frac{1}{k+2}x_0 + \frac{k+1}{k+2}\left(x_k - \alpha\mathbb{F}_{\mathcal{B}_k}(x_k)\right)$, using the same constant mini-batch size as S-Dual-OHM. We refer to this algorithm as S-OHM (with constant B). This is included to isolate the effect of using the dual-anchor update rather than the usual anchor acceleration and directly contrast the two mechanisms.

We also include the algorithms carefully designed for reducing the expected last-iterate residual. First, the HALPERN-PAGE algorithm from [9] uses the S-OHM update but replaces $\mathbb{F}_{\mathcal{B}_k}(x_k)$ by a PAGE-type estimator which uses a larger batch size s_1 with probability $p_k = \min\left\{1, \frac{2}{k+2}\right\}$ and otherwise uses a sample operator difference with batch size s_2. Second, we use the single-loop simplification of RAIN algorithm from [12] used in their experiments, which can be written as

$$z_{k+1/2} = z_k - \eta\big(\widehat{\mathbb{F}}(z_k;\xi_k) + r_k(z_k)\big), \qquad z_{k+1} = z_k - \eta\big(\widehat{\mathbb{F}}(z_{k+1/2};\xi_{k+1/2}) + r_k(z_{k+1/2})\big)$$

where $r_k(x) = \lambda\gamma(S_kx - m_k), S_k = \sum_{t<k}(1+\gamma)^t, m_k = \sum_{t<k}(1+\gamma)^t z_t$, and $\eta, \lambda, \gamma > 0$ are hyperparameters. Finally, HALPERN-VR-FINITE from [10] is similar to HALPERN-PAGE but applicable only to finite sum problems of the form $\mathbb{F} = \frac{1}{n}\sum_{i=1}^n \mathbb{F}(\cdot;\xi_i)$ and uses either full-batch (true) operator $\mathbb{F}$ or batches of size $\lceil\sqrt{n}\rceil$.

We tune all baseline algorithms and S-Dual-OHM under the same total sample budget $Q = NB$ via grid search, where computing $\mathbb{F}_{\mathcal{B}_k}$ with $|\mathcal{B}_k| = B$ counts as B operator evaluations. The only hyperparameter B for S-Dual-OHM is tuned over $B \in \{1, 10, 20, 50, 100\}$, and S-OHM (Constant B) is then run with the same selected B.[1] The HALPERN-PAGE parameters are tuned over $s_1 \in \{1, 10, 20, 50, 100\}$ and $s_2 \in \{1, 5, 10, 20\}$, and RAIN hyperparameters are tuned over $\eta \in \{0.005, 0.01, 0.05, 0.1, 1\}$, $\lambda \in \{0.001, 0.01, 0.1, 1\}$, and $\gamma \in \{0.001, 0.01, 0.1, 1\}$. SGDA and SEG step-sizes were tuned over $\alpha \in \{0.005, 0.01, 0.05, 0.1, 1\}$, the same step-size grid as RAIN. We run each experiment with 10 independent random seeds and report the average residual norm $\|\mathbb{F}(x_k)\|$ together with the shaded region indicating the empirical 5th–95th percentile band.

Experiment 1: Worst-case nonexpansive operator. We use the affine lower-bound construction of [40]: define $\mathbb{H}\colon \mathbb{R}^d \to \mathbb{R}^d$ by $\mathbb{H}(x_1, \ldots, x_d) = \left(x_d - \frac{2}{\sqrt{d}}, -x_1, -x_2, \ldots, -x_{d-1}\right)$, and after

[1] While this emphasizes the distinction between the anchor and the dual-anchor updates, for fully controlled comparison, we additionally provide experiments where B for S-OHM is tuned separately in Appendix B.

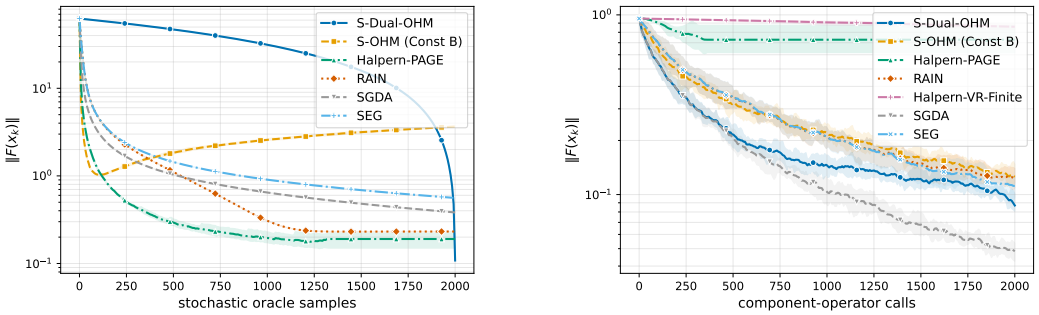

Figure 1: Plot of residual norm versus stochastic samples. **(Left)** Worst-case nonexpansive operator construction (Experiment 1). **(Right)** Finite-sum random cocoercive operator construction (Experiment 2). Solid curves indicate means over 10 independent runs, and shaded bands denote empirical 5th–95th percentiles.

choosing a shift vector $s \in \mathbb{R}^d$, let $\mathbb{F}(x) = \mathbb{H}(x-s)+(x-s)$ and $\mathbb{T} = \mathbb{I}-\mathbb{F}$. Then $\mathbb{F}$ is $\frac{1}{2}$-cocoercive and $\mathbb{T}$ is nonexpansive. We use Gaussian noise oracle $\widehat{\mathbb{F}}(x;\xi) = \mathbb{F}(x) + \xi$ with $\xi \sim \mathcal{N}\left(0, \frac{\sigma^2}{d} I_d\right)$, and use a budget of 2000 samples. We set $d = 2001$, $x_0 = 0$, $\alpha = 1$, $\sigma = 0.1$ and $s \sim \mathcal{N}(0, I_d)$.

The left panel of Figure 1 illustrates that S-Dual-OHM is not an anytime algorithm, so its intermediate residuals can be worse than those of the other algorithms; however, in the end, it achieves the smallest last-iterate residual among all algorithms compared. This is in sharp contrast with S-OHM with constant B, which fails to converge and exhibits diverging residual. This demonstrates that Halpern-type algorithms can indeed be prone to accumulation of stochastic errors. However, with variance reduction (HALPERN-PAGE) or recursively adjusted anchor (RAIN), the residual convergence gets stabilized. Finally, SEG and SGDA converge steadily although they were not specifically designed for residual minimization.

Experiment 2: Finite-sum cocoercive operator. We consider a finite-sum problem

$$\mathbb{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{F}_i(x), \qquad \mathbb{F}_i(x) = \begin{pmatrix} q_i(q_i^\top x_{1:r}) \\ 0 \end{pmatrix} \in \mathbb{R}^d,$$

where each $q_i \in \mathbb{R}^r$ is sampled uniformly from the unit sphere. Here each $\mathbb{F}_i$ is 1-cocoercive and linear, and therefore $\mathbb{F}$ is cocoercive with $L \le 1$ but not strongly monotone when $r < d$ since the common null space of $\mathbb{F}_i$ is nontrivial. A stochastic oracle call samples component indices and returns the average of the queried $\mathbb{F}_i$'s. The operator variance is bounded over the region that trajectories of all algorithms stay within during all experiment runs. We use a budget of 2000 stochastic samples, and set $n = 200$, $d = 200$, $r = 199$ and $\mathbb{T} = \mathbb{I} - \alpha\mathbb{F}$ with $\alpha = 1$. We take $x_0 \sim 10 \cdot \mathcal{N}(0, I_d)$. This is a generic, non-worst-case problem that basic methods like SGDA can solve effectively. Still, S-Dual-OHM is also fairly competent, and importantly, outperforms its primal counterpart S-OHM. Notably, HALPERN-VR-FINITE makes very small progress because the total sample budget 2000 is restrictive with our choice $n = 200$, even though it has the best asymptotic dependence on ϵ in theory.

Experiment 3: SCSC minimax problem with Huber-type regularizers. We consider

$$\Phi(x,y) = (1-\delta)\sum_i g_\nu(x_i) + \delta x^\top y - (1-\delta)\sum_i g_\nu(y_i) + \frac{\mu}{2}\|x\|^2 - \frac{\mu}{2}\|y\|^2,$$

for $z = (x,y) \in \mathbb{R}^d \times \mathbb{R}^d$, where $g_\nu(u) = \begin{cases} \frac{1}{2}u^2, & |u| \le \nu \\ \nu|u| - \frac{1}{2}\nu^2, & |u| > \nu \end{cases}$. The saddle operator $\mathbb{F}(z) = (\nabla_x\Phi(x,y), -\nabla_y\Phi(x,y))$ is then given as $\mathbb{F}(z) = \begin{pmatrix} (1-\delta)\,\mathrm{clip}(x,-\nu,\nu) + \delta y + \mu x \\ -\delta x + (1-\delta)\,\mathrm{clip}(y,-\nu,\nu) + \mu y \end{pmatrix}$, which is μ-strongly monotone and M-Lipschitz with $M = 1+\mu$. We use the Gaussian noise oracle $\widehat{\mathbb{F}}(z;\xi) = \mathbb{F}(z) + \xi$ with $\xi \sim \mathcal{N}\left(0, \frac{\sigma^2}{2d} I_{2d}\right)$, and use a budget of 10000 stochastic samples. We set

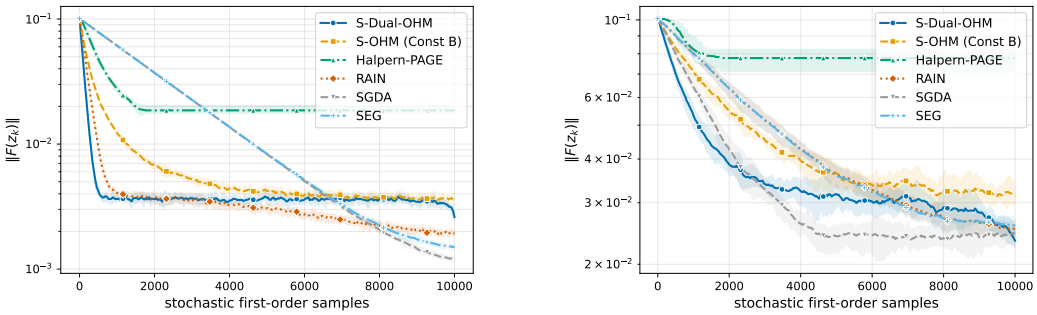

Figure 2: Plot of residual norm versus stochastic samples for SCSC Huber-type minimax problem (Experiment 3). **(Left)** Low-variance regime $\sigma = 0.05$. **(Right)** High-variance regime $\sigma = 1.5$. Solid curves are means over 10 independent runs, and shaded bands indicate empirical 5th–95th percentiles.

$d = 50$, $\delta = 10^{-2}$, $\nu = 5 \times 10^{-5}$, $\mu = 0.1$, $L = \frac{M^2}{\mu}$ and $z_0 = (x_0, y_0)$ is a uniform random unit vector. We use $\alpha = \frac{1}{L}$ for S-Dual-OHM, S-OHM and HALPERN-PAGE.

We simulate both low noise ($\sigma = 0.05$) and high noise ($\sigma = 1.5$) regimes. When $\sigma = 0.05$, we observe rapid progression in earlier iterations of S-Dual-OHM, but then it stagnates (Figure 2). This is because the noise term is small, so the best batch size choice is $B^* = 1$, i.e., mini-batching does not benefit the final convergence. In this regime, RAIN, SGDA, SEG converges more steadily than S-Dual-OHM, although we still observe that S-OHM is inferior to S-Dual-OHM. With larger noise with $\sigma = 1.5$, we observe that mini-batching becomes effective and $B^* = 10$ is selected; in this case, dual-anchor with mini-batching exhibits comparable performance to classical methods and deliberately designed near-optimal algorithm RAIN.

6 Conclusion, Limitations and Open Questions

We demonstrate that dual-anchor acceleration from deterministic root-finding or fixed-point problems admits a direct stochastic extension, unlike anchor-type accelerated algorithms. This is based on the observation that distinct deterministic accelerations are intrinsically different in their behavior under stochasticity. We believe that our work enables interesting future work on acceleration and optimal complexities for stochastic optimization.

Two immediate limitations of our work, which thus remain as open questions, are: **(1)** whether the dual-anchor mechanism extends to root-finding problems with monotone and Lipschitz operators $\mathbb{F}$ and **(2)** whether we can further improve the complexity to the optimal level [12] $\mathcal{O}(\epsilon^{-2} \log \epsilon^{-1})$ via dual-anchor acceleration. As the dual-anchor acceleration also exists in the monotone Lipschitz inclusion setting [56], we believe that the extension would be possible with novel insights not relying on nonexpansivity of $\mathbb{I} - \alpha\mathbb{F}$. We also expect our result to serve as a baseline that can be combined with broader techniques in stochastic optimization for potentially achieving the optimal complexity.

Acknowledgments

TaeHo Yoon's contribution to this work was supported by NSF CCF 2504626. Nicolas Loizou's contribution to this work was supported by NSF CCF 2504626, AFOSR YIP FA9550-26-1-B353, and NSF CAREER 2542902.

References

[1] James K Alcala, Yat Tin Chow, and Mahesh Sunkula. Moving anchor extragradient methods for smooth structured minimax problems. *arXiv:2308.12359*, 2023.

[2] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(221):1–51, 2018.

[3] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer International Publishing, 2nd edition, 2017.

[4] Aleksandr Beznosikov, Eduard Gorbunov, Hugo Berard, and Nicolas Loizou. Stochastic gradient descent-ascent: Unified theory and new efficient methods. *International Conference on Artificial Intelligence and Statistics*, 2023.

[5] Radu I. Boţ and Enis Chenchene. Extragradient method with flexible anchoring: Strong convergence and fast residual decay. *SIAM Journal on Optimization*, 36(3):1420–1445, 2026.

[6] Radu Ioan Boţ, Ernö Robert Csetnek, and Dang-Khoa Nguyen. Fast Optimistic Gradient Descent Ascent (OGDA) method in continuous and discrete time. *Foundations of Computational Mathematics*, 2023.

[7] Mario Bravo and Roberto Cominetti. Stochastic fixed-point iterations for nonexpansive maps: Convergence and error bounds. *SIAM Journal on Control and Optimization*, 62(1):191–219, 2024.

[8] Mario Bravo and Juan Pablo Contreras. Stochastic Halpern iteration in normed spaces and applications to reinforcement learning. *Mathematical Programming*, 2026.

[9] Xufeng Cai, Chaobing Song, Cristóbal A Guzmán, and Jelena Diakonikolas. Stochastic halpern iteration with variance reduction for stochastic monotone inclusions. *Neural Information Processing Systems*, 2022.

[10] Xufeng Cai, Ahmet Alacaoglu, and Jelena Diakonikolas. Variance reduced halpern iteration for finite-sum monotone inclusions. *International Conference on Learning Representations*, 2024.

[11] Yang Cai and Weiqiang Zheng. Accelerated single-call methods for constrained min-max optimization. *International Conference on Learning Representations*, 2023.

[12] Lesi Chen and Luo Luo. Near-optimal algorithms for making the gradient small in stochastic minimax optimization. *Journal of Machine Learning Research*, 25(387):1–44, 2024.

[13] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. *International Conference on Learning Representations*, 2018.

[14] Damek Davis. Variance reduction for root-finding problems. *Mathematical Programming*, 197 (1):375–410, 2023.

[15] Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75, 2014.

[16] Jelena Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. *Conference on Learning Theory*, 2020.

[17] Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer-Verlag, 2003.

[18] Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Neural Information Processing Systems*, 2014.

[20] Eduard Gorbunov, Hugo Berard, Gauthier Gidel, and Nicolas Loizou. Stochastic extragradient: General analysis and improved rates. *International Conference on Artificial Intelligence and Statistics*, 2022.

[21] Benjamin Halpern. Fixed points of nonexpanding maps. *Bulletin of the American Mathematical Society*, 73(6):957–961, 1967.

[22] Donghwan Kim. Accelerated proximal point method for maximally monotone operators. *Mathematical Programming*, 190(1–2):57–87, 2021.

[23] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12(4):747–756, 1976.

[24] M. A. Krasnosel'skii. Two remarks on the method of successive approximations. *Uspekhi Matematicheskikh Nauk*, 10(1):123–127, 1955.

[25] Jongmin Lee and Ernest K. Ryu. Accelerating value iteration with anchoring. *Neural Information Processing Systems*, 2023.

[26] Jongmin Lee, Mario Bravo, and Roberto Cominetti. Near-optimal sample complexity for MDPs via anchoring. *Proceedings of the 42nd international conference on machine learning*, 2025.

[27] Sucheol Lee and Donghwan Kim. Fast extra gradient methods for smooth structured nonconvex–nonconcave minimax problems. *Neural Information Processing Systems*, 2021.

[28] Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtarik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. *International Conference on Machine Learning*, 2021.

[29] Felix Lieder. On the convergence rate of the Halpern-iteration. *Optimization Letters*, 15(2): 405–418, 2021.

[30] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. *Neural Information Processing Systems*, 2015.

[31] Tianyi Lin, Chi Jin, and Michael I. Jordan. Near-optimal algorithms for minimax optimization. *Conference on Learning Theory*, 125, 2020.

[32] Nicolas Loizou, Hugo Berard, Gauthier Gidel, Ioannis Mitliagkas, and Simon Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Neural Information Processing Systems*, 2021.

[33] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018.

[34] William Robert Mann. Mean value methods in iteration. *Proceedings of the American Mathematical Society*, 4(3):506–510, 1953.

[35] Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, C.-S. Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *International Conference on Learning Representations*, 2019.

[36] George J. Minty. Monotone (nonlinear) operators in Hilbert space. *Duke Mathematical Journal*, 29(3):341–346, 1962.

[37] Martin Morin and Pontus Giselsson. Cocoercivity, smoothness and bias in variance-reduced stochastic gradient methods. *Numerical Algorithms*, 91(2):749–772, 2022.

[38] Arkadi 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.

[39] Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. *International Conference on Machine Learning*, 2017.

[40] Jisun Park and Ernest K Ryu. Exact optimal accelerated complexity for fixed-point iterations. *International Conference on Machine Learning*, 2022.

[41] Nicholas Pischke and Thomas Powell. Asymptotic Regularity of a Generalised Stochastic Halpern Scheme. *Journal of Optimization Theory and Applications*, 210(1):3, 2026.

[42] R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[43] Ernest K Ryu and Wotao Yin. *Large-Scale Convex Optimization via Monotone Operators*. Cambridge University Press, 2022.

[44] Ernest K. Ryu, Kun Yuan, and Wotao Yin. ODE analysis of stochastic gradient methods with optimism and anchoring for minimax problems and GANs. *arXiv:1905.10899*, 2019.

[45] Shoham Sabach and Shimrit Shtern. A first order method for solving convex bilevel optimization problems. *SIAM Journal on Optimization*, 27(2):640–660, 2017.

[46] Gesualdo Scutari, Francisco Facchinei, Jong-Shi Pang, and Daniel P. Palomar. Real and complex monotone communication games. *IEEE Transactions on Information Theory*, 60(7):4197–4231, 2014.

[47] Guido Stampacchia. Formes bilineaires coercitives sur les ensembles convexes. *Comptes Rendus Hebdomadaires Des Seances De L Academie Des Sciences*, 258(18):4413, 1964.

[48] Jaewook J. Suh, Jisun Park, and Ernest K. Ryu. Continuous-time analysis of anchor acceleration. *Neural Information Processing Systems*, 2023.

[49] Quoc Tran-Dinh. From Halpern's fixed-point iterations to Nesterov's accelerated interpretations for root-finding problems. *Computational Optimization and Applications*, 87(1):181–218, 2024.

[50] Quoc Tran-Dinh. VFOSA: Variance-reduced fast operator splitting algorithms for generalized equations. *Journal of Machine Learning Research*, 26(254):1–68, 2025.

[51] Quoc Tran-Dinh and Yang Luo. Halpern-type accelerated and splitting algorithms for monotone inclusions. *arXiv:2110.08150*, 2021.

[52] Rainer Wittmann. Approximation of fixed points of nonexpansive mappings. *Archiv der Mathematik*, 58(5):486–491, 1992.

[53] TaeHo Yoon and Benjamin Grimmer. A Theory of Composition and Duality of Extremal Optimal Fixed-Point Algorithms. *arXiv:2605.02231*, 2026.

[54] TaeHo Yoon and Ernest K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. *International Conference on Machine Learning*, 2021.

[55] TaeHo Yoon and Ernest K. Ryu. Accelerated minimax algorithms flock together. *SIAM Journal on Optimization*, 35(1):180–209, 2025.

[56] TaeHo Yoon, Jaeyeon Kim, Jaewook J. Suh, and Ernest K. Ryu. Optimal acceleration for minimax and fixed-point problems is not unique. *International Conference on Machine Learning*, 2024.

[57] TaeHo Yoon, Sayantan Choudhury, and Nicolas Loizou. Multiplayer federated learning: Reaching equilibrium with less communication. *Advances in Neural Information Processing Systems*, 2026.

[58] TaeHo Yoon, Ernest K. Ryu, and Benjamin Grimmer. H-invariance theory: A complete characterization of minimax optimal fixed-point algorithms. *Mathematical Programming*, 2026.

[59] Siqi Zhang, Sayantan Choudhury, Sebastian Stich, and Nicolas Loizou. Communication-efficient gradient descent-accent methods for distributed variational inequalities: Unified analysis and local updates. *International Conference on Learning Representations*, 2024.

Appendix

Contents

1 Introduction **1**

2 Related Work **2**

3 Preliminaries and Assumptions **3**

3.1 Monotonicity, cocoercivity, and nonexpansiveness of operators 3

3.2 Stochastic oracle models 4

3.3 Known acceleration in fixed-point problems and its extension 4

4 Stochastic Dual-Anchor Acceleration **5**

4.1 Proof outline for Theorem 4.1 5

4.1.1 Intuition for why S-Dual-OHM is stable under noise 6

4.2 Faster convergence under strong monotonicity 7

5 Numerical Experiments **8**

6 Conclusion, Limitations and Open Questions **10**

A Missing Proofs **15**

A.1 Proof of Lemma 4.2 15

A.2 Proof of Lemma 4.3 16

A.3 Proof of Lemma 4.4 16

A.4 Proof of Lemma 4.6 19

A.5 Proof of Lemma 4.7 19

B Additional Experiments **20**

A Missing Proofs

We denote by $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots$ the natural filtration generated by each iteration of the algorithm (selection of batches $\mathcal{B}_i$), i.e., $\mathcal{F}_k = \sigma(\mathcal{B}_0, \mathcal{B}_1, \dots, \mathcal{B}_{k-1})$.

A.1 Proof of Lemma 4.2

Proof. The proof is purely algebraic, and does not use any property of $\mathbb{F}$ as an operator or stochastic property of the mini-batches $\mathcal{B}_j$. The case $N = 1$ is immediate (we have the vacuous identity $0 = 0$), so assume $N \geq 2$.

For $j = 1, \dots, N$, define

$$G_j = \mathbb{F}_{\mathcal{B}_{j-1}}(x_{j-1}), \qquad g_j = \frac{\alpha}{2} G_j, \qquad u_j = x_{j-1} - g_j.$$

Here, $\mathcal{B}_{N-1}$ is a ghost (auxiliary) batch introduced only for the analysis, and is not used by the algorithm update; in fact, one may ignore this ghost batch (i.e., safely assume that $\mathcal{B}_{N-1}$ is the full batch) and simply let $G_N = \mathbb{F}(x_{N-1})$. With this notation, we have

$$\mathbb{T}_{\mathcal{B}_{j-1}}(x_{j-1}) = x_{j-1} - \alpha G_j = x_{j-1} - 2g_j.$$

Because $\mathbb{T}_{\mathcal{B}_{-1}}(x_{-1}) = x_0$, the $k = 0$ update of S-Dual-OHM is

$$x_1 = x_0 + \frac{N-1}{N}\big(\mathbb{T}_{\mathcal{B}_0}(x_0) - x_0\big).$$

Let

$$a_k = \frac{N-k-1}{N-k}, \qquad z_0 = 0, \qquad z_{k+1} = \mathbb{T}_{\mathcal{B}_k}(x_k) - x_{k+1}$$

for $k = 0, \dots, N-2$. Then, from the update rule,

$$\begin{aligned} z_{k+1} &= \mathbb{T}_{\mathcal{B}_k}(x_k) - x_k - a_k\big(\mathbb{T}_{\mathcal{B}_k}(x_k) - \mathbb{T}_{\mathcal{B}_{k-1}}(x_{k-1})\big) \\ &= (1-a_k)(x_k - 2g_{k+1}) - x_k + a_k \mathbb{T}_{\mathcal{B}_{k-1}}(x_{k-1}) \\ &= a_k\left(\mathbb{T}_{\mathcal{B}_{k-1}}(x_{k-1}) - x_k\right) - 2(1-a_k)g_{k+1} \\ &= \frac{N-k-1}{N-k} z_k - \frac{2}{N-k} g_{k+1}. \end{aligned} \tag{9}$$

Then, by definition of z_{k+1},

$$x_{k+1} = \mathbb{T}_{\mathcal{B}_k}(x_k) - z_{k+1} = x_k - 2g_{k+1} - z_{k+1}. \tag{10}$$

Now define, for $k = 0, \dots, N-1$,

$$V_k = -\frac{N-k-1}{N-k}\left\|z_k + 2g_N\right\|^2 + \frac{2}{N-k}\left\langle z_k + 2g_N, x_k - x_{N-1}\right\rangle.$$

Note that $V_{N-1} = 0$. Next, we analyze $V_k - V_{k+1}$. To simplify the notations, fix $k \in \{0, \dots, N-2\}$, and write

$$m = N-k, \qquad w_k = z_k + 2g_N, \qquad d_k = g_{k+1} - g_N, \qquad p_k = x_k - x_{N-1}.$$

From (9), we have

$$w_{k+1} - 2g_N = \frac{N-k-1}{N-k}(w_k - 2g_N) - \frac{2}{N-k} g_{k+1} \implies w_{k+1} = \frac{m-1}{m} w_k - \frac{2}{m} d_k$$

and (10) yields

$$p_{k+1} = x_{k+1} - x_{N-1} = (x_k - x_{N-1}) - (z_{k+1} + 2g_N) - 2(g_{k+1} - g_N) = p_k - w_{k+1} - 2d_k.$$

Plugging the last identity into the expansion of $V_k - V_{k+1}$ below, we proceed as

$$\begin{aligned} V_k - V_{k+1} &= -\frac{m-1}{m}\left\|w_k\right\|^2 + \frac{2}{m}\left\langle w_k, p_k\right\rangle + \frac{m-2}{m-1}\left\|w_{k+1}\right\|^2 - \frac{2}{m-1}\left\langle w_{k+1}, p_{k+1}\right\rangle \\ &= -\frac{m-1}{m}\left\|w_k\right\|^2 + \frac{m}{m-1}\left\|w_{k+1}\right\|^2 + \left\langle \frac{2}{m} w_k - \frac{2}{m-1} w_{k+1}, p_k \right\rangle + \frac{4}{m-1}\left\langle w_{k+1}, d_k\right\rangle \\ &= \frac{4}{m(m-1)}\left\langle p_k, d_k\right\rangle - \frac{4}{m(m-1)}\left\|d_k\right\|^2 \\ &= \frac{4}{(N-k)(N-k-1)}\left\langle u_{k+1} - u_N, g_{k+1} - g_N\right\rangle, \end{aligned}$$

where in the third equality we substituted $w_{k+1} = \frac{m-1}{m} w_k - \frac{2}{m} d_k$ and canceled out the terms, and in the last equality we used

$$p_k - d_k = x_k - x_{N-1} - (g_{k+1} - g_N) = u_{k+1} - u_N.$$

Summing over $k = 0, \dots, N-2$ and using $V_{N-1} = 0$ yields

$$V_0 = \sum_{j=1}^{N-1} \frac{4}{(N-j+1)(N-j)} \langle u_j - u_N, g_j - g_N \rangle .$$

On the other hand, since $z_0 = 0$,

$$V_0 = -\frac{N-1}{N} \|2g_N\|^2 + \frac{2}{N} \langle 2g_N, x_0 - x_{N-1} \rangle = -\frac{4}{N} \left((N-1) \|g_N\|^2 + \langle g_N, x_{N-1} - x_0 \rangle \right).$$

Therefore, equating the two expressions for V_0 and multiplying $\frac{N}{4}$ throughout, we obtain

$$0 = (N-1) \|g_N\|^2 + \langle g_N, x_{N-1} - x_0 \rangle + \sum_{j=1}^{N-1} \frac{N}{(N-j)(N-j+1)} \langle u_j - u_N, g_j - g_N \rangle . \quad (11)$$

Finally, substituting $u_j = x_{j-1} - g_j$ and $g_j = \frac{\alpha}{2} G_j$, we have

$$\begin{aligned} \langle u_j - u_N, g_j - g_N \rangle &= \langle x_{j-1} - x_{N-1}, g_j - g_N \rangle - \|g_j - g_N\|^2 \\ &= \frac{\alpha}{2} \left(\langle x_{j-1} - x_{N-1}, G_j - G_N \rangle - \frac{\alpha}{2} \|G_j - G_N\|^2 \right), \end{aligned}$$

and

$$(N-1) \|g_N\|^2 = \frac{(N-1)\alpha^2}{4} \|G_N\|^2, \qquad \langle g_N, x_{N-1} - x_0 \rangle = \frac{\alpha}{2} \langle G_N, x_{N-1} - x_0 \rangle .$$

Hence, recalling $G_j = \mathbb{F}_{\mathcal{B}_{j-1}}(x_{j-1})$ and $\lambda_{N,j} = \frac{N\alpha}{2(N-j)(N-j+1)}$, (11) is exactly the claimed identity. □

A.2 Proof of Lemma 4.3

We can rewrite Lemma 4.2 as

$$\begin{aligned} 0 \geq \mathbb{E} \left[\frac{(N-1)\alpha^2}{4} \|\mathbb{F}(x_{N-1})\|^2 + \frac{\alpha}{2} \langle \mathbb{F}(x_{N-1}), x_{N-1} - x_0 \rangle \right] \\ + \sum_{j=1}^{N-1} \lambda_{N,j} \mathbb{E} \left[\langle x_{j-1} - x_{N-1}, \mathbb{F}_{\mathcal{B}_{j-1}}(x_{j-1}) - \mathbb{F}(x_{N-1}) \rangle - \frac{\alpha}{2} \left\| \mathbb{F}_{\mathcal{B}_{j-1}}(x_{j-1}) - \mathbb{F}(x_{N-1}) \right\|^2 \right]. \end{aligned} \quad (12)$$

Now let $e_{j-1} = \mathbb{F}_{\mathcal{B}_{j-1}}(x_{j-1}) - \mathbb{F}(x_{j-1})$ be the error term at iteration $j-1$. Then we can write

$$\begin{aligned} &\mathbb{E} \left[\langle x_{j-1} - x_{N-1}, \mathbb{F}_{\mathcal{B}_{j-1}}(x_{j-1}) - \mathbb{F}(x_{N-1}) \rangle - \frac{\alpha}{2} \left\| \mathbb{F}_{\mathcal{B}_{j-1}}(x_{j-1}) - \mathbb{F}(x_{N-1}) \right\|^2 \right] \\ &= \mathbb{E} \left[\langle x_{j-1} - x_{N-1}, \mathbb{F}(x_{j-1}) - \mathbb{F}(x_{N-1}) \rangle - \frac{\alpha}{2} \|\mathbb{F}(x_{j-1}) - \mathbb{F}(x_{N-1})\|^2 \right] \\ &\quad + \mathbb{E} \left[\langle x_{j-1} - x_{N-1}, e_{j-1} \rangle - \alpha \langle \mathbb{F}(x_{j-1}) - \mathbb{F}(x_{N-1}), e_{j-1} \rangle - \frac{\alpha}{2} \|e_{j-1}\|^2 \right] \\ &\geq \mathbb{E} \left[\langle -x_{N-1} + \alpha \mathbb{F}(x_{N-1}), e_{j-1} \rangle \right] - \frac{\alpha \sigma^2}{2B} \end{aligned}$$

where the last inequality uses cocoercivity of $\mathbb{F}$, $\mathbb{E}[e_{j-1} \,|\, \mathcal{F}_{j-1}] = 0$ and $x_{j-1} - \alpha \mathbb{F}(x_{j-1})$ is deterministic on $\mathcal{F}_{j-1}$, and $\mathbb{E} \left[\|e_{j-1}\|^2 \,\middle|\, \mathcal{F}_{j-1} \right] \leq \frac{\sigma^2}{B}$. Substituting the above into (12) and using $\mathbb{T} = \mathbb{I} - \alpha \mathbb{F}$, we obtain (6).

A.3 Proof of Lemma 4.4

Lemma A.1. For $k = 1, \dots, N-1$, the S-Dual-OHM iterates can be expressed as

$$x_k = \frac{1}{N}x_0 + \sum_{t=0}^{k-2} \frac{1}{(N-t-1)(N-t)} \mathbb{T}_{\mathcal{B}_t}(x_t) + \frac{N-k}{N-k+1} \mathbb{T}_{\mathcal{B}_{k-1}}(x_{k-1}). \tag{13}$$

Proof. We use induction on k. For $k = 1$, $x_1 = x_0 + \frac{N-1}{N}\big(\mathbb{T}_{\mathcal{B}_0}(x_0) - x_0\big) = \frac{1}{N}x_0 + \frac{N-1}{N}\mathbb{T}_{\mathcal{B}_0}(x_0)$, which agrees with (13). Now assume (13) holds for some $k \in \{1, \dots, N-2\}$. Then, by induction hypothesis,

$$\begin{aligned} x_{k+1} &= x_k + \frac{N-k-1}{N-k}\Big(\mathbb{T}_{\mathcal{B}_k}(x_k) - \mathbb{T}_{\mathcal{B}_{k-1}}(x_{k-1})\Big) \\ &= \frac{1}{N}x_0 + \sum_{t=0}^{k-2} \frac{1}{(N-t-1)(N-t)} \mathbb{T}_{\mathcal{B}_t}(x_t) + \left(\frac{N-k}{N-k+1} - \frac{N-k-1}{N-k}\right) \mathbb{T}_{\mathcal{B}_{k-1}}(x_{k-1}) \\ &\quad + \frac{N-k-1}{N-k} \mathbb{T}_{\mathcal{B}_k}(x_k). \end{aligned}$$

The induction (hence the proof) is then complete, as $\frac{N-k}{N-k+1} - \frac{N-k-1}{N-k} = \frac{1}{(N-k)(N-k+1)}$. □

Lemma A.2 (Leave-one-out stability)**.** For $s \in \{0, \dots, N-2\}$, let $\{x_k^{(s)}\}_{k=0}^{N-1}$ be the trajectory obtained from S-Dual-OHM by replacing only the minibatch $\mathcal{B}_s$ by an independent copy $\mathcal{B}_s'$, while keeping all other minibatches the same in the two runs. Then, for every $k = s+1, \dots, N-1$,

$$\mathbb{E}\left[\left\|x_k - x_k^{(s)}\right\|^2\right] \le \frac{\alpha^2\sigma^2}{B}\left(1 + \frac{(N-k)(N-k-1)}{(N-s)(N-s-1)}\right). \tag{14}$$

In particular,

$$\mathbb{E}\left[\left\|x_{N-1} - x_{N-1}^{(s)}\right\|^2\right] \le \frac{\alpha^2\sigma^2}{B}. \tag{15}$$

Proof. Define $w_{k,t} = \begin{cases} \dfrac{1}{(N-t-1)(N-t)}, & 0 \le t \le k-2, \\ \dfrac{N-k}{N-k+1}, & t = k-1 \end{cases}$ so that by Lemma A.1,

$$x_k = \frac{1}{N}x_0 + \sum_{t=0}^{k-1} w_{k,t}\mathbb{T}_{\mathcal{B}_t}(x_t), \qquad x_k^{(s)} = \frac{1}{N}x_0 + \sum_{t=0}^{k-1} w_{k,t}\mathbb{T}_{\mathcal{B}_t^{(s)}}(x_t^{(s)}),$$

where $\mathcal{B}_t^{(s)} = \mathcal{B}_t$ for $t \ne s$, and $\mathcal{B}_s^{(s)} = \mathcal{B}_s'$. Since $\sum_{t=0}^{k-1} w_{k,t} = 1 - \frac{1}{N}$, by Jensen's inequality,

$$\begin{aligned} d_k^{(s)} := \mathbb{E}\left[\left\|x_k - x_k^{(s)}\right\|^2\right] &\le \mathbb{E}\left[\sum_{t=0}^{k-1} w_{k,t}\left\|\mathbb{T}_{\mathcal{B}_t}(x_t) - \mathbb{T}_{\mathcal{B}_t^{(s)}}\big(x_t^{(s)}\big)\right\|^2\right] \\ &= \sum_{t=s}^{k-1} w_{k,t}\mathbb{E}\left[\left\|\mathbb{T}_{\mathcal{B}_t}(x_t) - \mathbb{T}_{\mathcal{B}_t^{(s)}}\big(x_t^{(s)}\big)\right\|^2\right]. \end{aligned} \tag{16}$$

Let $\nu^2 = \frac{\alpha^2\sigma^2}{B}$ for simplicity. We will show by induction on k that

$$d_k^{(s)} \le \nu^2\left(1 + \frac{(N-k)(N-k-1)}{(N-s)(N-s-1)}\right). \tag{17}$$

For $k = s+1$, from (16) we have

$$d_{s+1}^{(s)} \le w_{s+1,s}\mathbb{E}\left[\left\|\mathbb{T}_{\mathcal{B}_s}(x_s) - \mathbb{T}_{\mathcal{B}_s'}\big(x_s\big)\right\|^2\right] \le \frac{N-s-1}{N-s} \cdot 2\nu^2$$

which agrees with (17). Here we used the fact that $\mathbb{T}_{\mathcal{B}_s}(x_s)$ and $\mathbb{T}_{\mathcal{B}'_s}(x_s)$ are estimators of $\mathbb{T}(x_s)$ of variance $\leq \frac{\alpha^2\sigma^2}{B} = \nu^2$, and they are conditionally independent on $\mathcal{F}_s$. Next, let $s+2 \leq \ell \leq N-1$, and suppose (17) holds for all $t = s+1, \ldots, \ell-1$. Note that in (16), we have $\mathcal{B}_t^{(s)} = \mathcal{B}_t$ for $t > s$, so by Assumption 3.2,

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbb{T}_{\mathcal{B}_t}(x_t) - \mathbb{T}_{\mathcal{B}_t^{(s)}}\big(x_t^{(s)}\big)\right\|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left\|\mathbb{T}_{\mathcal{B}_t}(x_t) - \mathbb{T}_{\mathcal{B}_t}\big(x_t^{(s)}\big)\right\|^2 \,\middle|\, \mathcal{G}_t^{(s)}\right]\right] \\ &\leq \mathbb{E}\left[\mathbb{E}\left[\left\|x_t - x_t^{(s)}\right\|^2 \,\middle|\, \mathcal{G}_t^{(s)}\right]\right] = \mathbb{E}\left[\left\|x_t - x_t^{(s)}\right\|^2\right] = d_t^{(s)}\end{aligned}$$

where $\mathcal{G}_t^{(s)} = \sigma(\mathcal{B}_0, \ldots, \mathcal{B}_{t-1}, \mathcal{B}'_s)$ is the σ-algebra enlarged from $\mathcal{F}_t$ to make $x_t^{(s)}$ measurable. Applying this to (16) with the induction hypothesis, we obtain

$$\begin{aligned}d_\ell^{(s)} &\leq 2w_{\ell,s}\nu^2 + \sum_{t=s+1}^{\ell-1} w_{\ell,t} d_t^{(s)} \\ &\leq 2w_{\ell,s}\nu^2 + \sum_{t=s+1}^{\ell-1} w_{\ell,t}\nu^2\left(1 + \frac{(N-t)(N-t-1)}{(N-s)(N-s-1)}\right) \\ &= \nu^2\left[2w_{\ell,s} + \sum_{t=s+1}^{\ell-1} w_{\ell,t}\right] + \frac{\nu^2}{(N-s)(N-s-1)}\sum_{t=s+1}^{\ell-1} w_{\ell,t}(N-t)(N-t-1).\end{aligned} \tag{18}$$

Now using the explicit formula for $w_{\ell,t}$, we can directly verify

$$2w_{\ell,s} + \sum_{t=s+1}^{\ell-1} w_{\ell,t} = 1 - \frac{1}{N-s} + \frac{1}{(N-s)(N-s-1)},$$

and

$$\begin{aligned}&\sum_{t=s+1}^{\ell-1} w_{\ell,t}(N-t)(N-t-1) = \sum_{t=s+1}^{\ell-2} 1 + \frac{N-\ell}{N-\ell+1}(N-\ell+1)(N-\ell) \\ &= \ell - s - 2 + (N-\ell)^2 = N - s - 2 + (N-\ell)(N-\ell-1).\end{aligned}$$

Hence

$$\begin{aligned}d_\ell^{(s)} &\leq \nu^2\left(1 - \frac{1}{N-s}\right) + \frac{\nu^2}{(N-s)(N-s-1)}\Big(N - s - 1 + (N-\ell)(N-\ell-1)\Big) \\ &= \nu^2\left(1 + \frac{(N-\ell)(N-\ell-1)}{(N-s)(N-s-1)}\right),\end{aligned}$$

which completes the induction, proving (14). Then (15) follows immediately by setting $k = N-1$. □

Now we are ready to prove Lemma 4.4. Fix $j \in \{1, \ldots, N-1\}$, and let $x_{N-1}^{(j-1)}$ denote the leave-one-out terminal iterate from Lemma A.2, obtained by replacing only the minibatch $\mathcal{B}_{j-1}$ by an independent copy. By the fact that $x_{N-1}^{(j-1)}$ is conditionally independent with e_{j-1} (which depends only on $\mathcal{B}_{j-1}$) on $\mathcal{F}_{j-1}$ and the tower property, we have $\mathbb{E}\left[\left\langle \mathbb{T}\big(x_{N-1}^{(j-1)}\big), e_{j-1}\right\rangle\right] = 0$. Hence

$$\begin{aligned}\mathbb{E}\left[\langle e_{j-1}, \mathbb{T}x_{N-1}\rangle\right] &= \mathbb{E}\left[\left\langle e_{j-1}, \mathbb{T}(x_{N-1}) - \mathbb{T}\big(x_{N-1}^{(j-1)}\big)\right\rangle\right] \\ &\leq \sqrt{\mathbb{E}\left[\|e_{j-1}\|^2\right]\ \mathbb{E}\left[\left\|\mathbb{T}(x_{N-1}) - \mathbb{T}\big(x_{N-1}^{(j-1)}\big)\right\|^2\right]} \\ &\leq \sqrt{\frac{\sigma^2}{B}\mathbb{E}\left[\left\|x_{N-1} - x_{N-1}^{(j-1)}\right\|^2\right]} \\ &\leq \sqrt{\frac{\sigma^2}{B}\cdot\frac{\alpha^2\sigma^2}{B}} = \frac{\alpha\sigma^2}{B}.\end{aligned}$$

A.4 Proof of Lemma 4.6

Let

$$s_k = \mathbb{E}\left\|x_k - y_k\right\|^2, \qquad M_k = \max_{0\le i\le k} s_i.$$

It suffices to show that $M_k \le \frac{\alpha^2\sigma^2}{B(1-\gamma^2)}$. For $k=0$, we have $x_0 = y_0$, so we have $s_0 = M_0 = 0$ and the result is trivial. Let $k \ge 1$. Applying Lemma A.1 for both S-Dual-OHM and its full-batch version (Dual-OHM), we obtain

$$x_k - y_k = \sum_{t=0}^{k-1} w_{k,t}\left[\mathbb{T}(x_t) - \mathbb{T}(y_t) + e_t\right]$$

where $e_t = \mathbb{T}_{\mathcal{B}_t}(x_t) - \mathbb{T}(x_t)$ and $w_{k,t} = \begin{cases} \dfrac{1}{(N-t-1)(N-t)}, & 0 \le t \le k-2 \\ \dfrac{N-k}{N-k+1}, & t = k-1 \end{cases}$. Note that here $\mathbb{E}\left[e_t \,|\, \mathcal{F}_t\right] = 0$ and $\mathbb{E}\left[\left\|e_t\right\|^2 \,\middle|\, \mathcal{F}_t\right] \le \nu^2 := \frac{\alpha^2\sigma^2}{B}$. Using $\sum_{t=0}^{k-1} w_{k,t} = 1 - \frac{1}{N} < 1$ and Jensen's inequality gives

$$s_k = \mathbb{E}\left[\left\|x_k - y_k\right\|^2\right] \le \sum_{t=0}^{k-1} w_{k,t}\mathbb{E}\left[\left\|\mathbb{T}(x_t) - \mathbb{T}(y_t) + e_t\right\|^2\right]$$

$$\le \sum_{t=0}^{k-1} w_{k,t}\left(\mathbb{E}\left[\left\|\mathbb{T}(x_t) - \mathbb{T}(y_t)\right\|^2 + \left\|e_t\right\|^2\right]\right) \le \sum_{t=0}^{k-1} w_{k,t}\left(\gamma^2\mathbb{E}\left[\left\|x_t - y_t\right\|^2\right] + \nu^2\right) \le \gamma^2 M_{k-1} + \nu^2$$

where for the second inequality, we used conditional unbiasedness together with the tower property:

$$\mathbb{E}\left[\langle e_t, \mathbb{T}(x_t) - \mathbb{T}(y_t)\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\langle e_t, \mathbb{T}(x_t) - \mathbb{T}(y_t)\rangle \,|\, \mathcal{F}_t\right]\right] = 0$$

which holds because $\mathbb{T}(x_t) - \mathbb{T}(y_t)$ is deterministic conditioned on $\mathcal{F}_t$ and $\mathbb{E}\left[e_t \,|\, \mathcal{F}_t\right] = 0$. Now starting with $M_0 = 0$ and using induction on k, the above shows that

$$M_k \le \max\left\{M_{k-1}, \gamma^2 M_{k-1} + \nu^2\right\} \le \max\left\{M_{k-1}, \gamma^2\frac{\nu^2}{1-\gamma^2} + \nu^2\right\} = \frac{\nu^2}{1-\gamma^2}.$$

A.5 Proof of Lemma 4.7

Observe that for $k = 1, 2, \ldots,$

$$\left\|y_{k+1} - y_k\right\| = \frac{N-k-1}{N-k}\left\|\mathbb{T}(y_k) - \mathbb{T}(y_{k-1})\right\| \le \gamma\left\|y_k - y_{k-1}\right\|$$

$$\implies \left\|y_{k+1} - y_k\right\| \le \gamma^k\left\|y_1 - y_0\right\| = \gamma^k\left\|\frac{N-1}{N}(\mathbb{T}(y_0) - y_0)\right\| \le \gamma^k(1+\gamma)\left\|y_0 - x_\star\right\|$$

where the last inequality uses $\left\|\mathbb{T}(y_0) - y_0\right\| \le \left\|\mathbb{T}(y_0) - x_\star\right\| + \left\|x_\star - y_0\right\| \le (1+\gamma)\left\|y_0 - x_\star\right\|$ (recall that we assume $x_0 = y_0$). Thus

$$\left\|y_{N-1} - y_k\right\| \le \sum_{j=k}^{N-2}\left\|y_{j+1} - y_j\right\| \le (1+\gamma)\left\|y_0 - x_\star\right\|\sum_{j=k}^{N-2}\gamma^j \le \frac{1+\gamma}{1-\gamma}\gamma^k\left\|x_0 - x_\star\right\|.$$

B Additional Experiments

In the experiments of Section 5, we used for S-OHM the same batch size B selected for S-Dual-OHM. This was designed to isolate the effect of replacing the anchor update by the dual-anchor update while keeping all the other configurations identical. However, to verify that the observed empirical advantage of S-Dual-OHM is not an artifact of this experiment design, we repeat all four experiments while separately tuning the batch size of S-OHM over the same grid $B \in \{1, 10, 20, 50, 100\}$, under the same total sample budget $Q = NB$. All the other experimental settings are unchanged. For simplicity, we only compare the results from S-Dual-OHM with tuned B and runs of S-OHM using the shared and tuned batch sizes.

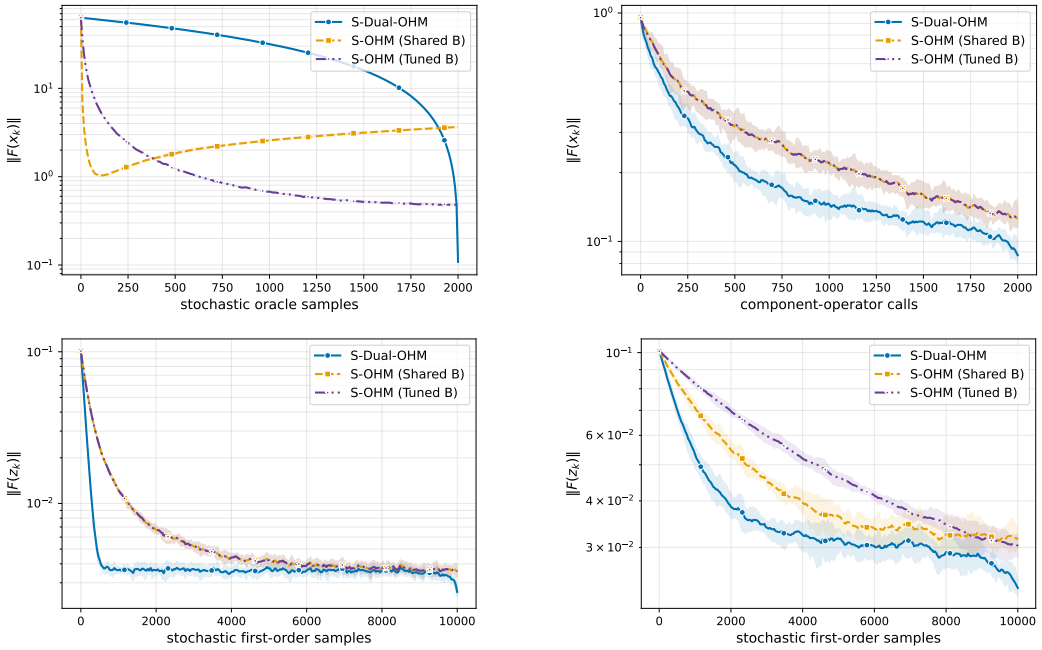

Figure 3: Comparison after independently tuning the constant batch size B of S-OHM. **(Top left)** Worst-case nonexpansive operator (Experiment 1). **(Top right)** Finite-sum cocoercive operator (Experiment 2). **(Bottom left)** SCSC Huber-type minimax problem (Experiment 3), in the low-variance regime with $\sigma = 0.05$. **(Bottom right)** The same minimax problem from Experiment 3 in the high-variance regime with $\sigma = 1.5$. The plots labeled as S-OHM (Tuned B) use the best $B \in \{1, 10, 20, 50, 100\}$ selected under the given total sample budget. Solid curves are means over 10 independent runs, and shaded bands denote empirical 5th–95th percentiles.

As shown in Figure 3, independent tuning selects the same batch size for S-OHM and S-Dual-OHM in Experiment 2 and the low-variance regime in Experiment 3. In Experiment 1 and the high-variance regime in Experiment 3, the batch sizes independently chosen for S-OHM differ. Nevertheless, even with its own optimized batch size, the final residual attained by S-OHM remains strictly larger than that of S-Dual-OHM. These results from controlled experiments further supports the empirical benefit/robustness of the dual-anchor mechanism over anchoring in stochastic settings.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our Theorems 1 and 2 show the oracle complexities of $\mathcal{O}(\epsilon^{-3})$ and $\widetilde{\mathcal{O}}(\epsilon^{-2})$ respectively, which are the main theoretical results we claim.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We do it in our Section 6, *Conclusion, Limitations and Open Questions.*

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Section 3 states all necessary concepts and assumptions for the analysis, and each main theorem in Section 4 accurately references the required assumptions.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all details of the experiments in Section 5, including the problem definitions and parameter setups, with all details of the algorithms' definitions and hyperparameters, and how they are tuned.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the code from which all experimental results can be reproduced.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We provide all details of the experiments in Section 5, including the problem definitions and parameter setups, with all details of the algorithms' definitions and hyperparameters, and how they are tuned.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We handle stochastic optimization, and we clearly explain that we plot mean performance plots together with 5th–95th percentile intervals.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Section 5 we explain that all experiments were run on a single laptop, with specs specified.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work is primarily theoretical and does not have any foreseeable societal impact or harmful consequences.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our work is primarily theoretical and does not have any foreseeable societal impact or harmful consequences.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: N/A

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: N/A

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: N/A

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: N/A

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: N/A

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: N/A

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.