LOWER BOUNDS AND PROXIMALLY ANCHORED SGD FOR NON-CONVEX MINIMIZATION UNDER UNBOUNDED VARIANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Analysis of Stochastic Gradient Descent and its variants typically relies on the assumption of uniformly bounded variance, a condition that frequently fails in practical non-convex settings, such as neural network training, and in several elementary optimization settings. While several relaxations are explored in the literature, the classical Blum-Gladyshev (BG-0) condition, which permits the variance to grow quadratically with distance, has recently been shown to be the weakest condition. However, the study of the oracle complexity of stochastic first-order non-convex optimization under BG-0 has remained underexplored. In this paper, we address this gap and establish information-theoretic lower bounds, proving that finding an ϵ-stationary point requires $\Omega(\epsilon^{-6})$ stochastic BG-0 oracle queries for smooth functions and $\Omega(\epsilon^{-4})$ queries under mean-square smoothness. These limits demonstrate an unavoidable degradation from classical bounded-variance complexities, i.e., $\Omega(\epsilon^{-4})$ and $\Omega(\epsilon^{-3})$ for smooth and mean-square smooth cases, respectively. To match these lower bounds, we consider Proximally Anchored STochastic Approximation (PASTA), a unified algorithmic framework whose problem-dependent instantiations combine dynamic batching and gradient estimation with Halpern anchoring and Tikhonov regularization. We prove that suitable PASTA instantiations achieve minimax optimal complexities across numerous non-convex regimes, including standard smooth, mean-square smooth, weakly convex, star-convex, and Polyak-Lojasiewicz functions, entirely under an unbounded domain and unbounded stochastic gradients.

1 INTRODUCTION

We consider the canonical unconstrained stochastic optimization problem

$$\min_{x\in\mathbb{R}^p} f(x) = \mathbb{E}_\xi[\tilde{f}(x,\xi)], \tag{1}$$

where $f:\mathbb{R}^p\to\mathbb{R}$ is a generally non-convex stochastic objective function. Because exact global minimization of non-convex objectives is intractable in the worst case, the standard algorithmic goal relaxes to identifying an ϵ-stationary point.

The theoretical analysis of Stochastic Gradient Descent (SGD) and its variance-reduced variants generally relies on the bounded variance assumption $\mathbb{E}[\|\widetilde{\nabla} f(x) - \nabla f(x)\|^2] \le \sigma^2$ for an unbiased estimator $\widetilde{\nabla} f(x)$. While convenient, a uniformly bounded variance fails to hold even in elementary settings such as unconstrained least squares or deep neural network training, where the gradient variance inherently grows with the distance from the optimal solution. While several relaxations are explored in the literature, the classical Blum-Gladyshev (BG-0) condition (Blum, 1954; Gladyshev, 1965) is shown to be the weakest viable variance assumption (Alacaoglu et al., 2025). To capture this structural growth, BG-0 posits that, for a given reference anchor x_0, the variance is permitted to scale quadratically: $\mathbb{E}[\|\widetilde{\nabla} f(x) - \nabla f(x)\|^2] \le B^2\|x - x_0\|^2 + G^2$. The extra variance term $B^2\|x - x_0\|^2$ introduces an undesirable quadratic term in the analyses. Over unbounded domains, this extra variance term breaks the contractive mechanisms of traditional convergence proofs, which rely on the descent lemma, leading to divergence unless highly restrictive assumptions such as an

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

artificially bounded domain are imposed. Let us formalize a *generalization of the existing BG-0 oracle* below, which is leveraged for all of our theoretical results.

Assumption 1 (BG-0 Oracle (Blum, 1954; Gladyshev, 1965; Alacaoglu et al., 2025))**.** *Let $\mathcal{X} \subset \mathbb{R}^p$ be a closed and convex set and consider some $x_0 \in \mathcal{X}$. The stochastic oracle $\widetilde{\nabla} f(x)$ is unbiased, i.e., $\mathbb{E}[\widetilde{\nabla} f(x)] = \bar{g}$ for some $\bar{g} \in \partial f(x)$. Also, there exist constants $B \geq 0$ and $G \geq 0$ such that*

$$\inf_{\bar{g} \in \partial f(x)} \mathbb{E}\left[\left\|\widetilde{\nabla} f(x) - \bar{g}\right\|^2\right] \leq B^2\|x - x_0\|^2 + G^2 \quad \forall x \in \mathcal{X}. \tag{2}$$

For a differentiable function f, $\bar{g}$ reduces to the unique gradient $\nabla f(x)$ such that the variance satisfies $\mathbb{E}\left[\|\widetilde{\nabla} f(x) - \nabla f(x)\|^2\right] \leq B^2\|x - x_0\|^2 + G^2$.

Please see Appendix A for a detailed discussion of the related work.

In convex settings, recent literature (Neu & Okolo, 2024; Alacaoglu et al., 2025) has successfully leveraged Halpern anchoring (Halpern, 1967) to stabilize trajectories; by drawing the current state toward a static reference point x_0 with an anchoring sequence $\beta_t \in (0, 1)$ alongside the step size $\eta_t > 0$, the proximal anchored update is defined by $x_{t+1} = \beta_t x_0 + (1 - \beta_t)x_t - \eta_t \widetilde{\nabla} f(x_t)$. This mechanism, along with the assumption of convexity (which inherently enables controlling the growth of the distance), enables us to deal with the extra variance term $B^2\|x - x_0\|^2$ and to recover the classical rates under the bounded variance assumption.

However, in contrast with convex functions, the non-convex regimes generally lack global contractive mechanisms to inherently counteract the expanding noise. As all known, optimal convergence rates and information-theoretic lower bounds for non-convex stochastic optimization (Arjevani et al., 2023) rely crucially on the uniform bounded-variance assumption; it has remained open whether the classical rates can even be recovered under the BG-0 condition, and more broadly, what the true information-theoretic limits of non-convex optimization are in this relaxed variance regime.

Contributions. Intuitively, one should expect worse complexity bounds under the BG-0 condition for non-convex minimization. If the variance scales quadratically with distance, an adversarial landscape can force an algorithm to traverse a vast distance from the initialization before encountering the stationary region. As the optimizer traverses this distance, the noise magnitude may explode as a function of ϵ^{-1}, exponentially degrading the signal-to-noise ratio, and as a consequence, the attainable oracle complexity. Somewhat interestingly, we formally leverage this intuition to prove that the celebrated bounded-variance lower bounds of Arjevani et al. (2023) do not hold under BG-0; the limits of performance are worse.

To match these new limits, we consider a unified algorithmic framework that enables the study of various non-convex regimes without bounded domain and stochastic gradient assumptions. To circumvent the incurred bias of traditional Halpern tracking, which may prevent convergence under non-convexity, we enforce a coupling between the anchoring parameter and the step size, i.e., $\beta_t = \lambda\eta_t$ for some $\lambda > 0$. Under this parameter coupling, the progression of the proximal anchored update becomes algebraically equivalent to a standard step on the strongly convex Tikhonov-regularized surrogate $F_\lambda(x) = f(x) + \frac{\lambda}{2}\|x - x_0\|^2$, for a sufficiently large λ. We introduce this Tikhonov-regularized surrogate not merely as a stabilization heuristic, but as a principled (and to our understanding, necessary) approach to optimize diverse non-convex functions, especially *weakly convex* ones, a rich class that, despite its potentially misleading name, generalizes standard smooth non-convex functions. This specific coupling absorbs the extra variance in Assumption 1 by shifting the optimization target to the unique global minimizer of the regularized subproblem. Because targeting $F_\lambda(x)$ induces a stationary bias relative to the true objective f, we embed optimizing the Tikhonov-regularized surrogate as an inner loop within an epoch-based framework, sequentially updating the anchor x_0 to mitigate the regularization bias. We call this unified framework Proximally Anchored STochastic Approximation (PASTA), which is summarized as Framework 1. From the perspective of fixed-point methods, PASTA acts as the interpolation between two classical methodologies: configuring the algorithm for a single epoch (i.e., $S = 1$) recovers the anchored Halpern iteration (Halpern, 1967) (with our specific coupling), while limiting the inner-loop to a single step (i.e., $K = 1$) reduces the dynamics to the Krasnoselskii-Mann iteration (Krasnosel'skiĭ, 1955; Mann, 1953).

With this motivation and discussion, we now summarize our key theoretical contributions:

- **Information-Theoretic Lower Bounds:** We establish the oracle complexity lower bounds for finding ϵ-stationary points of smooth functions under the BG-0 condition over unbounded domains. By leveraging the exploding variance in remote landscape regions, we prove that any randomized algorithm requires $\Omega(\epsilon^{-6})$ stochastic gradient evaluations, a degradation from the classical $\Omega(\epsilon^{-4})$ bounded-variance result. Furthermore, under the additional assumption of mean-square smoothness, we establish a tight $\Omega(\epsilon^{-4})$ lower bound, again a degradation from the classical $\Omega(\epsilon^{-3})$ limit.
- **Minimax Optimal Upper Bounds:** We prove that PASTA successfully matches these new, degraded lower bounds without requiring bounded domains and bounded stochastic gradient assumptions. We show that specific instantiations of PASTA achieve the optimal $\mathcal{O}(\epsilon^{-6})$ complexity for smooth functions. Moreover, by integrating variance reduction mechanisms, PASTA recovers the optimal $\mathcal{O}(\epsilon^{-4})$ complexity under mean-square smoothness.
- **Geometric Universality:** We extend the convergence analysis of PASTA across numerous non-convex regimes. Beyond standard smoothness, we provide optimal convergence guarantees under weak convexity Davis & Drusvyatskiy (2019), star-convexity (Nesterov & Polyak, 2006), and the Polyak-Lojasiewicz condition (Polyak, 1963), demonstrating the universal effectiveness of PASTA against unbounded variance in diverse non-convex stochastic optimization setups.

Notation: We consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and denote the expectation with respect to the underlying probability measure by $\mathbb{E}[\cdot]$. The conditional expectation with respect to the filtration $\mathcal{F}_{t-1}$, representing the history up to the start of iteration t, is denoted by $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. Throughout, $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denote the Euclidean norm and inner product, respectively.

2 Non-convex lower bounds with BG-0 oracle

In this section, we establish oracle complexity lower bounds for finding ϵ-stationary points of smooth non-convex functions under the BG-0 variance condition over unbounded domains. For smooth functions, $\mathbb{E}[\|\nabla f(x)\|^2] \leq \epsilon^2$ serves as a natural convergence criterion. Classical lower bounds for non-convex stochastic optimization rely on the bounded variance assumption, yielding an optimal rate of $\Omega(\epsilon^{-4})$ for generally smooth functions and $\Omega(\epsilon^{-3})$ when mean-square smoothness holds (Arjevani et al., 2023). We demonstrate that allowing the variance to grow quadratically with the distance inherently degrades the information-theoretic limits of the optimization landscape. Let us first state smoothness formally, and then we state our lower bound under Assumption 1.

Assumption 2 (Smoothness). *A function $f : \mathbb{R}^p \to \mathbb{R}$ is L-smooth if for all $x, y \in \mathbb{R}^p$ it holds that $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.*

Theorem 1 (Lower Bound under Smoothness). *For any $L > 0$, $\Delta > 0$, $B \geq 0$, $G \geq 0$, and $\epsilon \leq \sqrt{\frac{L\Delta}{1536\ell_1}}$ for some numerical constant $\ell_1 > 0$, any randomized algorithm requires $\Omega\big(\frac{B^2 L\Delta^3}{\epsilon^6} + \frac{G^2 L\Delta}{\epsilon^4}\big)$ stochastic oracle queries in expectation to find a point x satisfying $\mathbb{E}[\|\nabla F(x)\|] \leq \epsilon$ for a function F that satisfies $F(x_0) - \inf_x F(x) \leq \Delta$, and Assumptions 1 and 2.*

The proof is in Appendix B.1. The emergence of the $\Omega(\epsilon^{-6})$ complexity represents a degradation compared to the classical $\Omega(\epsilon^{-4})$ rate proved by the influential work of Arjevani et al. (2023). To understand the source of this lower bound, it is instructive to contrast our hard instance construction with the seminal bounded-variance framework established by Arjevani et al. (2023). Under a uniform variance bound σ^2, the standard lower bound works by constructing a high-dimensional probabilistic zero-chain $F : \mathbb{R}^T \to \mathbb{R}$ (see (19) for its definition). In this chain, the true gradient signal is masked by noise, requiring $\Omega(\sigma^2/\epsilon^2)$ queries in expectation to discover each subsequent relevant coordinate. To satisfy the stationarity condition, an algorithm must activate $\Theta(\Delta L/\epsilon^2)$ such coordinates, thereby leading to the $\Omega(\Delta L\sigma^2/\epsilon^4)$ bound.

Our novel construction, on the other hand, leverages the dependence of the BG-0 oracle on distances against the optimization algorithm. First, we consider a suitable scaling of the zero-chain $F_{\text{scaled}} : \mathbb{R}^T \to \mathbb{R}$, critically requiring a shift not used by Arjevani et al. (2023) (see (23) for its definition). Then, we isolate the information bottleneck by coupling a one-dimensional travel function $f_0(u)$ with the zero-chain subspace (see (21) for its definition). The composite objective $F(u, y) = f_0(u) + \phi(u)F_{\text{scaled}}(y)$ requires that the algorithm traverse $D = \Theta(\Delta/\epsilon)$ along the travel coordinate u to drop the gradient norm below the threshold ϵ. We use a smooth activation function $\phi(u)$ to

ensure that the gradient subspace corresponding to the high-dimensional zero-chain $F_{\text{scaled}}(y)$ remains inactive until this mandatory traversal is completed (see (22) for its definition).

By the time the algorithm reaches the active region of the zero-chain, its distance from the initialization x_0 satisfies $\|x - x_0\|^2 \geq \Theta(\Delta^2/\epsilon^2)$. The BG-0 condition thus dictates that the variance at this remote location is $\tilde{\sigma}^2 = \Theta(B^2\Delta^2/\epsilon^2 + G^2)$. By substituting this big variance budget into the standard zero-chain mechanics of Arjevani et al. (2023), the probability of the oracle revealing the true gradient direction shrinks from $\Theta(\epsilon^2)$ to $\Theta(\epsilon^4)$. The algorithm is thus forced to optimize the non-convex structure exactly where the signal-to-noise ratio is worst, yielding the $\Omega(\epsilon^{-6})$ limit.

We now focus on the case of having Mean-Square Smoothness (MSS), which enables improving the lower bound. The condition is defined next, followed by our second lower bound.

Assumption 3 (Mean-Square Smoothness)**.** *A function $f : \mathbb{R}^p \to \mathbb{R}$ is $\bar{L}$-smooth in the mean square sense if for all $x, y \in \mathbb{R}^p$ it holds that $\mathbb{E}[\|\widetilde{\nabla} f(x) - \widetilde{\nabla} f(y)\|^2] \leq \bar{L}^2\|x - y\|^2$.*

Theorem 2 (Lower Bound under Mean-Square Smoothness)**.** *For any $\bar{L} > 0$, $\Delta > 0$, $G \geq 0$, $0 \leq B \leq \frac{\varsigma \bar{L}}{96\bar{\ell}_1}$ for some numerical constants $\bar{\ell}_1 > 0$, $\varsigma > 0$, and with $0 < \epsilon \leq \frac{\sqrt{-2G^2 + \sqrt{4G^4 + \varsigma^2(\vartheta - B^2\Delta^2)}}}{4\varsigma}$ where $\vartheta = \frac{\varsigma^2 \bar{L}^2 \Delta^2}{9216\bar{\ell}_1^2}$, any randomized algorithm requires $\Omega\big(\frac{B\bar{L}\Delta^2}{\epsilon^4} + \frac{G\bar{L}\Delta}{\epsilon^3} + \frac{B^2\Delta^2}{\epsilon^4} + \frac{G^2}{\epsilon^2}\big)$ stochastic oracle queries in expectation to find a point x satisfying $\mathbb{E}[\|\nabla F(x)\|] \leq \epsilon$ for a function F that satisfies $F(x_0) - \inf_x F(x) \leq \Delta$, and Assumptions 1 and 3.*

The proof is in Appendix B.2. Theorem 2 formally establishes that while MSS helps improve the complexity, the distance-dependent variance allowed by BG-0 can still yield an $\Omega(\epsilon^{-4})$ dependence, compared with the classical $\mathcal{O}(\epsilon^{-3})$ bounds attainable under uniform variance and MSS. The bottleneck from the extra variance term allowed by BG-0 manifests in two terms. The term $\Omega(B\bar{L}\Delta^2/\epsilon^4)$ characterizes the sequential optimization complexity. To maintain the global MSS condition characterized by $\bar{L}$ despite the extra variance at distance D, the deterministic smoothness L of the underlying zero-chain must be controlled. Doing so couples the chain length to the variance limit, forcing more sequential queries to overcome the progressive information masking. Conversely, the term $\Omega(B^2\Delta^2/\epsilon^4)$ represents a statistical estimation floor arising from a distinct mean-estimation argument. Any algorithm must evaluate gradients at the traversal boundary imposed by our construction to verify the ϵ-stationarity condition. Because the oracle introduces noise with variance $\Theta(B^2\Delta^2/\epsilon^2)$, computing a valid gradient vector to precision ϵ requires $\Omega(B^2\Delta^2/\epsilon^4)$ independent stochastic samples.

3 LET'S COOK PASTA!

We now introduce the PASTA family, summarized in Framework 1. The framework specifies a parameterized update, and each theorem below gives a complete instantiation tailored to its problem class. The primary algorithmic challenge imposed by the BG-0 oracle condition is the unbounded growth of the stochastic gradient variance as the sequence of iterates drifts away from the initialization. Traditional unanchored stochastic gradient methods lack the necessary contractive mechanism to suppress this extra variance, leading to divergence over unbounded domains. PASTA instantiations mitigate this issue through problem-dependent choices of the sampling rule and gradient estimator, together with a Halpern-style anchoring mechanism (Neu & Okolo, 2024; Alacaoglu et al., 2025) embedded within a multi-epoch proximal point architecture.

Let us first talk about the update step used in the inner loop. Under MSS, classical methods such as SGD and SGD with momentum are sub-optimal, so they will also be sub-optimal under the BG-0 oracle. Optimal methods rely on variance reduction. Two well-known methods include PAGE and STORM (Li et al., 2021; Cutkosky & Orabona, 2019). Focusing on the former without loss of generality in this paper, to ensure we meet the lower bound established by Theorem 2 under MSS, the recursive estimator replacing the vanilla estimator of SGD takes the form

$$g_t = \begin{cases} \frac{1}{N_t}\sum_{i=1}^{N_t} \widetilde{\nabla} f(x_t, \xi_t^{(i)}) & \text{with probability } p_t \\ g_{t-1} + \frac{1}{b_t}\sum_{i=1}^{b_t} \left(\widetilde{\nabla} f(x_t, \zeta_t^{(i)}) - \widetilde{\nabla} f(x_{t-1}, \zeta_t^{(i)})\right) & \text{with probability } 1 - p_t \end{cases} \tag{3}$$

where N_t is a large batch size and b_t is a small secondary batch size. Note that when $p_t = 1$, the update reduces to that of simple SGD.

Framework 1 Proximally Anchored STochastic Approximation (PASTA)

1: **Parameters:** Initial point x_0, epoch length K, total epochs S, step size $\eta_t > 0$, anchoring parameter β_t, reset parameter p_t.
2: **for** $s = 1, 2, \ldots, S$ **do**
3: **for** $t = 0, 1, \ldots, K-1$ **do**
4: Compute g_t according to (3) by sampling the BG-0 oracle N_t or b_t times depending on $u \sim \mathrm{Ber}(p_t)$
5: Do the coupled proximal anchored update $x_{t+1} = \beta_t x_{s-1} + (1-\beta_t)x_t - \eta_t g_t$.
6: **end for**
7: Set the output of epoch s to the last iterate $x_s = x_K$.
8: **end for**

As mentioned, PASTA operates over S consecutive epochs, each consisting of K inner stochastic gradient steps. At the beginning of each epoch s, the algorithm establishes a fixed local anchor x_{s-1}. During the inner iterations, PASTA forms an update vector according to (3) and executes the coupled update $x_{t+1} = \beta_t x_{s-1} + (1-\beta_t)x_t - \eta_t g_t$. Note that, when $p_t = 1$, the update vector reduces to a simple mini-batch stochastic gradient g_t of size N_t. In that case, this update step interpolates between the static anchor x_{s-1} and standard SGD.

A key feature of PASTA, particularly for weakly convex analysis, is the coupling of the anchoring parameter β_t and the step size η_t. By enforcing the relation $\beta_t = \lambda\eta_t$ for some regularization parameter $\lambda > 0$, the update naturally reformulates as an attempt to optimize the Tikhonov-regularized surrogate objective $F_\lambda(x) = f(x) + \frac{\lambda}{2}\|x - x_{s-1}\|^2$. This regularization induces λ-strong convexity into the local optimization landscape. The induced strong convexity dynamically counteracts the extra variance term $B^2\|x - x_0\|^2$ in the BG-0 assumption, thereby bounding the maximum experienced variance. However, optimizing the surrogate $F_\lambda(x)$ inherently introduces a stationary bias relative to the true objective $f(x)$. PASTA resolves this by systematically resetting the anchor to the final iterate of the previous epoch, $x_s = x_K$. This outer-loop mechanism mirrors the classical inexact proximal point algorithm, progressively removing the regularization bias and driving the sequence toward a true ϵ-stationary point of the original non-convex objective. By configuring the epoch length K, the batch size N_t, and other parameters, PASTA can recover minimax optimal complexities across a wide spectrum of non-convex regimes, as we show in the next section.

4 Non-convex upper bounds with BG-0 oracle

Having established the lower bounds under the BG-0 condition in Section 2, we now demonstrate that PASTA achieves these limits with the right instantiation of the parameters. Beyond standard smoothness and MSS, we also consider other classes of non-convex functions. Different problem geometries activate different components of the PASTA framework: the smooth and MSS results rely on dynamic batching, while the weakly convex result uses proximal anchoring.

4.1 Optimal complexity under smoothness

Under standard L-smoothness, our lower bound is $\Omega(\epsilon^{-6})$. To achieve this limit, we note that we can directly control the BG-0 variance by varying the sampling budget. In particular, we use a single epoch ($S = 1$) and eliminate the anchor ($\beta_t = 0$), such that with $p_t = 1$, PASTA simplifies to standard unanchored SGD. To prevent the extra variance term allowed by the BG-0 oracle from causing convergence issues, we scale the batch size dynamically at each step, proportional to the squared distance from the initialization. This ensures the variance of the mini-batch estimator remains bounded by a global constant, achieving the $\mathcal{O}(\epsilon^{-6})$ limit through sampling adaptation, leading to the following result.

Theorem 3 (Convergence of PASTA via Dynamic Batching - Smooth Non-convex Case)**.** *Let Assumptions 1 and 2 hold. Let the sequence $\{x_t\}$ be generated by PASTA with $S = 1$, $\beta_t = 0$, and $p_t = 1$. Given a constant target variance $\sigma^2 > 0$, set batch size N_t, the step size η, and total iterations K according to*

$$N_t = \max\left\{1, \left\lceil \frac{B^2\|x_t - x_0\|^2 + G^2}{\sigma^2} \right\rceil\right\}, \quad \eta = \min\left\{\frac{1}{L}, \frac{\epsilon^2}{2L\sigma^2}\right\}, \quad K = \left\lceil \frac{4\Delta}{\eta\epsilon^2} \right\rceil, \qquad (4)$$

where $\Delta \geq f(x_0) - \inf_x f(x)$. *Then,* $\frac{1}{K}\sum_{t=0}^{K-1} \mathbb{E}[\|\nabla f(x_t)\|^2] \leq \epsilon^2$ *and the expected total stochastic oracle complexity evaluates to* $\mathcal{O}(\epsilon^{-6})$.

The proof is in Appendix C.1, where the exact expected oracle complexity can be found in (58). Theorem 3 matches the established lower bound in Theorem 1.

4.2 Optimality under convexity: geometric-statistical uncertainty principle

While our focus is on non-convex minimization, a natural question is whether the PASTA instantiation we explored above (i.e., the good old unanchored SGD with dynamic batching) can recover the rates established in Alacaoglu et al. (2025) for convex functions. There, Alacaoglu et al. (2025) demonstrates that SGD with Halpern anchoring attains $\mathcal{O}(\epsilon^{-2})$ complexity under the BG-0 oracle assumption. Theorem 9 and 10 in Appendix C.2 provide an affirmative answer to the preceding question, showing *without Halpern anchoring*, dynamic batching is enough to ensure for the ergodic average $\bar{x}_K = \frac{1}{K}\sum_{t=0}^{K-1} x_t$, we have $\mathbb{E}[f(\bar{x}_K) - f(x^*)] \leq \epsilon$ and the expected total stochastic oracle complexity evaluates to $\mathcal{O}(\epsilon^{-2})$.

Our results in Appendix C.2, along with those of Alacaoglu et al. (2025), demonstrate two distinct instantiations of PASTA to obtain optimal complexity under the BG-0 oracle and convexity. A natural question is whether these two instances are related, as one is anchored and uses a single oracle call per iteration, whereas the other is unanchored and uses dynamic batching. In what follows, we show that these two distinct instantiations are conjugate and satisfy an uncertainty principle.

The proofs of Theorem 9 and 10 (see Appendix C.2 for details), presented in Appendices C.3 and C.4, show that the key mechanism required to guarantee convergence under the BG-0 oracle is a bounded expected drift of the form $\mathbb{E}_t[\|x_{t+1} - x^*\|^2] \leq \left(1 - \frac{\beta_t}{2}\right)\|x_t - x^*\|^2 + \mathcal{E}_t$ for some residual $\mathcal{E}_t$.

Under σ^2-bounded variance, the residual $\mathcal{E}_t$ is bounded by a global constant controlled by η and the variance σ^2, enabling convergence. However, the BG-0 oracle permits the variance to scale quadratically with the trajectory distance. If the residual error $\mathcal{E}_t$ retains the unbounded term $\|x_t - x_0\|^2$, convergence cannot be established. Therefore, in the following result (proof in Appendix C.5), we establish conditions to obtain a bounded residual under the BG-0 oracle.

Theorem 4 (Geometric-Statistical Uncertainty Principle)**.** *Let Assumptions 1 and 2 hold. Furthermore, let* f *be convex (See Assumption 6 and set* $\rho = 0$*). Consider the sequence* $\{x_t\}$ *generated by* $x_{t+1} = \beta_t x_0 + (1 - \beta_t)x_t - \eta g_t$*, where* g_t *evaluates* $N_t \geq 1$ *independent stochastic queries (i.e., a PASTA instance with* $S = 1$ *and* $p_t = 1$*). Then, to ensure*

$$\mathbb{E}_t\|x_{t+1} - x^*\|^2 \leq \left(1 - \frac{\beta_t}{2}\right)\|x_t - x^*\|^2 + \mathcal{E}_t, \tag{5}$$

where the residual $\mathcal{E}_t = \beta_t\|x_0 - x^*\|^2 + \frac{\eta^2 G^2}{N_t}$ *is independent of the unbounded term* $\|x_t - x_0\|^2$*, the following uncertainty principle must hold*

$$\beta_t(1 - 2\beta_t)N_t \geq \eta^2 B^2, \qquad \eta \leq \frac{\sqrt{\beta_t}}{2L}. \tag{6}$$

The relation $\beta_t(1 - 2\beta_t)N_t \geq \eta^2 B^2$ is to be interpreted as an uncertainty principle. The left side represents PASTA's stabilization capacity, determined by the geometric parameter $\beta_t \ll 1$ and the statistical parameter N_t. The right side quantifies the normalized variance growth over a single step. The lower bound indicates that eliminating geometric anchoring, i.e., $\beta_t \to 0$, necessitates an unbounded expansion in statistical sampling, i.e., $N_t \to \infty$, to counteract the extra noise permitted by the BG-0 oracle (corresponding to the PASTA instantiation explored here). Conversely, using $N_t = 1$ requires setting $\beta_t \sim \eta^2 B^2$ to counteract the extra noise permitted by the BG-0 oracle (corresponding to the PASTA instantiation explored in Alacaoglu et al. (2025)). Therefore, the uncertainty principle rules out the simplest instance, i.e., SGD with $\beta_t \to 0$ and $N_t = 1$, simultaneously. Interestingly, the scaling $\beta_t \sim \eta^2$ when $\eta \sim 1/\sqrt{t}$ matches the classical Halpern iteration anchoring (Halpern, 1967).

4.3 Optimal complexity under mean-square smoothness

When the stochastic oracle additionally satisfies the MSS condition, the gradients exhibit sufficient correlation to employ recursive variance reduction. Under standard bounded variance, techniques

such as PAGE drop the oracle complexity from $\mathcal{O}(\epsilon^{-4})$ to $\mathcal{O}(\epsilon^{-3})$. However, as dictated by our lower bound in Theorem 2, the best we can expect to do with variance reduction under the BG-0 oracle is $\mathcal{O}(\epsilon^{-4})$. To achieve this limit, we instantiate PASTA with $p_t < 1$. This choice is essential, as demonstrated by the following key result from Li et al. (2021), to realize variance reduction.

Lemma 1 (PAGE Estimation Error (Li et al., 2021)). *Let the estimator g_t be generated by (3), and define the estimation error $V_t = \|g_t - \nabla f(x_t)\|^2$. Under Assumption 3, it holds that*

$$\mathbb{E}_t[V_t] \leq p_t \mathbb{E}_t \left[\left\| \frac{1}{N_t} \sum_{i=1}^{N_t} \widetilde{\nabla} f(x_t, \xi_t^{(i)}) - \nabla f(x_t) \right\|^2 \right] + (1-p_t) \left(V_{t-1} + \frac{\bar{L}^2}{b_t} \|x_t - x_{t-1}\|^2 \right). \quad (7)$$

The primary challenge under the BG-0 oracle is that recursive estimators bound variance using the distance between consecutive iterates, while the BG-0 noise scales with the absolute distance to the origin. We thus ensure PASTA utilizes dynamic large batches, an idea akin to what was explored in Theorem 3, leading to the following result, the proof of which is in Appendix C.6, where one can find the exact expected oracle complexity in (104).

Theorem 5 (Convergence of PASTA under MSS). *Let Assumptions 1 and 3 hold. Let the sequence $\{x_t\}$ be generated by PASTA with $\beta_t = 0$, $S = 1$. Set*

$$N_{ref} = \max\left\{1, \left\lceil \frac{8B^2\Delta^2}{\epsilon^4} + \frac{2G^2}{\epsilon^2} \right\rceil \right\},\ b = \left\lceil \sqrt{N_{ref}} \right\rceil,\ p = \frac{b}{N_{ref}},\ \eta = \frac{1}{4\bar{L}},\ K = \left\lceil \frac{16\Delta\bar{L}}{\epsilon^2} + \frac{1}{p} \right\rceil, \quad (8)$$

where $f(x_0) - \inf_x f(x) \leq \Delta$. Set batch size N_0 and set batch size N_t for $t \geq 1$ as

$$N_0 = \max\left\{1, \lceil 2G^2\epsilon^{-2} \rceil\right\}, \quad N_t = \max\left\{ N_{ref}, \left\lceil \frac{2(B^2\|x_t - x_0\|^2 + G^2)}{\epsilon^2} \right\rceil \right\}. \quad (9)$$

Then, $\frac{1}{K}\sum_{t=0}^{K-1} \mathbb{E}[\|\nabla f(x_t)\|^2] \leq \epsilon^2$ and the expected total stochastic oracle complexity is $\mathcal{O}\left(\epsilon^{-4}\right)$.

4.4 Optimal complexity under global curvature

We now shift our focus to optimization landscapes characterized by a degree of global curvature toward the optimal set, specifically those satisfying the Polyak-Lojasiewicz (PL) condition and star-convexity formalized next. We note that given the additional structure, $\mathbb{E}[f(x) - f(x^*)] \leq \epsilon$ serves as our natural convergence criterion.

Assumption 4 (Polyak-Lojasiewicz Condition). *A function $f : \mathbb{R}^p \to \mathbb{R}$ satisfies the Polyak-Lojasiewicz (PL) condition with constant $\mu > 0$, if $\|\nabla f(x)\|^2 \geq 2\mu(f(x) - \min f)$ for all x.*

Assumption 5 (Star-Convexity). *A function $f : \mathbb{R}^p \to \mathbb{R}$ is μ-star-convex on $\mathbb{R}^p$ with coefficient $\mu > 0$, if $\langle \nabla f(x), x - x^* \rangle \geq f(x) - f(x^*) + \frac{\mu}{2}\|x - x^*\|^2$, for some global minimizer x^* and for all x.*

In these regimes, the extra variance allowed by the BG-0 oracle does not degrade the asymptotic convergence rates. The intuition underlying this recovery relies on the Quadratic Growth (QG) property. Both the PL condition and star-convexity imply QG (Karimi et al., 2016) (with star-convexity also implying PL), which posits $\|x - x^*\|^2 \leq \frac{2}{\mu}(f(x) - f(x^*))$. Expanding the variance, we can write $\|x_t - x_0\|^2 \leq 2\|x_t - x^*\|^2 + 2\|x_0 - x^*\|^2$. By invoking QG, we map this variance directly back to the functional suboptimality. Consequently, the extra noise injected by the oracle is upper-bounded by a combination of the current optimization error and the initialization error. Because the gradient signal in PL and star-convex landscapes provides a linear contraction, it naturally absorbs the multiplicative variance amplification all by itself, leading to the following results.

Theorem 6 (Convergence of PASTA under PL Condition). *Let Assumptions 1, 2, and 4 hold. Let the sequence $\{x_t\}$ be generated by PASTA with $N_t = 1$, $S = 1$, $p_t = 1$, and $\beta_t = 0$. Set*

$$\eta = \min\left\{ \frac{1}{L}, \frac{\mu\epsilon}{3LG^2}, \frac{\mu^2\epsilon}{12LB^2\Delta} \right\}, \quad K = \left\lceil \frac{2}{\eta\mu} \left[\log\left(\frac{3\Delta}{\epsilon} \right) \right]_+ \right\rceil, \text{ where } [a]_+ := \max\{a, 0\} \quad (10)$$

where $\Delta \geq f(x_0) - \inf_x f(x)$. Then, $\mathbb{E}[f(x_K) - f(x^*)] \leq \epsilon$, evaluating to a total stochastic oracle complexity of $\mathcal{O}(\epsilon^{-1}\log(\epsilon^{-1}))$.

The proof is in Appendix C.7, with the exact oracle complexity in (113). We now establish an analogous result for star-convexity directly. While star-convexity also implies QG, the descent analysis operates directly on the sequence $\mathbb{E}[\|x_t - x^*\|^2]$ rather than the functional sequence $\mathbb{E}[f(x_t) - f(x^*)]$, the proof of which is in Appendix C.8, where the exact oracle complexity can be found in (121).

Theorem 7 (Convergence of PASTA under Star-Convexity)**.** *Let Assumptions 1, 2, and 5 hold. Let the sequence $\{x_t\}$ be generated by PASTA with $N_t = 1$, $S = 1$, $p_t = 1$, and $\beta_t = 0$. Set*

$$\eta = \min\left\{\frac{1}{L}, \frac{\mu}{4B^2}, \frac{\mu^2\epsilon}{12LB^2\Delta}, \frac{\mu\epsilon}{3LG^2}\right\}, \qquad K = \left\lceil \frac{2}{\eta\mu}\left[\log\left(\frac{3L\Delta}{\mu\epsilon}\right)\right]_+\right\rceil, \tag{11}$$

where $\Delta \geq f(x_0) - \inf_x f(x)$. Then, $\mathbb{E}[f(x_K) - f(x^*)] \leq \epsilon$, evaluating to a total stochastic oracle complexity of $\mathcal{O}(\epsilon^{-1}\log(\epsilon^{-1}))$.

4.5 Optimal complexity under weak convexity

We now turn to the standard weakly convex setting, following Davis & Drusvyatskiy (2019), and state the assumptions used in our analysis.

Assumption 6 (Weak Convexity)**.** *A function $f : \mathbb{R}^p \to \mathbb{R}$ is ρ-weakly convex for some constant $\rho \geq 0$ if $f(x) + \frac{\rho}{2}\|x\|^2$ is convex. Note that when $\rho = 0$, the function is convex in the typical sense. Also, any non-convex smooth function is L-weakly convex.*

Assumption 7 (Lipschitz Continuity)**.** *A function $f : \mathbb{R}^p \to \mathbb{R}$ is C-Lipschitz continuous if $\|g\| \leq C$ for all x and $g \in \partial f(x)$.*

Unlike the PL or star-convex cases, weakly convex functions, which generalize non-convex L-smooth functions, lack global curvature pointing toward the optimal set. Consequently, the unanchored SGD approach utilized in the preceding sections does not provide the contraction needed for our analysis. We thus use PASTA in its epoch-based form with a nonzero $\beta_t = \lambda\eta_t$. As discussed in Section 3, because f is ρ-weakly convex, selecting the coupling parameter $\lambda > \rho$ ensures that the surrogate $F_\lambda(x) = f(x) + \frac{\lambda}{2}\|x - x_{s-1}\|^2$ is μ-strongly convex with coefficient $\mu = \lambda - \rho > 0$, a property akin to the curvature provided by the PL condition or star-convexity. However, minimizing the surrogate $F_\lambda(x)$ only yields the minimizer of the regularized subproblem, resulting in a biased stationary point of the true objective f. To remove this bias, PASTA employs an epoch-based structure. By running the inner loop for K iterations and subsequently updating the anchor $x_s = x_K$, the outer loop effectively implements an inexact Stochastic Proximal Point algorithm.

For weakly convex functions, we employ the Moreau envelope as a continuous surrogate to quantify stationarity. We recall some definitions from Davis & Drusvyatskiy (2019).

Definition 1 (Proximal Operator)**.** *For a proper, lower semicontinuous function $f : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ and a parameter $\lambda > 0$, the proximal mapping is defined by*

$$\operatorname{prox}_{f/\lambda}(x) = \operatorname*{argmin}_{y \in \mathbb{R}^p}\left\{f(y) + \frac{\lambda}{2}\|y - x\|^2\right\}. \tag{12}$$

Definition 2 (Moreau Envelope)**.** *Associated with the proximal operator, the Moreau envelope of f with parameter $\lambda > 0$ is defined as*

$$\varphi_{1/\lambda}(x) = \min_{y \in \mathbb{R}^p}\left\{f(y) + \frac{\lambda}{2}\|y - x\|^2\right\}. \tag{13}$$

The Moreau envelope acts as a regularized, smoothed approximation of f. When f is ρ-weakly convex and $\lambda > \rho$, the objective of the proximal mapping is strongly convex, ensuring that $\operatorname{prox}_{f/\lambda}(x)$ is unique and well-defined everywhere. The following lemma connects the gradient of the Moreau envelope to the proximal step.

Lemma 2. *Let f be a ρ-weakly convex function and $\lambda > \rho$. Then, $\varphi_{1/\lambda}(x)$ is smooth with constant $L_\varphi = \max\{\lambda, \frac{\lambda\rho}{\lambda - \rho}\}$, and its gradient is given by $\nabla\varphi_{1/\lambda}(x) = \lambda\big(x - \operatorname{prox}_{f/\lambda}(x)\big)$. Furthermore, $\|\nabla\varphi_{1/\lambda}(x)\|$ serves as a valid surrogate for stationarity. If $\|\nabla\varphi_{1/\lambda}(x)\| \leq \epsilon$, then there exists a point $\hat{x} = \operatorname{prox}_{f/\lambda}(x)$ such that $\|x - \hat{x}\| \leq \lambda^{-1}\epsilon$ and $\operatorname{dist}(0, \partial f(\hat{x})) \leq \epsilon$. Finally, it holds that $\varphi_{1/\lambda}(x_0) - \min_x \varphi_{1/\lambda}(x) \leq f(x_0) - \inf_x f(x)$ for all x_0 and $\lambda > 0$.*

Note that since any non-convex and smooth function is weakly convex, the lower bound established in Theorem 1 is also applicable here, and the best complexity one could hope for is $\mathcal{O}(\epsilon^{-6})$.

Having defined our notation of stationarity, the following theorem formalizes the optimal $\mathcal{O}(\epsilon^{-6})$ complexity of PASTA, the proof of which is in Appendix C.9.

Theorem 8 (Convergence of PASTA under Weak Convexity)**.** *Let Assumptions 1, 6, and 7 hold. Let the sequence $\{x_s\}$ be generated by PASTA running for S epochs of length K. Assume the coupling parameter satisfies $\lambda > \rho$ with strong convexity coefficient $\mu = \lambda - \rho > 0$, and define the anchoring sequence as $\beta_t = \lambda\eta$. Set the step size η, epoch count S, and epoch length K as*

$$\eta = \min\left\{\frac{\mu^3}{96\lambda^2(3\lambda^2-\rho^2)}, \frac{\mu^3\epsilon^2}{256C^2\lambda^2(3\lambda^2-\rho^2)}\right\},\ S = \left\lceil\frac{32\lambda^2\Delta}{\mu\epsilon^2}\right\rceil,\ K = \left\lceil 1 + \frac{1}{\eta\mu}\log\left(\frac{12(3\lambda^2-\rho^2)}{\mu^2}\right)\right\rceil. \tag{14}$$

Set the initial batch size N_0 at $t = 0$ and the constant inner loop batch size $N_t = N$ for $t \geq 1$ as

$$N_0 = \left\lceil \max\left\{\left(\frac{4}{3} + \frac{2\mu^2}{3(3\lambda^2-\rho^2)}\right)B^2\eta^2S^2, \frac{16G^2\eta^2}{3\lambda^2\epsilon^2}\right\}\right\rceil \tag{15}$$

and

$$N = \left\lceil \max\left\{\left(\frac{96(3\lambda^2-\rho^2)}{\mu^3} + \frac{48}{\mu}\right)B^2\eta S^2, \frac{128G^2\eta(3\lambda^2-\rho^2)}{\mu^3\lambda^2\epsilon^2}, \frac{72B^2\eta(3\lambda^2-\rho^2)}{\mu^3}\right\}\right\rceil, \tag{16}$$

where $\Delta \geq f(x_0) - \inf_x f(x)$. Then, $\frac{1}{S}\sum_{s=1}^{S}\mathbb{E}[\|\nabla\varphi_{1/\lambda}(x_{s-1})\|^2] \leq \epsilon^2$ and the total stochastic oracle complexity evaluates to $S(N_0 + KN) = \mathcal{O}(\epsilon^{-6})$.

Remark 1. *Traditional Halpern iteration relies on an uncoupled, decaying anchoring sequence (e.g., $\beta_t = \frac{1}{t+2}$). Analyzing such uncoupled dynamics yields harmonic summations that introduce a logarithmic dependence on ϵ^{-1} (see Theorem 3.6 in Alacaoglu et al. (2025) for an example). By coupling $\beta_t = \lambda\eta$ and holding η constant, the inner loop contracts linearly at a rate of $(1-\eta\mu)$ without accumulating logarithmic artifacts. Further, since the subproblem is μ-strongly convex, the distance to the exact subproblem minimizer decays exponentially. The inner loop length $K = \mathcal{O}(\epsilon^{-2})$ is selected to reduce the optimization error of the surrogate until it is dominated by the statistical floor. Finally, to deal with the extra variance term $B^2\|x_t - x_0\|^2$, the batch sizes N_0 and N scale with S^2. Intuitively, this makes sense as the maximum distance the iterates can traverse over S epochs scales with S; hence, ensuring the extra variance term $B^2\|x_t - x_0\|^2$ remains bounded by a global constant requires batch sizes of $\mathcal{O}(S^2) = \mathcal{O}(\epsilon^{-4})$.*

Remark 2. *By definition, any non-convex L-smooth function is L-weakly convex. To derive the convergence guarantees for the L-smooth functions using Theorem 8, we set $\rho = L$ and $\lambda = 2L$, which means $\mu = \lambda - \rho = L$ for the regularized surrogate. The proximal mapping $\hat{x} = \mathrm{prox}_{f/\lambda}(x)$ implies $\nabla f(\hat{x}) + \lambda(\hat{x} - x) = 0$, which, by Lemma 2, means $\nabla\varphi_{1/\lambda}(x) = \lambda(x - \hat{x}) = \nabla f(\hat{x})$. Using the L-smoothness of f and the triangle inequality we have $\|\nabla f(x)\| \leq \|\nabla f(x) - \nabla f(\hat{x})\| + \|\nabla f(\hat{x})\| \leq L\|x - \hat{x}\| + \|\nabla\varphi_{1/\lambda}(x)\|$. Thus, $\|\nabla f(x)\| \leq \frac{3}{2}\|\nabla\varphi_{1/2L}(x)\|$. Consequently, using $\tilde{\epsilon}^2 = \frac{4}{9}\epsilon^2$, Theorem 8 implies the same $\mathcal{O}(\epsilon^{-6})$ for smooth functions, which we established in Theorem 3 using dynamic batching, but here using an epoch-based instantiation of PASTA.*

5 Conclusion

We resolved the oracle complexity of non-convex minimization under the unbounded variance characterized by the BG-0 condition, recently shown to be the weakest variance condition. By establishing $\Omega(\epsilon^{-6})$ and $\Omega(\epsilon^{-4})$ lower bounds for smooth and mean-square smooth functions, we formally quantified the information-theoretic penalty incurred when going beyond (effectively) the bounded variance assumption. To bridge this gap, we studied problem-dependent PASTA instantiations that control the distance-dependent noise through suitable choices of batching, gradient estimation, anchoring, and epoch structure. In the proximally anchored instantiation, carefully coupling the anchoring parameter and step size induces curvature through Tikhonov regularization that counteracts the unbounded noise allowed by BG-0 oracles. Our analyses show that suitable PASTA instantiations attain minimax optimal rates for smooth, weakly convex, star-convex, and Polyak-Lojasiewicz functions without relying on bounded domain assumptions.

AI USE STATEMENT

We used AI tools to assist with drafting sections of the paper, refining the writing, and checking proofs. The authors reviewed and edited the resulting text and take full responsibility for the paper's content and correctness.

ETHICS STATEMENT

This work is theoretical and does not involve human subjects, personal data, proprietary data, or experiments on deployed systems. We do not introduce or release datasets, and we do not identify any direct ethical risks arising from the results.

REPRODUCIBILITY STATEMENT

All assumptions and convergence criteria are stated explicitly in the main text. Complete proofs of the lower and upper bounds, including the hard-instance constructions, auxiliary lemmas, parameter choices, and oracle-complexity calculations, are provided in the appendices. The PASTA framework and the parameter settings used for each problem class are also specified in the main text and appendices. The paper is theoretical and contains no empirical results requiring datasets or experimental code.

REFERENCES

Alekh Agarwal, Peter L Bartlett, Pradeep Ravikumar, and Martin J Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

Ahmet Alacaoglu, Yura Malitsky, and Stephen J Wright. Towards weaker variance assumptions for stochastic optimization. *arXiv preprint arXiv:2504.09951*, 2025.

Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pp. 699–707, 2016.

Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding local minima via first-order oracles. In *Advances in Neural Information Processing Systems*, pp. 3716–3726, 2018.

Yossi Arjevani and Ohad Shamir. Dimension-free iteration complexity of finite sum optimization problems. In *Advances in Neural Information Processing Systems*, pp. 3540–3548, 2016.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1): 165–214, 2023.

Hilal Asi and John C Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

Francis Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Julius R Blum. Approximation methods which converge with probability one. *The Annals of Mathematical Statistics*, pp. 382–386, 1954.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale learning. *SIAM Review*, 60(2):223–311, 2018.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 177:193–230, 2019.

Yiyue Chen, Abolfazl Hashemi, and Haris Vikalo. Communication-efficient variance-reduced decentralized stochastic optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 67(12):6583–6594, 2021.

Shisheng Cui and Uday V Shanbhag. On the analysis of variance-reduced and randomized projection variants of single projection schemes for monotone stochastic variational inequality problems. *Set-Valued and Variational Analysis*, 29:453–499, 2021.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, 2019.

Rudrajit Das, Anish Acharya, Abolfazl Hashemi, Sujay Sanghavi, Inderjit S Dhillon, and Ufuk Topcu. Faster non-convex federated learning via global and local momentum. In *Uncertainty in Artificial Intelligence*, pp. 496–506. PMLR, 2022.

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

Jelena Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *Conference on Learning Theory*, pp. 1428–1451. PMLR, 2020.

Justin Domke, Robert Gower, and Guillaume Garrigos. Provable convergence guarantees for black-box variational inference. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66289–66327, 2023.

Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. *arXiv preprint arXiv:1910.01845*, 2019.

Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1):503–558, 2019.

John C Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

Cong Fang, Chris Jiao Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pp. 689–699, 2018.

Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex SGD escaping from saddle points. In *Conference on Learning Theory*, pp. 1192–1234. PMLR, 2019.

Dylan J Foster, Ayush Sekhari, Ohad Shamir, Nathan Srebro, Karthik Sridharan, and Blake Woodworth. The complexity of making the gradient small in stochastic convex optimization. In *Conference on Learning Theory*, pp. 1319–1345. PMLR, 2019.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points: online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pp. 797–842, 2015.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

EG Gladyshev. On stochastic approximation. *Theory of Probability & Its Applications*, 10(2): 275–278, 1965.

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 680–690. PMLR, 2020.

Eduard Gorbunov, Abdurakhmon Sadiev, Marina Danilova, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability convergence for composite and distributed stochastic minimization and variational inequalities with heavy-tailed noise. *arXiv preprint arXiv:2310.01860*, 2023.

Benjamin Grimmer. Convergence rates for deterministic and stochastic subgradient methods without Lipschitz continuity. *SIAM Journal on Optimization*, 29(2):1350–1365, 2019.

Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning*, pp. 3964–3975. PMLR, 2021.

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

Abolfazl Hashemi. A unified model for large-scale inexact fixed-point iteration: A stochastic optimization perspective. *IEEE Transactions on Automatic Control*, 70(4):2435–2449, 2024.

Yankun Huang and Qihang Lin. Oracle complexity of single-loop switching subgradient methods for non-smooth weakly convex functional constrained optimization. *Advances in Neural Information Processing Systems*, 36:61327–61340, 2023.

Sasila Ilandarideva, Anatoli Juditsky, Guanghui Lan, and Tianjiao Li. Accelerated stochastic approximation with state-dependent noise. *arXiv preprint arXiv:2307.01497*, 2023.

Andrew Jacobsen and Ashok Cutkosky. Unconstrained online learning with unbounded losses. In *International Conference on Machine Learning*, pp. 14590–14630. PMLR, 2023.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of machine learning research*, 18(223):1–42, 2018.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International conference on machine learning*, pp. 1724–1732. PMLR, 2017.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811. Springer, 2016.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023.

Ahmed Khaled, Othmane Sebbouh, Nicolas Loizou, Robert M Gower, and Peter Richtárik. Unified analysis of stochastic gradient methods for composite convex and smooth optimization. *Journal of Optimization Theory and Applications*, 199(2):499–540, 2023.

Mark Aleksandrovich Krasnosel'skiĭ. Two remarks on the method of successive approximations. *Uspekhi matematicheskikh nauk*, 10(1):123–127, 1955.

Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pp. 38–53, 1973.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *arXiv preprint arXiv:1602.04915*, 2016.

Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via SCSG methods. In *Advances in Neural Information Processing Systems*, pp. 2348–2358, 2017.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International conference on machine learning*, pp. 6286–6295. PMLR, 2021.

Wei Liu and Yangyang Xu. A single-loop spider-type stochastic subgradient method for expectation-constrained nonconvex nonsmooth optimization. *arXiv preprint arXiv:2501.19214*, 2025.

Vien Mai and Mikael Johansson. Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *International conference on machine learning*, pp. 6630–6639. PMLR, 2020.

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

Arkadi Nemirovski and David B Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2004.

Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Gergely Neu and Nneka Okolo. Dealing with unbounded gradients in stochastic saddle-point optimization. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 37508–37530, 2024.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International conference on machine learning*, pp. 2613–2621. PMLR, 2017.

Ta Duy Nguyen, Alina Ene, and Huy L Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tails. *arXiv preprint arXiv:2304.01119*, 2023.

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Maxim Raginsky and Alexander Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *IEEE Transactions on Information Theory*, 57(10):7036–7056, 2011.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1571–1578, 2012.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, pp. 314–323, 2016.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pp. 71–79. PMLR, 2013.

Matus Telgarsky. Stochastic linear optimization never overfits with quadratically-bounded losses on general data. In *Conference on Learning Theory*, pp. 5453–5488. PMLR, 2022.

Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. In *Advances in Neural Information Processing Systems*, pp. 2899–2908, 2018.

Antesh Upadhyay, Arda Fazla, and Abolfazl Hashemi. Beyond bounded variance: Variance-reduced normalized methods for nonconvex optimization under blum-gladyshev noise. *arXiv preprint arXiv:2605.15314*, 2026a.

Antesh Upadhyay, Sang Bin Moon, and Abolfazl Hashemi. Fedsgm: A unified framework for constraint aware, bidirectionally compressed, multi-step federated optimization. *arXiv preprint arXiv:2601.16897*, 2026b.

Mengdi Wang and Dimitri P Bertsekas. Stochastic first-order methods with random constraint projection. *SIAM Journal on Optimization*, 26(1):681–717, 2016.

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv preprint arXiv:1810.10690*, 2018.

Blake Woodworth and Nathan Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems*, pp. 3639–3647, 2016.

Yi Xu, Jin Rong, and Tianbao Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. In *Advances in Neural Information Processing Systems*, pp. 5530–5540, 2018.

TaeHo Yoon and Ernest K Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $O(1/k^2)$ rate on squared gradient norm. In *International Conference on Machine Learning*, pp. 12098–12109. PMLR, 2021.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pp. 423–435. Springer, 1997.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. *Journal of Machine Learning Research*, 21(103):1–63, 2020.

Appendix Table of Contents

A Related Work **15**

B Proofs of Lower Bounds in Section 2 **16**

B.1 Proof of Theorem 1 18

B.2 Proof of Theorem 2 20

C Proofs of Upper Bounds in Section 4 **21**

C.1 Proof of Theorem 3 21

C.2 Results under Convexity 23

C.3 Proof of Theorem 9 23

C.4 Proof of Theorem 10 25

C.5 Proof of Theorem 4 26

C.6 Proof of Theorem 5 26

C.7 Proof of Theorem 6 28

C.8 Proof of Theorem 7 28

C.9 Proof of Theorem 8 29

A Related Work

Variance assumptions. Classical analyses of SGD assume uniformly bounded stochastic subgradients (Nemirovski et al., 2009; Bach & Moulines, 2011; Rakhlin et al., 2012; Shamir & Zhang, 2013), or globally bounded variance (Bottou et al., 2018; Jain et al., 2018). To mitigate this, several relaxations have been proposed. The Expected Smoothness or ABC condition (Gorbunov et al., 2020; Khaled & Richtárik, 2023; Khaled et al., 2023; Ilandarideva et al., 2023; Grimmer, 2019) bounds the variance by a constant plus a term proportional to the gradient norm and suboptimality. While less restrictive, Alacaoglu et al. (2025) demonstrates that the BG-0 condition is weaker and more universally applicable. The origins of the BG-0 condition trace back to the foundational asymptotic analyses of Robbins & Monro (1951); Blum (1954); Gladyshev (1965), and it has seen a recent resurgence in both asymptotic (Wang & Bertsekas, 2016; Cui & Shanbhag, 2021) and non-asymptotic contexts (Asi & Duchi, 2019; Jacobsen & Cutkosky, 2023; Telgarsky, 2022; Domke et al., 2023; Neu & Okolo, 2024). Other *orthogonal, complementary relaxations* include heavy-tailed noise settings (Gorbunov et al., 2023; Nguyen et al., 2023; Gurbuzbalaban et al., 2021), which we do not consider here.

Information-theoretic lower bounds. In the noiseless convex setting, Nemirovski & Yudin (1983) and Nesterov (2004) established the $\Omega(\sqrt{1/\epsilon})$ bounds, while Carmon et al. (2019) proved the $\Omega(\epsilon^{-2})$ bound for smooth non-convex functions. For stochastic convex optimization, a long line of literature has mapped the dimension-dependent and independent boundaries via statistical estimation reductions (Raginsky & Rakhlin, 2011; Agarwal et al., 2012; Woodworth & Srebro, 2016; Foster et al., 2019).

In the stochastic non-convex regime, Arjevani et al. (2023) proved that under the strictly bounded variance assumption, any randomized algorithm requires $\Omega(\epsilon^{-4})$ queries to find an ϵ-stationary point. Furthermore, under the addition of mean-square smoothness, Arjevani et al. (2023) establish an accelerated $\Omega(\epsilon^{-3})$ lower bound. Earlier algorithm-specific bounds include Drori & Shamir (2019), while finite-sum lower bounds were explored in Arjevani & Shamir (2016); Fang et al. (2018). Our lower bounds conceptually depart from Arjevani et al. (2023) by leveraging the unbounded nature of the BG-0 oracle. Because the BG-0 variance explodes in remote regions of the landscape, the zero-chain constructions used in Arjevani et al. (2023) lead to signal-to-noise degradation. By formally

embedding a coercive traversal manifold into the zero-chain, we prove that the bounds of Arjevani et al. (2023) degrade under the BG-0 condition, establishing strictly worse $\Omega(\epsilon^{-6})$ and $\Omega(\epsilon^{-4})$ limits.

Stochastic non-convex and weakly convex minimization. The study of SGD for non-convex functions gained modern prominence with the work of Ghadimi & Lan (2013), who established the optimal $\mathcal{O}(\epsilon^{-4})$ oracle complexity for finding ϵ-stationary points of smooth functions. Subsequent research heavily explored the capacity of SGD and its perturbed variants to escape saddle points and find local minima (Ge et al., 2015; Jin et al., 2017; Lee et al., 2016; Duchi & Ruan, 2018).

Beyond standard smoothness, *weakly convex* stochastic optimization, which encompasses a broader class of non-smooth, non-convex problems prevalent in robust statistics and deep learning, has drawn attention. The seminal work of Davis & Drusvyatskiy (2019) developed the stochastic model-based framework, establishing $\mathcal{O}(\epsilon^{-4})$ rates for the stochastic proximal point and prox-linear methods by tracking the stationarity of the Moreau envelope, a natural convergence criterion for the optimization of weakly convex functions. This paved the way for extensive studies into proximal, compositional, constrained, and distributed non-convex methods (Duchi & Ruan, 2018; Drusvyatskiy & Paquette, 2019; Mai & Johansson, 2020; Huang & Lin, 2023; Liu & Xu, 2025; Upadhyay et al., 2026b). However, the theoretical guarantees in these works implicitly or explicitly rely on globally-bounded variance or bounded domain assumptions, which we relax in this paper.

Variance reduction and high-order methods. Large effort has been devoted to surpassing the standard $\mathcal{O}(\epsilon^{-4})$ SGD rate using variance reduction or higher-order information. In the variance reduction category, algorithms based on SVRG, SAGA, SPIDER, SNVRG, and PAGE leverage mean-square smoothness to attain the optimal $\mathcal{O}(\epsilon^{-3})$ rate in various settings including centralized and distributed optimization and fixed-point calculation (Allen-Zhu & Hazan, 2016; Reddi et al., 2016; Lei et al., 2017; Fang et al., 2018; Zhou et al., 2020; Wang et al., 2018; Cutkosky & Orabona, 2019; Nguyen et al., 2017; Li et al., 2021; Chen et al., 2021; Das et al., 2022; Hashemi, 2024). Alternatively, high-order methods utilize Lipschitz continuous Hessians to achieve rates of $\mathcal{O}(\epsilon^{-3.5})$ or better via cubic regularization or perturbed acceleration (Xu et al., 2018; Allen-Zhu & Li, 2018; Tripuraneni et al., 2018; Fang et al., 2019). Crucially, *all* of these accelerated rates depend on the uniform bounded-variance assumption. Our work extends variance reduction mechanics to the BG-0 regime for stochastic non-convex optimization under mean-square smoothness, proving that while acceleration is possible (from $\mathcal{O}(\epsilon^{-6})$ to $\mathcal{O}(\epsilon^{-4})$), the unbounded noise inherently prevents attaining the $\mathcal{O}(\epsilon^{-3})$ rates under bounded variance and mean-square smoothness. Recent work by Upadhyay et al. (2026a) develops complementary normalized momentum and variance-reduced methods under BG-0, attaining $\mathcal{O}(\epsilon^{-6})$ complexity under smoothness and $\mathcal{O}(\epsilon^{-4})$ under MSS with sharp initialization, matching the lower-bound orders established here, further demonstrating the tightness of the lower-bound orders established here.

Halpern anchoring and fixed-point dynamics. As discussed, PASTA is connected to fixed-point methods. Finding a stationary point maps naturally to finding a fixed point of a non-expansive mapping. Classical approaches include the Krasnoselskii-Mann iteration (Krasnosel'skiĭ, 1955; Mann, 1953), which lacks strong convergence guarantees, and the Halpern iteration (Halpern, 1967), which anchors the sequence to a fixed initial point to force strong convergence. Recently, Halpern anchoring has been leveraged in convex and min-max optimization to achieve accelerated rates (Diakonikolas, 2020; Yoon & Ryu, 2021). Neu & Okolo (2024) and Alacaoglu et al. (2025) demonstrated that Halpern-like updates can counteract the divergence of BG-0 noise in convex and bilinear min-max settings. However, traditional Halpern tracking introduces logarithmic convergence penalties due to uncoupled parameter decay. PASTA, on the other hand, couples the step size and Halpern anchoring parameter via $\beta_t = \lambda\eta_t$. This seamlessly transforms the geometric anchoring into an exact stochastic proximal step on a Tikhonov-regularized surrogate, providing the requisite strong convexity to mitigate the extra variance term in general non-convex and weakly convex landscapes without logarithmic decay artifacts.

B Proofs of Lower Bounds in Section 2

Before presenting the proofs of our lower bounds, we recall the framework of Arjevani et al. (2023).

Definition 3 (Progress Function)**.** *For any vector $y \in \mathbb{R}^T$ and a threshold $\alpha \geq 0$, the progress function* $\text{prog}_\alpha(y)$ *identifies the highest coordinate index whose absolute magnitude exceeds the*

threshold. Formally, defining a dummy variable $y_0 \equiv \infty$, we have

$$\text{prog}_\alpha(y) := \max\{i \in \{0, 1, \ldots, T\} : |y_i| > \alpha\}. \tag{17}$$

Definition 4 (Zero-Respecting Algorithms)**.** *An optimization algorithm A is termed zero-respecting if it only explores the coordinate subspace spanned by the gradient feedback previously observed from the oracle. Formally, an algorithm interacting with an oracle O_F belongs to the class of zero-respecting algorithms, denoted $\mathcal{A}_{\text{zr}}$, if for any generated query sequence $x^{(1)}, x^{(2)}, \ldots$, the support satisfies*

$$\text{supp}(x^{(t)}) \subseteq \bigcup_{i=1}^{t-1} \text{supp}(\mathsf{O}_F(x^{(i)})) \quad \forall t \geq 1. \tag{18}$$

Following Arjevani et al. (2023), we choose the underlying deterministic function F to be a hard zero-chain construction. For each dimension $T \in \mathbb{N}$, we define

$$\bar{F}_T(x) := -\Psi(1)\Phi(x_1) + \sum_{i=2}^{T} \left[\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)\right], \tag{19}$$

where the non-linear component functions Ψ and Φ are constructed as

$$\Psi(t) = \begin{cases} 0, & t \leq 1/2, \\ \exp\left(1 - \frac{1}{(2t-1)^2}\right), & t > 1/2. \end{cases} \quad \text{and} \quad \Phi(t) = \sqrt{e}\int_{-\infty}^{t} e^{-\frac{1}{2}\eta^2} d\eta. \tag{20}$$

The function $\bar{F}_T$ is a deterministic zero-chain. It inherently maintains a large gradient magnitude unless all coordinates have been activated and optimized to be sufficiently large (i.e., $\text{prog}_1(x) \geq T$). We formally enumerate its relevant geometric properties below.

The bottleneck of any zero-respecting algorithm operating on a probability-p zero-chain is established by the expected hitting time to activate subsequent coordinates.

Lemma 3 (Lemma 1 in Arjevani et al. (2023))**.** *Let $g(x, z)$ be a probability-p zero-chain gradient estimator for $F : \mathbb{R}^T \to \mathbb{R}$, and let O be any oracle with $\mathsf{O}_F(x, z) = (F(x), g(x, z))$. Let $\{x_{\mathsf{A}}^{(k)}\}$ be the sequence of queries generated by any zero-respecting algorithm $\mathsf{A} \in \mathcal{A}_{\text{zr}}$ interacting with O_F. Then, with probability at least $1 - \delta$, the progress is bounded by $\max_{k\in[K]} \text{prog}_0(x_{\mathsf{A}}^{(k)}) < T$ for any query budget $K \leq \frac{T - \log(1/\delta)}{2p}$.*

Lemma 4 (Lemma 2 in Arjevani et al. (2023))**.** *The function $\bar{F}_T$ satisfies the following properties. (1) $\bar{F}_T(0) - \inf_x \bar{F}_T(x) \leq \Delta_0 \cdot T$, where $\Delta_0 = 12$. (2) The gradient of $\bar{F}_T$ is ℓ_1-Lipschitz continuous, where $\ell_1 = 152$. (3) For all $x \in \mathbb{R}^T$, $\|\nabla \bar{F}_T(x)\|_\infty \leq \gamma_\infty$, where $\gamma_\infty = 23$. (4) For all $x \in \mathbb{R}^T$, $\text{prog}_0(\nabla \bar{F}_T(x)) \leq \text{prog}_{1/2}(x) + 1$. (5) For all $x \in \mathbb{R}^T$, if $\text{prog}_1(x) < T$ then $\|\nabla \bar{F}_T(x)\| \geq |\nabla_{\text{prog}_1(x)+1} \bar{F}_T(x)| > 1$.*

To simulate the difficulty of stochastic optimization, the deterministic chain is evaluated using a stochastic oracle that masks the true gradient.

A simple extension to Lemma 4 utilizing the structure of $\bar{F}_T(x)$ in (19) to provide a total variation bound is as follows.

Remark 3 (Total Variation Bound)**.** *Let $\bar{F}_T$ be defined as in (19), then for all $x \in \mathbb{R}^T$, we have $|\bar{F}_T(x)| \leq e\sqrt{2\pi e}\, T < 12T$. Consequently, $\sup_x \bar{F}_T(x) \leq 12T$ and $\inf_x \bar{F}_T(x) \geq -12T$, and hence $\sup_x \bar{F}_T(x) - \inf_x \bar{F}_T(x) \leq 24T = 2\Delta_0 \cdot T$.*

Proof of Remark 3. Recall the definition of $\Psi(t)$ and $\Phi(t)$ from (20). It then immediately follows that $0 \leq \Psi(t) \leq e$ and $0 \leq \Phi(t) \leq \sqrt{e}\int_{-\infty}^{\infty} e^{-\eta^2/2}\, d\eta = \sqrt{2\pi e}$. Let $M := \sqrt{2\pi e}$. Now, we focus on $\bar{F}_T(x)$. We bound the first term first.

Since $\Psi(1) = 1$, we have

$$|-\Psi(1)\Phi(x_1)| = |\Phi(x_1)| \leq M.$$

Next, fix any $i \in \{2, \ldots, T\}$ and consider the summand $\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)$. Because $\Psi(t)$ is nonzero only when $t > \frac{1}{2}$, at most one of $\Psi(a)$ and $\Psi(-a)$ can be nonzero for any

$a \in \mathbb{R}$. Indeed, if both were nonzero, then one would simultaneously have $a > \frac{1}{2}$ and $-a > \frac{1}{2}$, which is impossible. Therefore, for each $i \geq 2$, at most one of the two terms in the summand is active. It follows that

$$|\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)| \leq eM.$$

Summing this all, we get $|\bar{F}_T(x)| \leq M + (T-1)eM \leq TeM \leq 12T$. The rest just follows as presented in the remark above. ■

Lemma 5 (Lemma 3 in Arjevani et al. (2023)). *The stochastic gradient estimator $\bar{g}_T$ is a probability-p zero-chain, is unbiased for $\nabla \bar{F}_T$, and has variance $\mathbb{E}[\|\bar{g}_T(x,z) - \nabla \bar{F}_T(x)\|^2] \leq \varsigma^2 \cdot \frac{1-p}{p}$ for all $x \in \mathbb{R}^T$, where $\varsigma = 23$.*

To establish limits under the MSS condition, we rely on an advanced zero-chain estimator that reduces the variance of gradient differences.

Lemma 6 (Lemma 4 in Arjevani et al. (2023)). *There exists a stochastic gradient estimator $\bar{g}_T(x,z)$ for $\bar{F}_T(x)$ that is a probability-p zero-chain, is unbiased for $\nabla \bar{F}_T$, and satisfies the variance bound $\mathbb{E}[\|\bar{g}_T(x,z) - \nabla \bar{F}_T(x)\|^2] \leq \varsigma^2 \frac{1-p}{p}$ with $\varsigma = 23$, and the mean-square smoothness property $\mathbb{E}[\|\bar{g}_T(x,z) - \bar{g}_T(y,z)\|^2] \leq \frac{\bar{\ell}_1^2}{p}\|x-y\|^2$ for all $x, y \in \mathbb{R}^T$, where $\bar{\ell}_1 = 328$.*

Finally, we state a result which is an extension of classical mean estimation LeCam (1973); Yu (1997) that we need for the MS lower bound.

Lemma 7 (Lemma 10 in Arjevani et al. (2023)). *There exists a number $c_0 > 0$ such that for any dimension T and $\epsilon \leq \sqrt{\bar{L}\Delta}$, extracting an ϵ-stationary mean vector from an oracle with variance σ^2 requires $\geq c_0 \sigma^2 \epsilon^{-2}$ for some function with initial suboptimality Δ and smoothness constant $\bar{L}$.*

B.1 Proof of Theorem 1

Proof of Theorem 1. Let $D = \frac{\Delta}{4\epsilon}$. We define a coordinate function $f_0 \colon \mathbb{R} \to \mathbb{R}$ satisfying

$$f_0(0) = 0, \qquad f_0'(u) = \begin{cases} -2\epsilon, & u \leq D \\ -2\epsilon + \frac{L}{2}(u - D), & u \in (D, D + \frac{4\epsilon}{L}) \\ 0, & u \geq D + \frac{4\epsilon}{L}. \end{cases} \tag{21}$$

Integrating this derivative establishes $\inf_u f_0(u) \geq -\frac{\Delta}{2} - \frac{4\epsilon^2}{L}$. We define a twice-differentiable activation function $\phi \colon \mathbb{R} \to [0,1]$ such that

$$\phi(u) = \begin{cases} 0 & u \leq D/2 \\ 1 & u \geq D \\ 10t^3 - 15t^4 + 6t^5 & D/2 \leq u \leq D, \end{cases} \qquad t = \frac{8\epsilon u}{\Delta} - 1. \tag{22}$$

Evaluating the global extrema over this polynomial yields the derivative bounds $\|\phi'\|_\infty = \frac{15\epsilon}{\Delta}$ and $\|\phi''\|_\infty = \frac{640\epsilon^2}{\sqrt{3}\Delta^2}$.

Following Lemma 4, we construct a scaled and globally shifted high-dimensional zero-chain $F_{\text{scaled}} \colon \mathbb{R}^T \to \mathbb{R}$

$$F_{\text{scaled}}(y) = \frac{L\lambda^2}{2\ell_1}\left(\bar{F}_T\left(\frac{y}{\lambda}\right) - \sup_{x \in \mathbb{R}^T} \bar{F}_T(x)\right), \quad \lambda = \frac{4\ell_1\epsilon}{L}, \quad T = \left\lfloor \frac{L\Delta}{768\ell_1\epsilon^2} \right\rfloor \tag{23}$$

where $\ell_1 = 152$. By Remark 3, the total variation of the base zero-chain is bounded by $2\Delta_0 \cdot T$. Applying the restriction on T ensures $-\frac{\Delta}{4} \leq F_{\text{scaled}}(y) \leq 0$ globally. The unscaled gradient satisfies $\|\nabla \bar{F}_T(x)\|_\infty > 1$ for $\text{prog}_1(x) < T$. The scaled gradient expands to $\nabla F_{\text{scaled}}(y) = 2\epsilon \nabla \bar{F}_T(y/\lambda)$, thereby preserving $\|\nabla F_{\text{scaled}}(y)\|_\infty > 2\epsilon$ for $\text{prog}_1(y/\lambda) < T$.

We define a composite objective function $F \colon \mathbb{R} \times \mathbb{R}^T \to \mathbb{R}$ that couples the coordinate function with the zero-chain using the activation function according to

$$F(u, y) = f_0(u) + \phi(u) F_{\text{scaled}}(y). \tag{24}$$

The initialization is set to $x_0 = (0, 0)$. It is then immediate that $\inf_{u,y} F(u, y) \geq -\frac{3\Delta}{4} - \frac{4\epsilon^2}{L}$. Imposing $\epsilon \leq \sqrt{\frac{L\Delta}{1536\ell_1}}$ ensures $\frac{4\epsilon^2}{L} \leq \frac{\Delta}{384\ell_1} \leq \frac{\Delta}{4}$. Thus, $F(x_0) - \inf_{u,y} F(u, y) \leq \Delta$.

The composite Hessian matrix $\nabla^2 F(u, y)$ decomposes into a block-diagonal operator H_{diag} and an off-diagonal operator H_{off}. The spectral norm of H_{diag} is bounded by $\frac{L}{2} + \|\phi''\|_\infty |F_{\text{scaled}}(y)|$. Substituting the polynomial and chain bounds yields $\|H_{\text{diag}}\|_2 \leq \frac{L}{2} + \frac{160\epsilon^2}{\sqrt{3}\Delta}$. The spectral norm of the symmetric off-diagonal block H_{off} equates to the Euclidean norm of $\phi'(u)\nabla F_{\text{scaled}}(y)$. By Lemma 4, $\|\nabla \bar{F}_T(x)\|_\infty \leq 23$, and consequently using a norm bound $\|\nabla \bar{F}_T(x)\|_2 \leq 23\sqrt{T}$. Consequently, $\|\nabla F_{\text{scaled}}(y)\|_2 \leq 46\epsilon\sqrt{T} \leq \frac{46}{\sqrt{768\ell_1}}\sqrt{L\Delta}$. Thus, $\|H_{\text{off}}\|_2 \leq \frac{690\epsilon}{\sqrt{768\ell_1\Delta}}\sqrt{L}$. By Weyl's inequality

$$\|\nabla^2 F(u, y)\|_2 \leq \frac{L}{2} + \frac{160\epsilon^2}{\sqrt{3}\Delta} + \frac{690\epsilon}{\sqrt{768\ell_1\Delta}}\sqrt{L}. \tag{25}$$

Since we impose $\epsilon \leq \sqrt{\frac{L\Delta}{1536\ell_1}}$ we have $\|\nabla^2 F(u, y)\|_2 \leq L$, certifying global L-smoothness.

The expanded composite gradient vector evaluates to

$$\nabla F(u, y) = \begin{pmatrix} f_0'(u) + \phi'(u)F_{\text{scaled}}(y) \\ \phi(u)\nabla F_{\text{scaled}}(y) \end{pmatrix}. \tag{26}$$

For any $u \leq D$, $f_0'(u) = -2\epsilon$. The strict non-negativity of $\phi'(u)$ combined with the globally non-positive zero-chain $F_{\text{scaled}}(y) \leq 0$ enforces $\phi'(u)F_{\text{scaled}}(y) \leq 0$. The first gradient coordinate satisfies $f_0'(u) + \phi'(u)F_{\text{scaled}}(y) \leq -2\epsilon$, implying $\|\nabla F(u, y)\| \geq 2\epsilon > \epsilon$. Identifying any ϵ-stationary point then requires progressing to $u > D/2$. At these locations

$$\|x - x_0\|^2 \geq u^2 > \frac{\Delta^2}{64\epsilon^2}. \tag{27}$$

At these locations, under Assumption 1, the BG-0 stochastic oracle variance evaluates to

$$\tilde{\sigma}^2 = B^2\frac{\Delta^2}{64\epsilon^2} + G^2. \tag{28}$$

We deploy a composite stochastic gradient estimator $G(u, y, z)$ incorporating the zero-chain estimator $\bar{g}_T$ from Lemma 5

$$G(u, y, z) = \begin{pmatrix} f_0'(u) + \phi'(u)F_{\text{scaled}}(y) \\ \phi(u)2\epsilon\bar{g}_T(y/\lambda, z) \end{pmatrix}. \tag{29}$$

The variance of this estimator satisfies $\mathbb{E}[\|G(u, y, z) - \nabla F(u, y)\|^2] \leq \frac{4\varsigma^2\epsilon^2(1-p)}{p}$. Setting this upper bound equal to $\tilde{\sigma}^2$ yields

$$p = \frac{256\varsigma^2\epsilon^4}{B^2\Delta^2 + 64G^2\epsilon^2 + 256\varsigma^2\epsilon^4}. \tag{30}$$

By Lemma 3, determining a coordinate satisfying $\text{prog}_1(y/\lambda) \geq T$ demands an expected query complexity bounded strictly below by $\frac{T-1}{2p}$. The constraint $\epsilon \leq \sqrt{\frac{L\Delta}{1536\ell_1}}$ along with the definition of T ensures $T \geq 2$, such that $\frac{T-1}{2p} \geq \frac{T}{4p}$. Substituting the definition of p yields

$$\frac{T}{4p} = \frac{T}{4}\left(1 + \frac{B^2\Delta^2 + 64G^2\epsilon^2}{256\varsigma^2\epsilon^4}\right) \geq \frac{T(B^2\Delta^2 + 64G^2\epsilon^2)}{1024\varsigma^2\epsilon^4}. \tag{31}$$

Applying the dimensional bound $T \geq \frac{L\Delta}{1536\ell_1\epsilon^2}$ leads to the final expected query complexity

$$\frac{T}{4p} \geq \frac{B^2L\Delta^3}{1572864\ell_1\varsigma^2\epsilon^6} + \frac{G^2L\Delta}{24576\ell_1\varsigma^2\epsilon^4}. \tag{32}$$

To apply this bound for all randomized algorithms, following Arjevani et al. (2023), we embed the construction within a randomized orthogonal transformation $V = \text{diag}(1, U)$. Operating on the rotated objective $F(V^\top x)$ preserves Euclidean geometry entirely, directly extending the zero-respecting sequence bound, thereby furnishing the proof. ■

B.2 Proof of Theorem 2

Proof of Theorem 2. The proof is almost identical to that of Theorem 1, but we have to use Lemma 6 and a suitable reparameterization of the deterministic smooth constant as done in Arjevani et al. (2023). Additionally, we note that the restriction on B is essential to ensure $T \geq 2$ holds, something we need in our construction. We rewrite the near-identical calculation here for the ease of the reader.

Let $D = \frac{\Delta}{4\epsilon}$. We define a coordinate function $f_0 \colon \mathbb{R} \to \mathbb{R}$ satisfying $f_0(0) = 0$ with the derivative

$$f_0'(u) = \begin{cases} -2\epsilon, & u \leq D \\ -2\epsilon + \frac{L}{2}(u - D), & u \in (D, D + \frac{4\epsilon}{L}) \\ 0, & u \geq D + \frac{4\epsilon}{L}. \end{cases} \tag{33}$$

Integrating this derivative establishes $\inf_u f_0(u) \geq -\frac{\Delta}{2} - \frac{4\epsilon^2}{L}$. We define a twice-differentiable activation function $\phi \colon \mathbb{R} \to [0, 1]$ such that $\phi(u) = 0$ for $u \leq D/2$ and $\phi(u) = 1$ for $u \geq D$. Over the transition interval $u \in (D/2, D)$, we define the normalization $t = \frac{8\epsilon u}{\Delta} - 1$ and set $\phi(u) = 10t^3 - 15t^4 + 6t^5$. Evaluating the global extrema over this polynomial yields the derivative bounds $\|\phi'\|_\infty = \frac{15\epsilon}{\Delta}$ and $\|\phi''\|_\infty = \frac{640\epsilon^2}{\sqrt{3}\Delta^2}$.

At locations such that $u > D/2$ we have

$$\|x - x_0\|^2 \geq u^2 > \frac{\Delta^2}{64\epsilon^2}. \tag{34}$$

At these locations, under Assumption 1, the BG-0 stochastic oracle variance evaluates to

$$\tilde{\sigma}^2 = B^2 \frac{\Delta^2}{64\epsilon^2} + G^2. \tag{35}$$

We deploy a stochastic gradient estimator $G(u, y, z)$ incorporating the estimator $\bar{g}_T$ from Lemma 6

$$G(u, y, z) = \begin{pmatrix} f_0'(u) + \phi'(u) F_{\text{scaled}}(y) \\ \phi(u) 2\epsilon \bar{g}_T(y/\lambda, z) \end{pmatrix}. \tag{36}$$

The variance of the coupled stochastic gradient $G(u, y, z)$ is bounded by $\frac{4\varsigma^2\epsilon^2(1-p)}{p}$. Setting this variance to $\tilde{\sigma}^2$ yields

$$p = \frac{256\varsigma^2\epsilon^4}{B^2\Delta^2 + 64G^2\epsilon^2 + 256\varsigma^2\epsilon^4}. \tag{37}$$

To satisfy the global mean-squared smoothness condition parameterized by $\bar{L}$ in Assumption 3, we parameterize the effective deterministic smoothness L in Lemma 6 according to $L = \frac{\ell_1}{\bar{\ell}_1}\bar{L}\sqrt{p}$. Following Lemma 4, we construct the scaled and globally shifted zero-chain $F_{\text{scaled}} \colon \mathbb{R}^T \to \mathbb{R}$ using this modified deterministic smoothness L

$$F_{\text{scaled}}(y) = \frac{L\lambda^2}{2\ell_1}\left(\bar{F}_T\left(\frac{y}{\lambda}\right) - \sup_{x \in \mathbb{R}^T} \bar{F}_T(x)\right), \quad \lambda = \frac{4\ell_1\epsilon}{L}, \quad T = \left\lfloor \frac{\bar{L}\Delta\sqrt{p}}{768\bar{\ell}_1\epsilon^2} \right\rfloor. \tag{38}$$

By Remark 3, the total variation of the base zero-chain is bounded by $2\Delta_0 \cdot T$. The definition of T ensures $-\frac{\Delta}{4} \leq F_{\text{scaled}}(y) \leq 0$ globally. For the unscaled gradient we have $\|\nabla \bar{F}_T(x)\|_\infty > 1$ for $\text{prog}_1(x) < T$. Thus, $\nabla F_{\text{scaled}}(y) = 2\epsilon \nabla \bar{F}_T(y/\lambda)$, preserving $\|\nabla F_{\text{scaled}}(y)\|_\infty > 2\epsilon$ for $\text{prog}_1(y/\lambda) < T$.

The composite objective function $F \colon \mathbb{R} \times \mathbb{R}^T \to \mathbb{R}$ couples the coordinate sequence with the zero-chain

$$F(u, y) = f_0(u) + \phi(u) F_{\text{scaled}}(y). \tag{39}$$

The initialization is set to $x_0 = (0, 0)$. Also, $\inf_{u,y} F(u, y) \geq -\frac{3\Delta}{4} - \frac{4\epsilon^2}{L}$. The bound on ϵ leads to $\epsilon^2 \leq \frac{L\Delta}{1536\ell_1}$ such that along with the constraint on T we have $\frac{4\epsilon^2}{L} \leq \frac{\Delta}{384\ell_1} \leq \frac{\Delta}{4}$. Thus, $F(x_0) - \inf_{u,y} F(u, y) \leq \Delta$. Furthermore, substituting $\epsilon^2 \leq \frac{L\Delta}{1536\ell_1}$ bounds the spectral perturbations of the composite Hessian exactly as in Theorem 1, ensuring $\|\nabla^2 F(u, y)\|_2 \leq L$ and deterministic L-smoothness.

The expanded composite gradient vector evaluates to

$$\nabla F(u,y) = \begin{pmatrix} f_0'(u) + \phi'(u) F_{\text{scaled}}(y) \\ \phi(u) \nabla F_{\text{scaled}}(y) \end{pmatrix}. \tag{40}$$

For any $u \le D$, $f_0'(u) = -2\epsilon$. The strict non-negativity of $\phi'(u)$ combined with the globally non-positive zero-chain $F_{\text{scaled}}(y) \le 0$ enforces $\phi'(u)F_{\text{scaled}}(y) \le 0$. The first gradient coordinate satisfies $f_0'(u) + \phi'(u)F_{\text{scaled}}(y) \le -2\epsilon$, implying $\|\nabla F(u,y)\| \ge 2\epsilon > \epsilon$. Identifying any ϵ-stationary point, therefore, mathematically necessitates progressing to $u > D/2$.

By Lemma 3, determining a coordinate satisfying $\text{prog}_1(y/\lambda) \ge T$ requires an expected query complexity bounded strictly below by $\frac{T-1}{2p}$. The restriction on ϵ ensures $T \ge 2$, establishing $\frac{T-1}{2p} \ge \frac{T}{4p}$. Substituting the dimension T yields

$$\frac{T}{4p} = \frac{1}{4p}\left\lfloor \frac{\bar{L}\Delta\sqrt{p}}{768\bar{\ell}_1\epsilon^2} \right\rfloor \ge \frac{\bar{L}\Delta}{6144\bar{\ell}_1\epsilon^2\sqrt{p}}. \tag{41}$$

Noting $\frac{1}{\sqrt{p}} \ge \frac{\sqrt{B^2\Delta^2+64G^2\epsilon^2}}{16\varsigma\epsilon^2}$ and substituting it into (41) yields

$$\frac{T}{4p} \ge \frac{\bar{L}\Delta\sqrt{B^2\Delta^2+64G^2\epsilon^2}}{98304\bar{\ell}_1\varsigma\epsilon^4} \ge \frac{B\bar{L}\Delta^2}{98304\sqrt{2}\bar{\ell}_1\varsigma\epsilon^4} + \frac{G\bar{L}\Delta}{12288\sqrt{2}\bar{\ell}_1\varsigma\epsilon^3} \tag{42}$$

where we used $\sqrt{A+B} \ge \frac{1}{\sqrt{2}}(\sqrt{A}+\sqrt{B})$.

By Lemma 7, extracting an ϵ-stationary mean vector from an oracle with variance $\tilde{\sigma}^2$ requires $c_0\frac{\tilde{\sigma}^2}{\epsilon^2}$ independent queries as long as ϵ satisfies the imposed upper bound. Thus, when $u > D/2$, this evaluates to

$$\frac{\tilde{\sigma}^2}{\epsilon^2} = \frac{B^2\Delta^2}{64\epsilon^4} + \frac{G^2}{\epsilon^2}. \tag{43}$$

Integrating the last bound with the expected query complexity lower bound on $\frac{T}{4p}$, in a near-identical fashion as done in Arjevani et al. (2023) leads to the lower bound. Lifting this complexity to the universal class of all randomized algorithms is done exactly as explained in the proof of Theorem 1 by following the randomized rotation idea of Arjevani et al. (2023). ■

C Proofs of Upper Bounds in Section 4

C.1 Proof of Theorem 3

Proof of Theorem 3. Setting $S = 1$, $\beta_t = 0$, with $p_t = 1$ simplifies PASTA to $x_{t+1} = x_t - \eta g_t$. The mini-batch estimator g_t is evaluated utilizing a dynamic batch size N_t

$$N_t = \max\left\{1, \left\lceil \frac{B^2\|x_t - x_0\|^2 + G^2}{\sigma^2} \right\rceil \right\} \tag{44}$$

where $\sigma^2 > 0$ is some target variance fixed across iterations. By the independence of the stochastic queries conditioned on the filtration at t and the $\mathcal{F}_{t-1}$-measurability of x_t

$$\mathbb{E}_t[\|g_t - \nabla f(x_t)\|^2] \le \frac{B^2\|x_t - x_0\|^2 + G^2}{N_t} \le \sigma^2. \tag{45}$$

The next few steps are standard. Applying L-smoothness

$$f(x_{t+1}) \le f(x_t) - \eta\langle \nabla f(x_t), g_t\rangle + \frac{L\eta^2}{2}\|g_t\|^2. \tag{46}$$

Taking the conditional expectation and invoking unbiasedness

$$\mathbb{E}_t[f(x_{t+1})] \le f(x_t) - \eta\|\nabla f(x_t)\|^2 + \frac{L\eta^2}{2}\mathbb{E}_t\left[\|g_t\|^2\right]. \tag{47}$$

Expanding the second moment and utilizing the variance bound above,

$$\mathbb{E}_t[f(x_{t+1})] \leq f(x_t) - \eta\left(1 - \frac{L\eta}{2}\right)\|\nabla f(x_t)\|^2 + \frac{L\eta^2\sigma^2}{2}. \tag{48}$$

Since $\eta \leq \frac{1}{L}$, we have $1 - \frac{L\eta}{2} \geq \frac{1}{2}$. Taking the total expectation, recursively unrolling from $t = 0$ to $K - 1$, and rearranging

$$\sum_{t=0}^{K-1} \mathbb{E}[\|\nabla f(x_t)\|^2] \leq \frac{2(f(x_0) - \mathbb{E}[f(x_K)])}{\eta} + L\eta K\sigma^2 \leq \frac{2\Delta}{\eta} + KL\eta\sigma^2. \tag{49}$$

Recall that $\eta \leq \frac{\epsilon^2}{2L\sigma^2}$ and $K = \left\lceil \frac{4\Delta}{\eta\epsilon^2} \right\rceil$.

To find the total oracle complexity, we need to estimate $\|x_t - x_0\|$, which determines the batch size N_t. Using PASTA's update rule

$$x_t - x_0 = -\eta \sum_{j=0}^{t-1} \nabla f(x_j) - \eta \sum_{j=0}^{t-1} (g_j - \nabla f(x_j)). \tag{50}$$

Using $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and taking the expectation

$$\mathbb{E}[\|x_t - x_0\|^2] \leq 2\eta^2\mathbb{E}\Big[\Big\|\sum_{j=0}^{t-1} \nabla f(x_j)\Big\|^2\Big] + 2\eta^2\mathbb{E}\Big[\Big\|\sum_{j=0}^{t-1} (g_j - \nabla f(x_j))\Big\|^2\Big]. \tag{51}$$

The first term is bounded by $t\sum_{j=0}^{t-1} \|\nabla f(x_j)\|^2$. Substituting the bound from (49)

$$2\eta^2\mathbb{E}\Big[\Big\|\sum_{j=0}^{t-1} \nabla f(x_j)\Big\|^2\Big] \leq 2\eta^2 K \sum_{j=0}^{K-1} \mathbb{E}\big[\|\nabla f(x_j)\|^2\big] \leq 4\eta K\Delta + 2K^2L\eta^3\sigma^2. \tag{52}$$

On the other hand, by the unbiasedness of the oracle, the second term can be bounded by

$$2\eta^2\mathbb{E}\Big[\Big\|\sum_{j=0}^{t-1} (g_j - \nabla f(x_j))\Big\|^2\Big] = 2\eta^2 \sum_{j=0}^{t-1} \mathbb{E}\big[\|g_j - \nabla f(x_j)\|^2\big] \leq 2\eta^2 K\sigma^2. \tag{53}$$

Aggregating these two bounds yields

$$\mathbb{E}\big[\|x_t - x_0\|^2\big] \leq 4\eta K\Delta + 2K^2L\eta^3\sigma^2 + 2\eta^2 K\sigma^2. \tag{54}$$

To find an ϵ-stationary solution, note from (49)

$$\frac{1}{K}\sum_{t=0}^{K-1} \mathbb{E}\big[\|\nabla f(x_t)\|^2\big] \leq \frac{2\Delta}{K\eta} + L\eta\sigma^2 \leq \epsilon^2. \tag{55}$$

It remains to bound the total stochastic oracle complexity. We consider the two possible branches of the stepsize.

Case 1: $\eta = \frac{\epsilon^2}{2L\sigma^2}$, equivalently $\epsilon^2 \leq 2\sigma^2$.

In this case, since $K = \left\lceil \frac{4\Delta}{\eta\epsilon^2} \right\rceil$, the bound (54) simplifies to

$$\mathbb{E}\big[\|x_t - x_0\|^2\big] \leq \frac{32\Delta^2}{\epsilon^2} + \frac{4\Delta}{L} + \frac{6\Delta\epsilon^2}{L\sigma^2} + \frac{\epsilon^4}{2L^2\sigma^2} + \frac{\epsilon^6}{4L^2\sigma^4}. \tag{56}$$

The total stochastic oracle complexity integrates the dynamic batch sizes across the entire optimization path. Taking the total expectation,

$$\mathbb{E}[N_{total}] = \sum_{t=0}^{K-1} \mathbb{E}[N_t] \leq \sum_{t=0}^{K-1} \left(1 + \frac{B^2\mathbb{E}\big[\|x_t - x_0\|^2\big] + G^2}{\sigma^2}\right). \tag{57}$$

Substituting the bound on $\mathbb{E}\left[\|x_t - x_0\|^2\right]$ found above yields

$$\mathbb{E}[N_{total}] \leq \frac{256B^2\Delta^3 L}{\epsilon^6} + \frac{80B^2\Delta^2}{\epsilon^2\sigma^2} + \frac{32B^2\Delta^2}{\epsilon^4} + \frac{8B^2\Delta\epsilon^2}{L\sigma^4} + \frac{8B^2\Delta}{L\sigma^2} + \frac{B^2\epsilon^6}{4L^2\sigma^6} + \frac{B^2\epsilon^4}{2L^2\sigma^4} + \frac{8\Delta LG^2}{\epsilon^4} + \frac{8\Delta L\sigma^2}{\epsilon^4} + \frac{G^2}{\sigma^2} + 1, \tag{58}$$

and therefore $\mathbb{E}[N_{\text{total}}] = \mathcal{O}(\epsilon^{-6})$.

Case 2: $\eta = \frac{1}{L}$, equivalently $\epsilon^2 > 2\sigma^2$. In this case, $K = \left\lceil \frac{4\Delta L}{\epsilon^2} \right\rceil = \mathcal{O}(\epsilon^{-2})$. Using (54),

$$\mathbb{E}\left[\|x_t - x_0\|^2\right] \leq \frac{4K\Delta}{L} + \frac{2K^2\sigma^2}{L^2} + \frac{2K\sigma^2}{L^2} = \mathcal{O}(\epsilon^{-4}). \tag{59}$$

Therefore,

$$\begin{aligned}\mathbb{E}[N_{total}] = \sum_{t=0}^{K-1} \mathbb{E}[N_t] &\leq \sum_{t=0}^{K-1} \left(1 + \frac{B^2\mathbb{E}\left[\|x_t - x_0\|^2\right] + G^2}{\sigma^2}\right) \\ &\leq K\left(1 + \frac{G^2}{\sigma^2}\right) + \frac{B^2}{\sigma^2}\sum_{t=0}^{K-1} \mathbb{E}\left[\|x_t - x_0\|^2\right] = \mathcal{O}(\epsilon^{-2}) + \mathcal{O}(\epsilon^{-2})\mathcal{O}(\epsilon^{-4}) = \mathcal{O}(\epsilon^{-6}).\end{aligned} \tag{60}$$

Combining the two cases, we conclude that $\mathbb{E}[N_{total}] = \mathcal{O}(\epsilon^{-6})$, which finishes the proof. ■

C.2 Results under Convexity

Theorem 9 (Convergence of PASTA via Dynamic Batching - Smooth Convex Case)**.** *Let Assumptions 1 and 2 hold. Furthermore, let f be convex (See Assumption 6 with $\rho = 0$). Let the sequence $\{x_t\}$ be generated by PASTA with $S = 1$, $\beta_t = 0$, and $p_t = 1$. Given a constant target variance $\sigma^2 > 0$, set batch size N_t, the step size η, and total iterations K according to*

$$N_t = \max\left\{1, \left\lceil \frac{B^2\|x_t - x_0\|^2 + G^2}{\sigma^2} \right\rceil\right\}, \quad \eta = \min\left\{\frac{1}{2L}, \frac{\epsilon}{2\sigma^2}\right\}, \quad K = \left\lceil \frac{2\|x_0 - x^*\|^2}{\eta\epsilon} \right\rceil. \tag{61}$$

Then, for the ergodic average $\bar{x}_K = \frac{1}{K}\sum_{t=0}^{K-1} x_t$, $\mathbb{E}[f(\bar{x}_K) - f(x^*)] \leq \epsilon$ and the expected total stochastic oracle complexity evaluates to $\mathcal{O}(\epsilon^{-2})$.

The proof is in Appendix C.3, where the exact expected oracle complexity can be found in (72) and below. Using a similar technique, we can state the following theorem for convex functions that are not smooth, the proof of which is in Appendix C.4, where the exact expected oracle complexity can be found in (84).

Theorem 10 (Convergence of PASTA under Lipschitz Continuity and Convexity)**.** *Let Assumption 1 hold and assume f is convex (see Assumption 6 with $\rho = 0$ and C-Lipschitz (see Assumption 7). Let the sequence $\{x_t\}$ be generated by PASTA with $S = 1$, $\beta_t = 0$, and $p_t = 1$. Given a constant target variance $\sigma^2 > 0$, set the batch size N_t, the step size η, and total iterations K according to*

$$N_t = \max\left\{1, \left\lceil \frac{B^2\|x_t - x_0\|^2 + G^2}{\sigma^2} \right\rceil\right\}, \quad \eta = \frac{\epsilon}{C^2 + \sigma^2}, \quad K = \left\lceil \frac{(C^2 + \sigma^2)\|x_0 - x^*\|^2}{\epsilon^2} \right\rceil. \tag{62}$$

Then, for the ergodic average $\bar{x}_K = \frac{1}{K}\sum_{t=0}^{K-1} x_t$, $\mathbb{E}[f(\bar{x}_K) - f(x^*)] \leq \epsilon$ and the expected total stochastic oracle complexity evaluates to $\mathcal{O}(\epsilon^{-2})$.

Note that one could avoid using the Lipschitzness assumption if, following Neu & Okolo (2024); Alacaoglu et al. (2025) the variance condition in Assumption 1 is strengthened to $\mathbb{E}[\|\widetilde{\nabla} f(x)\|^2] \leq B^2\|x - x_0\|^2 + G^2$.

C.3 Proof of Theorem 9

Proof of Theorem 9. Under the parameterization $S = 1$, $p_t = 1$, and $\beta_t = 0$, PASTA reduces to $x_{t+1} = x_t - \eta g_t$. The dynamic batch size N_t and the independence of the oracle queries conditioned

on $\mathcal{F}_t$ ensure

$$\mathbb{E}_t\big[\|g_t - \nabla f(x_t)\|^2\big] \le \sigma^2. \tag{63}$$

Expanding the distance to the optimum and taking the conditional expectation yields

$$\mathbb{E}_t\big[\|x_{t+1} - x^*\|^2\big] = \|x_t - x^*\|^2 - 2\eta\langle \nabla f(x_t), x_t - x^*\rangle + \eta^2 \mathbb{E}_t\big[\|g_t\|^2\big]. \tag{64}$$

The convexity of f means $\langle \nabla f(x_t), x_t - x^*\rangle \ge f(x_t) - f(x^*)$. Also, by L-smoothness we have $\|\nabla f(x_t)\|^2 \le 2L(f(x_t) - f(x^*))$. Therefore,

$$\mathbb{E}_t\big[\|g_t\|^2\big] \le 2L(f(x_t) - f(x^*)) + \sigma^2. \tag{65}$$

Substituting these relations into the distance expansion yields

$$\mathbb{E}_t\big[\|x_{t+1} - x^*\|^2\big] \le \|x_t - x^*\|^2 - 2\eta(1 - L\eta)(f(x_t) - f(x^*)) + \eta^2\sigma^2. \tag{66}$$

Since $\eta \le \frac{1}{2L}$, we have $1 - L\eta \ge \frac{1}{2}$. Taking the total expectation and rearranging

$$\eta\mathbb{E}[f(x_t) - f(x^*)] \le \mathbb{E}\big[\|x_t - x^*\|^2\big] - \mathbb{E}\big[\|x_{t+1} - x^*\|^2\big] + \eta^2\sigma^2. \tag{67}$$

Summing from $t = 0$ to $K - 1$ and discarding the non-positive terms yields

$$\sum_{t=0}^{K-1} \mathbb{E}[f(x_t) - f(x^*)] \le \frac{\|x_0 - x^*\|^2}{\eta} + K\eta\sigma^2. \tag{68}$$

Dividing by K, by convexity, and noting the definition of the ergodic average $\bar{x}_K$

$$\mathbb{E}[f(\bar{x}_K) - f(x^*)] \le \frac{\|x_0 - x^*\|^2}{K\eta} + \eta\sigma^2. \tag{69}$$

Since $\eta \le \frac{\epsilon}{2\sigma^2}$ and $K = \left\lceil \frac{2\|x_0 - x^*\|^2}{\eta\epsilon} \right\rceil$ guarantees $\mathbb{E}[f(\bar{x}_K) - f(x^*)] \le \epsilon$.

To determine the expected total stochastic oracle complexity, recall (66) and discard the non-positive suboptimality term to bound $\mathbb{E}\big[\|x_{t+1} - x^*\|^2\big] \le \mathbb{E}\big[\|x_t - x^*\|^2\big] + \eta^2\sigma^2$. Unrolling this establishes $\mathbb{E}\big[\|x_t - x^*\|^2\big] \le \|x_0 - x^*\|^2 + t\eta^2\sigma^2$. Applying the inequality $\|a + b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ bounds

$$\mathbb{E}\big[\|x_t - x_0\|^2\big] \le 4\|x_0 - x^*\|^2 + 2t\eta^2\sigma^2. \tag{70}$$

The total expected stochastic oracle complexity amounts to

$$\mathbb{E}[N_{total}] = \sum_{t=0}^{K-1} \mathbb{E}[N_t] \le \sum_{t=0}^{K-1} \left(1 + \frac{B^2\mathbb{E}\big[\|x_t - x_0\|^2\big] + G^2}{\sigma^2}\right). \tag{71}$$

Substituting $\mathbb{E}\big[\|x_t - x^*\|^2\big] \le \|x_0 - x^*\|^2 + t\eta^2\sigma^2$ and utilizing $\sum_{t=0}^{K-1} t \le \frac{K^2}{2}$ yields

$$\mathbb{E}[N_{total}] \le K\left(1 + \frac{G^2 + 4B^2\|x_0 - x^*\|^2}{\sigma^2}\right) + B^2K^2\eta^2. \tag{72}$$

It remains to bound the last expression. We consider the two possible branches of the stepsize.

Case 1: $\eta = \frac{\epsilon}{2\sigma^2}$, equivalently $\epsilon \le \frac{\sigma^2}{L}$. In this case,

$$K = \left\lceil \frac{2\|x_0 - x^*\|^2}{\eta\epsilon} \right\rceil = \left\lceil \frac{4\sigma^2\|x_0 - x^*\|^2}{\epsilon^2} \right\rceil = \mathcal{O}(\epsilon^{-2}),$$

and

$$K^2\eta^2 = \mathcal{O}(\epsilon^{-4}) \cdot \mathcal{O}(\epsilon^2) = \mathcal{O}(\epsilon^{-2}).$$

Therefore, (72) implies $\mathbb{E}[N_{total}] = \mathcal{O}(\epsilon^{-2})$.

Case 2: $\eta = \frac{1}{2L}$, equivalently $\epsilon \ge \frac{\sigma^2}{L}$. In this case,

$$K = \left\lceil \frac{2\|x_0 - x^*\|^2}{\eta\epsilon} \right\rceil = \left\lceil \frac{4L\|x_0 - x^*\|^2}{\epsilon} \right\rceil = \mathcal{O}(\epsilon^{-1}),$$

and

$$K^2\eta^2 = \mathcal{O}(\epsilon^{-2}).$$

Therefore, (72) again implies $\mathbb{E}[N_{total}] = \mathcal{O}(\epsilon^{-2})$. Combining the two cases, furnishes the proof. ■

C.4 Proof of Theorem 10

Proof of Theorem 10. The parameterization $S = 1$, $p_t = 1$, and $\beta_t = 0$ reduces the algorithm to $x_{t+1} = x_t - \eta g_t$. By dynamic batch size definition and the conditional independence of the oracle queries

$$\mathbb{E}_t\big[\|g_t - \bar{g}_t\|^2\big] \le \sigma^2 \tag{73}$$

where $\bar{g}_t \in \partial f(x_t)$ represents the true subgradient. Expanding yields

$$\mathbb{E}_t\big[\|x_{t+1} - x^*\|^2\big] = \|x_t - x^*\|^2 - 2\eta\langle \bar{g}_t, x_t - x^*\rangle + \eta^2 \mathbb{E}_t\big[\|g_t\|^2\big]. \tag{74}$$

By convexity $-\langle \bar{g}_t, x_t - x^*\rangle \le -(f(x_t) - f(x^*))$ while by C-Lipschitz $\|\bar{g}_t\|^2 \le C^2$ such that

$$\mathbb{E}_t\big[\|g_t\|^2\big] = \|\bar{g}_t\|^2 + \mathbb{E}_t\big[\|g_t - \bar{g}_t\|^2\big] \le C^2 + \sigma^2. \tag{75}$$

Integrating these establishes

$$\mathbb{E}_t\big[\|x_{t+1} - x^*\|^2\big] \le \|x_t - x^*\|^2 - 2\eta(f(x_t) - f(x^*)) + \eta^2(C^2 + \sigma^2). \tag{76}$$

Taking the total expectation and rearranging

$$\mathbb{E}[f(x_t) - f(x^*)] \le \frac{\mathbb{E}\big[\|x_t - x^*\|^2\big] - \mathbb{E}\big[\|x_{t+1} - x^*\|^2\big]}{2\eta} + \frac{\eta(C^2 + \sigma^2)}{2}. \tag{77}$$

Summing from $t = 0$ to $K - 1$ and discarding the non-positive term $\mathbb{E}\big[\|x_K - x^*\|^2\big]$ yields

$$\sum_{t=0}^{K-1} \mathbb{E}[f(x_t) - f(x^*)] \le \frac{\|x_0 - x^*\|^2}{2\eta} + \frac{K\eta(C^2 + \sigma^2)}{2}. \tag{78}$$

Dividing by K, using convexity, and the definition of the ergodic average $\bar{x}_K$ establishes

$$\mathbb{E}[f(\bar{x}_K) - f(x^*)] \le \frac{\|x_0 - x^*\|^2}{2K\eta} + \frac{\eta(C^2 + \sigma^2)}{2}. \tag{79}$$

Substituting $\eta = \frac{\epsilon}{C^2+\sigma^2}$ and $K = \left\lceil \frac{(C^2+\sigma^2)\|x_0 - x^*\|^2}{\epsilon^2} \right\rceil$ ensures $\mathbb{E}[f(\bar{x}_K) - f(x^*)] \le \epsilon$.

To determine the total expected stochastic oracle complexity, recall (76) and remove the non-positive suboptimality term to obtain $\mathbb{E}\big[\|x_{t+1} - x^*\|^2\big] \le \mathbb{E}\big[\|x_t - x^*\|^2\big] + \eta^2(C^2 + \sigma^2)$. Unrolling this yields

$$\mathbb{E}\big[\|x_t - x^*\|^2\big] \le \|x_0 - x^*\|^2 + t\eta^2(C^2 + \sigma^2). \tag{80}$$

Applying $\|a + b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ we get

$$\mathbb{E}\big[\|x_t - x_0\|^2\big] \le 4\|x_0 - x^*\|^2 + 2t\eta^2(C^2 + \sigma^2). \tag{81}$$

The expected total stochastic oracle complexity amounts to

$$\mathbb{E}[N_{total}] = \sum_{t=0}^{K-1} \mathbb{E}[N_t] \le \sum_{t=0}^{K-1} \left(1 + \frac{B^2 \mathbb{E}\big[\|x_t - x_0\|^2\big] + G^2}{\sigma^2}\right). \tag{82}$$

Substituting the bound on $\mathbb{E}\big[\|x_t - x_0\|^2\big]$ and utilizing $\sum_{t=0}^{K-1} t \le \frac{K^2}{2}$

$$\mathbb{E}[N_{total}] \le K\left(1 + \frac{G^2 + 4B^2\|x_0 - x^*\|^2}{\sigma^2}\right) + \frac{B^2}{\sigma^2} K^2 \eta^2 (C^2 + \sigma^2). \tag{83}$$

Substituting the definitions of K and η yields

$$\mathbb{E}[N_{total}] \le \left\lceil \frac{(C^2+\sigma^2)\|x_0 - x^*\|^2}{\epsilon^2} \right\rceil \left(1 + \frac{G^2 + 4B^2\|x_0 - x^*\|^2}{\sigma^2}\right) + \frac{B^2\epsilon^2}{\sigma^2(C^2+\sigma^2)} \left\lceil \frac{(C^2+\sigma^2)\|x_0 - x^*\|^2}{\epsilon^2} \right\rceil^2, \tag{84}$$

which is evidently $\mathcal{O}(\epsilon^{-2})$. ■

C.5 Proof of Theorem 4

Proof of Theorem 4. Let $\bar{g}_t = \nabla f(x_t)$ and $g_t = \bar{g}_t + \xi_t$ for an unbiased noise term ξ_t. Expanding yields

$$\|x_{t+1} - x^*\|^2 = \|x_t - x^* - \eta\bar{g}_t - \beta_t(x_t - x_0) - \eta\xi_t\|^2 \tag{85}$$

Taking the conditional expectation

$$\mathbb{E}_t\left[\|x_{t+1} - x^*\|^2\right] = \|x_t - x^* - \eta\bar{g}_t - \beta_t(x_t - x_0)\|^2 + \eta^2\mathbb{E}_t\left[\|\xi_t\|^2\right] \tag{86}$$

Expanding establishes

$$\mathbb{E}_t\left[\|x_{t+1}-x^*\|^2\right] = \|x_t-x^*\|^2-2\langle\eta\bar{g}_t+\beta_t(x_t-x_0), x_t-x^*\rangle+\|\eta\bar{g}_t+\beta_t(x_t-x_0)\|^2+\eta^2\mathbb{E}_t\left[\|\xi_t\|^2\right] \tag{87}$$

By convexity $-2\eta\langle\bar{g}_t, x_t - x^*\rangle \leq 0$. Applying the polarization identity $-2\langle a, b\rangle = \|a - b\|^2 - \|a\|^2 - \|b\|^2$ yields

$$-2\beta_t\langle x_t - x_0, x_t - x^*\rangle = -\beta_t\|x_t - x_0\|^2 - \beta_t\|x_t - x^*\|^2 + \beta_t\|x_0 - x^*\|^2. \tag{88}$$

Using $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and the consequence of L-smoothness $\|\bar{g}_t\|^2 \leq L^2\|x_t - x^*\|^2$ we have

$$\|\eta\bar{g}_t + \beta_t(x_t - x_0)\|^2 \leq 2\eta^2L^2\|x_t - x^*\|^2 + 2\beta_t^2\|x_t - x_0\|^2. \tag{89}$$

Using BG-0 oracle and a dynamic mini-batch size N_t

$$\eta^2\mathbb{E}_t\left[\|\xi_t\|^2\right] \leq \frac{\eta^2B^2}{N_t}\|x_t - x_0\|^2 + \frac{\eta^2G^2}{N_t} \tag{90}$$

Combining these bounds yields

$$\begin{aligned}\mathbb{E}_t\left[\|x_{t+1} - x^*\|^2\right] \leq{}& \left(1 - \beta_t + 2\eta^2L^2\right)\|x_t - x^*\|^2 \\ &+ \left(-\beta_t + 2\beta_t^2 + \frac{\eta^2B^2}{N_t}\right)\|x_t - x_0\|^2 + \beta_t\|x_0 - x^*\|^2 + \frac{\eta^2G^2}{N_t}\end{aligned} \tag{91}$$

Imposing $\eta \leq \frac{\sqrt{\beta_t}}{2L}$ ensures the contraction coefficient $1 - \beta_t + 2\eta^2L^2$ is bounded from above by $1 - \frac{\beta_t}{2}$. To guarantee convergence, the coefficient of $\|x_t - x_0\|^2$ must remain non-positive. We thus set $-\beta_t + 2\beta_t^2 + \frac{\eta^2B^2}{N_t} \leq 0$ to get $\beta_t(1 - 2\beta_t)N_t \geq \eta^2B^2$, thereby leading to the residual error $\mathcal{E}_t = \beta_t\|x_0 - x^*\|^2 + \frac{\eta^2G^2}{N_t}$, and concluding the proof. ■

C.6 Proof of Theorem 5

Proof of Theorem 5. Define a reference large batch size N_{ref}, a small batch size b, a fixed large-batch probability p, and a fixed step size η as follows

$$N_{\text{ref}} = \max\left\{1, \left\lceil\frac{8B^2\Delta^2}{\epsilon^4} + \frac{2G^2}{\epsilon^2}\right\rceil\right\}, \quad b = \left\lceil\sqrt{N_{\text{ref}}}\right\rceil, \quad p = \frac{b}{N_{\text{ref}}}, \quad \eta = \frac{1}{4\bar{L}}. \tag{92}$$

At initialization, we compute $g_0 = \frac{1}{N_0}\sum_{i=1}^{N_0}\widetilde{\nabla}f(x_0; \xi_0^{(i)})$ using $N_0 = \max\left\{1, \lceil 2G^2\epsilon^{-2}\rceil\right\}$. We set $\beta = 0$ and $S = 1$, thereby using the update rule $x_{t+1} = x_t - \eta g_t$.

At each iteration $t \geq 1$, the dynamic large batch size is set to

$$N_t = \max\left\{N_{\text{ref}}, \left\lceil\frac{2(B^2\|x_t - x_0\|^2 + G^2)}{\epsilon^2}\right\rceil\right\}. \tag{93}$$

With probability p, the algorithm performs a reset and uses $g_t = \frac{1}{N_t}\sum_{i=1}^{N_t}\widetilde{\nabla}f(x_t; \xi_t^{(i)})$, whereas with probability $1 - p$, the estimator employs the small batch size and computes $g_t = g_{t-1} + \frac{1}{b}\sum_{i=1}^{b}\left(\widetilde{\nabla}f(x_t; \zeta_t^{(i)}) - \widetilde{\nabla}f(x_{t-1}; \zeta_t^{(i)})\right)$.

Define the random estimation error $V_t = \|g_t - \nabla f(x_t)\|^2$. Since x_t is $\mathcal{F}_{t-1}$-measurable, the variance of the large batch is bounded by $\frac{B^2\|x_t - x_0\|^2 + G^2}{N_t} \leq \frac{\epsilon^2}{2}$ using the definition of N_t. Leveraging the MSS condition and Lemma 1

$$\mathbb{E}_t[V_t] \leq p\frac{\epsilon^2}{2} + (1 - p)\left(V_{t-1} + \frac{\bar{L}^2\eta^2}{b}\|g_{t-1}\|^2\right). \tag{94}$$

Note any MSS function is $\bar{L}$-smoothness, such that using the descent lemma for smooth functions and the polarization identity $-2\langle a,b\rangle = \|a-b\|^2 - \|a\|^2 - \|b\|^2$

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|^2 - \frac{\eta}{2}(1-\bar{L}\eta)\|g_t\|^2 + \frac{\eta}{2}V_t. \tag{95}$$

We define the Lyapunov function $\Phi_t = f(x_t) - \inf_x f(x) + \frac{\eta}{2p}V_t$. Evaluating the conditional expectation of Φ_{t+1} and integrating the above descent bound with (94)

$$\mathbb{E}_{t+1}[\Phi_{t+1}] \leq f(x_t) - \inf_x f(x) - \frac{\eta}{2}\|\nabla f(x_t)\|^2 - \frac{\eta}{2}\left(1 - \bar{L}\eta - \frac{(1-p)\bar{L}^2\eta^2}{pb}\right)\|g_t\|^2 + \frac{\eta}{2p}V_t + \frac{\eta\epsilon^2}{4}. \tag{96}$$

By setting $\eta = \frac{1}{4\bar{L}}$ and utilizing $pb = \frac{b^2}{N_{\text{ref}}} \geq 1$, the coefficient of $\|g_t\|^2$ satisfies $1 - \frac{1}{4} - \frac{1-p}{16pb} \geq \frac{3}{4} - \frac{1}{16} = \frac{11}{16} \geq \frac{1}{2}$. Taking the full unconditional expectation over all random variables yields

$$\mathbb{E}[\Phi_{t+1}] \leq \mathbb{E}[\Phi_t] - \frac{\eta}{2}\mathbb{E}[\|\nabla f(x_t)\|^2] - \frac{\eta}{4}\mathbb{E}[\|g_t\|^2] + \frac{\eta\epsilon^2}{4}. \tag{97}$$

Telescoping this inequality K and rearranging

$$\sum_{t=0}^{K-1}\mathbb{E}[\|\nabla f(x_t)\|^2] + \frac{1}{2}\sum_{t=0}^{K-1}\mathbb{E}[\|g_t\|^2] \leq \frac{2\mathbb{E}[\Phi_0]}{\eta} + \frac{K\epsilon^2}{2}. \tag{98}$$

At initialization, we use N_0 oracle calls which means $\mathbb{E}[V_0] \leq \frac{\epsilon^2}{2}$, thereby we have $\mathbb{E}[\Phi_0] \leq \Delta + \frac{\eta\epsilon^2}{4p}$. Using $K = \lceil \frac{16\Delta\bar{L}}{\epsilon^2} + \frac{1}{p} \rceil$ in (98) we have

$$\frac{1}{K}\sum_{t=0}^{K-1}\mathbb{E}[\|\nabla f(x_t)\|^2] \leq \frac{2\Delta}{K\eta} + \frac{\epsilon^2}{2pK} + \frac{\epsilon^2}{2} \leq \frac{4\Delta + \frac{\eta\epsilon^2}{p}}{2\eta\left(\frac{4\Delta}{\eta\epsilon^2} + \frac{1}{p}\right)} + \frac{\epsilon^2}{2} = \epsilon^2. \tag{99}$$

To bound the dynamic batch size, using (98) we can write

$$\sum_{t=0}^{K-1}\mathbb{E}[\|g_t\|^2] \leq \frac{4\Delta}{\eta} + \frac{\epsilon^2}{p} + K\epsilon^2 \leq 16\Delta\bar{L} + \frac{\epsilon^2}{p} + \left(\frac{16\Delta\bar{L}}{\epsilon^2} + \frac{1}{p} + 1\right)\epsilon^2 = 32\Delta\bar{L} + \frac{2\epsilon^2}{p} + \epsilon^2. \tag{100}$$

As we did in the proof of Theorem 3,

$$\mathbb{E}[\|x_t - x_0\|^2] \leq t\eta^2\sum_{j=0}^{t-1}\mathbb{E}[\|g_j\|^2] \leq K\eta^2(32\Delta\bar{L} + \frac{2\epsilon^2}{p} + \epsilon^2). \tag{101}$$

The expected computational cost at step $t \geq 1$ is $\mathbb{E}[C_t] = p\mathbb{E}[N_t] + (1-p)b \leq p\mathbb{E}[N_t] + b$, which amounts to the following using the definition of our parameters

$$\begin{aligned}\mathbb{E}[C_t] &\leq p\left(N_{\text{ref}} + 1 + \frac{2B^2K\eta^2\left(32\Delta\bar{L} + \frac{2\epsilon^2}{p} + \epsilon^2\right) + 2G^2}{\epsilon^2}\right) + b \\ &= 2b + p + p\frac{2G^2}{\epsilon^2} + p\frac{2B^2}{\epsilon^2}K\eta^2\left(32\Delta\bar{L} + \frac{2\epsilon^2}{p} + \epsilon^2\right).\end{aligned} \tag{102}$$

The total expected stochastic oracle complexity can then be evaluated as follows

$$\mathbb{E}[N_{total}] \leq N_0 + \sum_{t=1}^{K-1}\mathbb{E}[C_t] \leq 2G^2\epsilon^{-2} + 1 + K\left(2b + p + p\frac{2G^2}{\epsilon^2} + p\frac{2B^2}{\epsilon^2}K\eta^2\left(32\Delta\bar{L} + \frac{2\epsilon^2}{p} + \epsilon^2\right)\right). \tag{103}$$

which, using the definition of our parameters, leads to

$$\begin{aligned}\mathbb{E}[N_{total}] \leq \frac{2G^2}{\epsilon^2} + 1 + \left(\frac{16\Delta\bar{L}}{\epsilon^2} + \sqrt{\frac{8B^2\Delta^2}{\epsilon^4} + \frac{2G^2}{\epsilon^2} + 1} + 1\right) \times \\ \left[2b + p + p\frac{2G^2}{\epsilon^2} + p\frac{2B^2}{16\bar{L}^2\epsilon^2}\left(\frac{16\Delta\bar{L}}{\epsilon^2} + \sqrt{\frac{8B^2\Delta^2}{\epsilon^4} + \frac{2G^2}{\epsilon^2} + 1} + 1\right)\left(32\Delta\bar{L} + \frac{2\epsilon^2}{p} + \epsilon^2\right)\right].\end{aligned} \tag{104}$$

Readily, the expected oracle complexity is $\mathcal{O}(\epsilon^{-4})$. ■

C.7 Proof of Theorem 6

Proof of Theorem 6. We set $S = 1$, $N_t = 1$, $p_t = 1$, and $\lambda = 0$ such that $x_{t+1} = x_t - \eta g_t$. By the L-smoothness of the objective f, the standard descent lemma yields

$$f(x_{t+1}) \leq f(x_t) - \eta\langle \nabla f(x_t), g_t\rangle + \frac{L\eta^2}{2}\|g_t\|^2. \tag{105}$$

Take the expectation conditioned on the current iterate x_t and by the unbiasedness of the stochastic gradient oracle

$$\mathbb{E}_t[f(x_{t+1})] \leq f(x_t) - \eta\|\nabla f(x_t)\|^2 + \frac{L\eta^2}{2}\mathbb{E}_t\big[\|g_t\|^2\big]. \tag{106}$$

Using the BG-0 variance condition and unbiasedness

$$\mathbb{E}_t\big[\|g_t\|^2\big] \leq \|\nabla f(x_t)\|^2 + B^2\|x_t - x_0\|^2 + G^2. \tag{107}$$

Substituting this bound into the descent inequality

$$\mathbb{E}_t[f(x_{t+1})] \leq f(x_t) - \eta\left(1 - \frac{L\eta}{2}\right)\|\nabla f(x_t)\|^2 + \frac{L\eta^2B^2}{2}\|x_t - x_0\|^2 + \frac{L\eta^2G^2}{2}. \tag{108}$$

We impose $\eta \leq \frac{1}{L}$ to ensure $1 - \frac{L\eta}{2} \geq \frac{1}{2}$. Subtracting $f(x^*)$ from both sides and using PL condition $-\|\nabla f(x_t)\|^2 \leq -2\mu(f(x_t) - f(x^*))$ establishes

$$\mathbb{E}_t[f(x_{t+1}) - f(x^*)] \leq (1 - \eta\mu)(f(x_t) - f(x^*)) + \frac{L\eta^2B^2}{2}\|x_t - x_0\|^2 + \frac{L\eta^2G^2}{2}. \tag{109}$$

As we discussed in the main paper, for an L-smooth function, the PL condition implies QG (Karimi et al., 2016), i.e. $\|x - x^*\|^2 \leq \frac{2}{\mu}(f(x) - f(x^*))$. Expanding $\|x_t - x_0\|^2 \leq 2\|x_t - x^*\|^2 + 2\|x_0 - x^*\|^2$ yields

$$\|x_t - x_0\|^2 \leq \frac{4}{\mu}(f(x_t) - f(x^*)) + \frac{4}{\mu}(f(x_0) - f(x^*)). \tag{110}$$

Substituting this bound into our previous result yields

$$\mathbb{E}_t[f(x_{t+1}) - f(x^*)] \leq \left(1 - \eta\mu + \frac{2L\eta^2B^2}{\mu}\right)(f(x_t) - f(x^*)) + \frac{2L\eta^2B^2}{\mu}(f(x_0) - f(x^*)) + \frac{L\eta^2G^2}{2}. \tag{111}$$

We impose $\frac{2L\eta^2B^2}{\mu} \leq \frac{\eta\mu}{2}$ to ensure the contraction coefficient is bounded by $1 - \frac{\eta\mu}{2}$. Unfolding recursively from $t = 0$ to K and applying the geometric series sum bound $\sum_{j=0}^{K-1}(1 - \frac{\eta\mu}{2})^j \leq \frac{2}{\eta\mu}$ yields

$$\mathbb{E}[f(x_K) - f(x^*)] \leq \left(1 - \frac{\eta\mu}{2}\right)^K (f(x_0) - f(x^*)) + \frac{4L\eta B^2}{\mu^2}(f(x_0) - f(x^*)) + \frac{L\eta G^2}{\mu}. \tag{112}$$

To guarantee the final suboptimality is less than ϵ, we impose conditions such that each of the three terms in the bound are less than $\frac{\epsilon}{3}$. This requires $\frac{4L\eta B^2}{\mu^2}(f(x_0) - f(x^*)) \leq \frac{\epsilon}{3}$, $\frac{L\eta G^2}{\mu} \leq \frac{\epsilon}{3}$, and $K \geq \frac{2}{\eta\mu}\log(\frac{3(f(x_0)-f(x^*))}{\epsilon})$. Since $S = 1$ and $N_t = 1$, the stochastic oracle complexity is equivalent to K and evaluates to

$$K = \left\lceil \max\left(\frac{2L}{\mu}\log\left(\frac{3\Delta}{\epsilon}\right), \frac{6LG^2}{\mu^2\epsilon}\log\left(\frac{3\Delta}{\epsilon}\right), \frac{24LB^2\Delta}{\mu^3\epsilon}\log\left(\frac{3\Delta}{\epsilon}\right)\right)\right\rceil, \tag{113}$$

which is evidently $\mathcal{O}(\epsilon^{-1}\log(\epsilon^{-1}))$. ■

C.8 Proof of Theorem 7

Proof of Theorem 7. We set $S = 1$, $N_t = 1$, $p_t = 1$, and $\lambda = 0$ such that $x_{t+1} = x_t - \eta g_t$. Expanding $\|x_{t+1} - x^*\|^2$ yields

$$\mathbb{E}_t\big[\|x_{t+1} - x^*\|^2\big] = \|x_t - x^*\|^2 - 2\eta\langle \nabla f(x_t), x_t - x^*\rangle + \eta^2\mathbb{E}_t\big[\|g_t\|^2\big]. \tag{114}$$

Applying the definition of μ-star-convexity directly bounds

$$-2\eta\langle\nabla f(x_t), x_t - x^*\rangle \le -2\eta(f(x_t) - f(x^*)) - \eta\mu\|x_t - x^*\|^2. \tag{115}$$

Using BG-0 variance condition, unbiasedness, and L-smoothness implication $\|\nabla f(x_t)\|^2 \le 2L(f(x_t) - f(x^*))$ leads to

$$\mathbb{E}_t\left[\|g_t\|^2\right] \le 2L(f(x_t) - f(x^*)) + B^2\|x_t - x_0\|^2 + G^2. \tag{116}$$

Substituting these bounds into the distance recurrence above, we have

$$\mathbb{E}_t\left[\|x_{t+1} - x^*\|^2\right] \le (1-\eta\mu)\|x_t - x^*\|^2 - 2\eta\,(1 - L\eta)\,(f(x_t) - f(x^*)) + \eta^2 B^2\|x_t - x_0\|^2 + \eta^2 G^2. \tag{117}$$

We impose $\eta \le \frac{1}{L}$ such that $1 - L\eta \ge 0$. Using $\|x_t - x_0\|^2 \le 2\|x_t - x^*\|^2 + 2\|x_0 - x^*\|^2$ yields

$$\mathbb{E}_t\left[\|x_{t+1} - x^*\|^2\right] \le \left(1 - \eta\mu + 2\eta^2 B^2\right)\|x_t - x^*\|^2 + 2\eta^2 B^2\|x_0 - x^*\|^2 + \eta^2 G^2. \tag{118}$$

We impose $\eta \le \frac{\mu}{4B^2}$ such that $2\eta^2 B^2 \le \frac{\eta\mu}{2}$. Thus the contraction multiplier is bounded by $1 - \frac{\eta\mu}{2}$. Recursively unrolling from $t = 0$ to K and using the geometric series sum bound $\sum_{j=0}^{K-1}\left(1 - \frac{\eta\mu}{2}\right)^j \le \frac{2}{\eta\mu}$ establishes

$$\mathbb{E}\left[\|x_K - x^*\|^2\right] \le \left(1 - \frac{\eta\mu}{2}\right)^K\|x_0 - x^*\|^2 + \frac{4\eta B^2}{\mu}\|x_0 - x^*\|^2 + \frac{2\eta G^2}{\mu}. \tag{119}$$

Using L-smoothness implication $f(x_t) - f(x^*) \le \frac{L}{2}\|x_t - x^*\|^2$ and μ-star-convexity at initialization $\|x_0 - x^*\|^2 \le \frac{2}{\mu}(f(x_0) - f(x^*))$ we have

$$\mathbb{E}[f(x_K) - f(x^*)] \le \frac{L}{\mu}\left(1 - \frac{\eta\mu}{2}\right)^K(f(x_0) - f(x^*)) + \frac{4L\eta B^2}{\mu^2}(f(x_0) - f(x^*)) + \frac{L\eta G^2}{\mu}. \tag{120}$$

To guarantee the last suboptimality is less than ϵ, we ensure each term in the bound is less than $\frac{\epsilon}{3}$, i.e., $\frac{4L\eta B^2}{\mu^2}(f(x_0) - f(x^*)) \le \frac{\epsilon}{3}$, $\frac{L\eta G^2}{\mu} \le \frac{\epsilon}{3}$, and $K \ge \frac{2}{\eta\mu}\log\left(\frac{3L(f(x_0) - f(x^*))}{\mu\epsilon}\right)$. Since $S = 1$ and $N_t = 1$, the oracle complexity matches the iteration count K evaluating to

$$K = \left\lceil \max\left(\frac{2L}{\mu}\log\left(\frac{3L\Delta}{\mu\epsilon}\right), \frac{8B^2}{\mu^2}\log\left(\frac{3L\Delta}{\mu\epsilon}\right), \frac{24LB^2\Delta}{\mu^3\epsilon}\log\left(\frac{3L\Delta}{\mu\epsilon}\right), \frac{6LG^2}{\mu^2\epsilon}\log\left(\frac{3L\Delta}{\mu\epsilon}\right)\right)\right\rceil, \tag{121}$$

which is readily $\mathcal{O}(\epsilon^{-1}\log(\epsilon^{-1}))$. ■

C.9 Proof of Theorem 8

Proof of Theorem 8. Recall the strongly convex Tikhonov-regularized surrogate for epoch s as $F_\lambda(x) = f(x) + \frac{\lambda}{2}\|x - x_{s-1}\|^2$ with unique minimizer $x_s^\star$. By setting $\beta_t = \lambda\eta$, the inner loop update evaluates to $x_{t+1} = x_t - \eta(g_t + \lambda(x_t - x_{s-1}))$. Applying the strong convexity of F_λ,

$$\mathbb{E}_t\left[\|x_{t+1} - x_s^\star\|^2\right] \le (1 - 2\eta\mu)\|x_t - x_s^\star\|^2 + \eta^2\mathbb{E}_t\left[\|g_t - \bar{g}_t\|^2\right] + 2\eta^2\|\bar{g}_t\|^2 + 2\eta^2\lambda^2\|x_t - x_{s-1}\|^2 \tag{122}$$

where $\bar{g}_t = \mathbb{E}_t[g_t] \in \partial f(x_t)$. For $t = 0$ we have $x_0 = x_{s-1}$. Substituting the variance bound from Assumption 1 and the Lipschitz bound $\|\bar{g}_t\|^2 \le C^2$ yields

$$\mathbb{E}_0\left[\|x_1 - x_s^\star\|^2\right] \le (1 - 2\eta\mu)\|x_{s-1} - x_s^\star\|^2 + \frac{\eta^2}{N_0}\left(B^2\|x_{s-1} - x_0\|^2 + G^2\right) + 2\eta^2 C^2. \tag{123}$$

For $t \ge 1$, expanding $\|x_t - x_0\|^2 \le 3\|x_t - x_s^\star\|^2 + 3\|x_s^\star - x_{s-1}\|^2 + 3\|x_{s-1} - x_0\|^2$ and $\|x_t - x_{s-1}\|^2 \le 2\|x_t - x_s^\star\|^2 + 2\|x_s^\star - x_{s-1}\|^2$ results in

$$\begin{aligned}\mathbb{E}_t\left[\|x_{t+1} - x_s^\star\|^2\right] \le{}& \left(1 - 2\eta\mu + \frac{3\eta^2 B^2}{N} + 4\eta^2\lambda^2\right)\|x_t - x_s^\star\|^2 + \left(\frac{3\eta^2 B^2}{N} + 4\eta^2\lambda^2\right)\|x_s^\star - x_{s-1}\|^2 \\ &+ \frac{3\eta^2 B^2}{N}\|x_{s-1} - x_0\|^2 + \frac{\eta^2 G^2}{N} + 2\eta^2 C^2.\end{aligned} \tag{124}$$

The condition on η ensures that the contraction coefficient is strictly bounded by $1 - \frac{\eta\mu}{2}$. Unrolling this recurrence from $t = 1$ to K for the final distance $e_s = x_K - x_s^\star$ yields

$$\begin{aligned}\mathbb{E}_s\big[\|e_s\|^2\big] \le (1-\eta\mu)^{K-1}\mathbb{E}_0\big[\|x_1 - x_s^\star\|^2\big] + \left(\frac{6\eta B^2}{\mu N} + \frac{8\eta\lambda^2}{\mu}\right)\|x_s^\star - x_{s-1}\|^2 \\ + \frac{6\eta B^2}{\mu N}\|x_{s-1} - x_0\|^2 + \frac{2\eta G^2}{\mu N} + \frac{4\eta C^2}{\mu}.\end{aligned} \tag{125}$$

Recall Lemma 2 which posits that the Moreau envelope is smooth such that using the descent lemma for it at $x_s = x_s^\star + e_s$ yields

$$\mathbb{E}[\varphi_{1/\lambda}(x_s)] \le \mathbb{E}[\varphi_{1/\lambda}(x_{s-1})] - \frac{\mu}{4}\|x_{s-1} - x_s^\star\|^2 + \frac{3\lambda^2 - \rho^2}{2\mu}\mathbb{E}_s\big[\|e_s\|^2\big]. \tag{126}$$

Using the definitions for K and N alongside the substitution of (125) into (126), and applying the Moreau envelope property $\|x_{s-1} - x_s^\star\|^2 = \lambda^{-2}\|\nabla\varphi_{1/\lambda}(x_{s-1})\|^2$ results in

$$\frac{\mu}{4\lambda^2}\mathbb{E}\big[\|\nabla\varphi_{1/\lambda}(x_{s-1})\|^2\big] \le \mathbb{E}[\varphi_{1/\lambda}(x_{s-1})] - \mathbb{E}[\varphi_{1/\lambda}(x_s)] + \Gamma\|x_{s-1} - x_0\|^2 + \Lambda \tag{127}$$

where

$$\Gamma = \frac{3\lambda^2 - \rho^2}{2\mu}B^2\left[(1-\eta\mu)^{K-1}\frac{\eta^2}{N_0} + \frac{6\eta}{\mu N}\right] \tag{128}$$

and

$$\Lambda = \frac{3\lambda^2 - \rho^2}{2\mu}\left[(1-\eta\mu)^{K-1}\left(\frac{\eta^2 G^2}{N_0} + 2\eta^2 C^2\right) + \frac{2\eta G^2}{\mu N} + \frac{4\eta C^2}{\mu}\right]. \tag{129}$$

Using a telescoping sum, we have

$$x_{s-1} - x_0 = \sum_{i=1}^{s-1}(x_i - x_{i-1}). \tag{130}$$

Using $\|\sum_{i=1}^n v_i\|^2 \le n\sum_{i=1}^n \|v_i\|^2$, yields

$$\|x_{s-1} - x_0\|^2 = \left\|\sum_{i=1}^{s-1}(x_i - x_{i-1})\right\|^2 \le (s-1)\sum_{i=1}^{s-1}\|x_i - x_{i-1}\|^2. \tag{131}$$

Taking the total expectation of (131) and summing over the epoch index s from 1 to S yields

$$\sum_{s=1}^{S}\mathbb{E}\big[\|x_{s-1} - x_0\|^2\big] \le \sum_{s=1}^{S}\left((s-1)\sum_{i=1}^{s-1}\mathbb{E}\big[\|x_i - x_{i-1}\|^2\big]\right). \tag{132}$$

Using $\sum_{i=1}^{s-1}\mathbb{E}\big[\|x_i - x_{i-1}\|^2\big] \le \sum_{i=1}^{S}\mathbb{E}\big[\|x_i - x_{i-1}\|^2\big]$ and $\sum_{s=1}^{S}(s-1) = \frac{S(S-1)}{2}$ we have

$$\sum_{s=1}^{S}\mathbb{E}\big[\|x_{s-1} - x_0\|^2\big] \le \frac{S^2}{2}\sum_{i=1}^{S}\mathbb{E}\big[\|x_i - x_{i-1}\|^2\big] = \frac{S^2}{2}\sum_{s=1}^{S}\mathbb{E}\big[\|x_s - x_{s-1}\|^2\big]. \tag{133}$$

by changing the dummy variable index in the upper bound from i to s. Using the bound $\|x_s - x_{s-1}\|^2 \le 2\|x_{s-1} - x_s^\star\|^2 + 2\|e_s\|^2$ and summing the relation established in (125) over S epochs

$$\begin{aligned}\sum_{s=1}^{S}\mathbb{E}\big[\|x_s - x_{s-1}\|^2\big] \le 2\left(1 + (1-\eta\mu)^{K-1} + \frac{6\eta B^2}{\mu N} + \frac{8\eta\lambda^2}{\mu}\right)\sum_{s=1}^{S}\mathbb{E}\big[\|x_{s-1} - x_s^\star\|^2\big] \\ + \frac{4\mu}{3\lambda^2 - \rho^2}\Gamma\sum_{s=1}^{S}\mathbb{E}\big[\|x_{s-1} - x_0\|^2\big] + \frac{4\mu S}{3\lambda^2 - \rho^2}\Lambda.\end{aligned} \tag{134}$$

Substituting the sum derived from rearranging (126) leads to

$$\begin{aligned}\sum_{s=1}^{S}\mathbb{E}\big[\|x_s - x_{s-1}\|^2\big] \le \left(\frac{16}{\mu} + \frac{8\mu}{3\lambda^2 - \rho^2}\right)\Delta \\ + \left(\frac{16}{\mu} + \frac{8\mu}{3\lambda^2 - \rho^2}\right)\Gamma\frac{S^2}{2}\sum_{s=1}^{S}\mathbb{E}\big[\|x_s - x_{s-1}\|^2\big] + \left(\frac{16}{\mu} + \frac{8\mu}{3\lambda^2 - \rho^2}\right)\Lambda S.\end{aligned} \tag{135}$$

Our batch size definitions ensure $\left(\frac{16}{\mu} + \frac{8\mu}{3\lambda^2 - \rho^2}\right) \Gamma S^2 \leq 1$ such that

$$\sum_{s=1}^{S} \mathbb{E}\left[\|x_s - x_{s-1}\|^2\right] \leq 2\left(\frac{16}{\mu} + \frac{8\mu}{3\lambda^2 - \rho^2}\right)\Delta + 2\left(\frac{16}{\mu} + \frac{8\mu}{3\lambda^2 - \rho^2}\right)\Lambda S \tag{136}$$

which yields

$$\sum_{s=1}^{S} \mathbb{E}\left[\|x_{s-1} - x_0\|^2\right] \leq S^2\left(\frac{16}{\mu} + \frac{8\mu}{3\lambda^2 - \rho^2}\right)\Delta + S^3\left(\frac{16}{\mu} + \frac{8\mu}{3\lambda^2 - \rho^2}\right)\Lambda. \tag{137}$$

Substituting this back into (127) yields

$$\begin{aligned}\frac{1}{S}\sum_{s=1}^{S} \mathbb{E}\left[\|\nabla\varphi_{1/\lambda}(x_{s-1})\|^2\right] &\leq \frac{4\lambda^2}{\mu}\left[\frac{\Delta}{S} + \Gamma S\left(\frac{16}{\mu} + \frac{8\mu}{3\lambda^2 - \rho^2}\right)\Delta + \Gamma S^2\left(\frac{16}{\mu} + \frac{8\mu}{3\lambda^2 - \rho^2}\right)\Lambda + \Lambda\right] \\ &\leq \frac{4\lambda^2}{\mu}\left(\frac{2\Delta}{S} + 2\Lambda\right).\end{aligned} \tag{138}$$

By our parameter selection we ensure $\frac{8\lambda^2\Delta}{\mu S} \leq \frac{\epsilon^2}{2}$ and $\frac{8\lambda^2\Lambda}{\mu} \leq \frac{\epsilon^2}{2}$, which establishes we have an ϵ-stationary solution, i.e. $\frac{1}{S}\sum_{s=1}^{S} \mathbb{E}\left[\|\nabla\varphi_{1/\lambda}(x_{s-1})\|^2\right] \leq \epsilon^2$. The total stochastic oracle complexity can be found according to

$$S(N_0 + KN) \leq \left(\frac{16\lambda^2\Delta}{\mu\epsilon^2} + 1\right)\left[N_0 + \left(2 + \frac{1}{\eta\mu}\log\left(\frac{12(3\lambda^2 - \rho^2)}{\mu^2}\right)\right)N\right]. \tag{139}$$

Substituting the definitions for η, N_0, and N leads to a precise, yet ugly expression (arguably, even uglier than (104)). Nonetheless, the oracle complexity is readily $\mathcal{O}(\epsilon^{-6})$. ■