Beyond Bounded Variance: Variance-Reduced Normalized Methods for Nonconvex Optimization under Blum–Gladyshev Noise

Antesh Upadhyay* **Arda Fazla*** **Abolfazl Hashemi***
*School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907
{aantesh, afazla, abolfazl}@purdue.edu

Abstract

Stochastic optimization theory often assumes that gradient noise has uniformly bounded variance. However, this assumption can be inadequate even in simple settings; e.g., it does not hold for linear least squares regression. The Blum–Gladyshev (BG-0) noise model instead allows the stochastic gradient variance to grow quadratically with the distance from the initialization and has recently been identified as the weakest known viable variance assumption. We study nonconvex stochastic optimization under BG-0 noise and prove that normalized momentum converges with one stochastic gradient per iteration, reaching an ε-stationary point with the optimal $\mathcal{O}(\varepsilon^{-6})$ complexity. We then show that, under mean-square smoothness, a variance-reduced normalized STORM method achieves the minimax optimal complexity of $\mathcal{O}(\varepsilon^{-4})$. We further extend our analysis to the symmetric generalized smoothness framework, which captures objectives whose local curvature can scale with the gradient norm. To our knowledge, these are the first convergence guarantees for normalized methods in nonconvex stochastic optimization under BG-0 noise without bounded domains.

1 Introduction

Large-scale nonconvex learning is typically formulated as the stochastic optimization problem

$$\min_{x\in\mathbb{R}^d} f(x), \qquad f(x) := \mathbb{E}_\xi[f(x;\xi)], \tag{1}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is generally nonconvex. The goal is to find an ε-stationary point, i.e., a point x satisfying $\mathbb{E}\|\nabla f(x)\|\le\varepsilon$.

Under standard smoothness and uniformly bounded stochastic noise, the complexity theory for this problem is well developed. In the deterministic setting, gradient descent achieves the optimal $\mathcal{O}(\varepsilon^{-2})$ iteration complexity for smooth nonconvex optimization [34, 35]. In the stochastic setting, SGD achieves the classical $\mathcal{O}(\varepsilon^{-4})$ sample complexity under bounded variance $\mathbb{E}_\xi\left\|\nabla f(x;\xi)-\nabla f(x)\right\|^2\le\sigma^2$, and this rate is unimprovable for general smooth nonconvex stochastic optimization [14, 2]. Under stronger mean-square smoothness or average-smoothness conditions, variance-reduced methods such as SVRG, SAGA, SARAH, SPIDER, SpiderBoost, SNVRG, PAGE, and STORM improve the complexity to the optimal $\mathcal{O}(\varepsilon^{-3})$ [23, 11, 36, 12, 42, 46, 31, 10].

These classical stochastic guarantees typically rely on a uniform bounded-noise model. While mathematically convenient, this assumption can fail even in elementary unconstrained stochastic optimization problems, such as least squares, and in modern deep learning, where the stochastic

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

gradient variance is not uniformly constant but can grow as the iterate moves farther from a reference point. This motivates the BG-0 noise model [4, 15], which is the weakest known viable variance assumption for theoretical analysis [1]. Under BG-0, the variance is allowed to grow quadratically with the distance from the initialization $\mathbb{E}_\xi \left\|\nabla f(x;\xi) - \nabla f(x)\right\|^2 \le B^2\|x - x^0\|^2 + G^2$. The classical bounded variance setting is recovered by taking $B = 0$, while the deterministic case corresponds to $B = G = 0$.

Although BG-0 is more permissive than bounded variance, it creates an inherent feedback mechanism. If the iterates move far from x^0, then the oracle becomes noisier; the larger stochastic noise can then perturb future directions and push the trajectory even farther from x^0. Thus, over unbounded domains, the variance scale $B^2\|x^k - x^0\|^2 + G^2$ cannot be treated as a harmless constant. Recently, Fazla et al. [13] showed that this difficulty is intrinsic: under BG-0 noise, the lower bound for smooth nonconvex optimization worsens from the classical $\Omega(\varepsilon^{-4})$ to $\Omega(\varepsilon^{-6})$, and even under mean-square smoothness, the best possible rate is $\Omega(\varepsilon^{-4})$ rather than the bounded variance $\Omega(\varepsilon^{-3})$ rate [2]. They also proposed PASTA, which matches these limits by combining Halpern-style anchoring [18], Tikhonov regularization, and dynamic batching, connecting BG-0 stochastic approximation with classical fixed-point schemes such as Halpern and Krasnoselskii–Mann iterations [32, 28, 18, 13].

In this paper, we take a complementary viewpoint. Rather than adding an explicit stabilizing anchor or increasing per-iteration batch sizes, we ask whether normalized stochastic methods with momentum or recursive variance reduction can handle BG-0 noise without bounded domains.

Why Normalization? The key observation is that normalization decouples the length of the update from the magnitude of the stochastic gradient estimator. Given an estimator v^k, define[1]

$$d^k = \frac{v^k}{\|v^k\|}, \qquad x^{k+1} = x^k - \gamma d^k.$$

Then, $\|x^{k+1} - x^k\| = \gamma\|d^k\| \le \gamma$, regardless of the size of $\|v^k\|$. Consequently,

$$\|x^k - x^0\| \le \sum_{t=0}^{k-1} \|x^{t+1} - x^t\| \le k\gamma.$$

Thus, along a normalized trajectory, the BG-0 variance scale satisfies $\left(B^2\|x^k - x^0\|^2 + G^2\right)^{1/2} \le G + Bk\gamma$. Normalization, therefore, transforms the potentially uncontrolled effect of large stochastic gradient magnitudes into controlled trajectory growth. In contrast, PASTA [13] controls the BG-0 variance via dynamic batching. There, the batch size must scale with the upper bound on the local variance, i.e., $N_k \propto \frac{B^2\|x^k - x^0\|^2 + G^2}{\sigma^2}$, up to problem-dependent accuracy and stepsize factors. Such a rule requires knowledge, or at least reliable estimates, of the variance-growth constants B and G, which may be difficult to obtain in practice.

Generalized Smoothness. A second challenge is that many learning objectives are not globally smooth, i.e., their local smoothness can grow with the gradient norm [22, 44]. This is a natural setting for normalized updates [7]. We therefore study BG-0 noise under the deterministic class $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ and its stochastic analogue $\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ from Chen et al. [7], for $\alpha \in (0, 1]$. The class $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ contains standard smooth functions ($\mathcal{L}$), asymmetric generalized smooth functions ($\mathcal{L}^*_{\mathrm{asym}}$), and Hessian-based generalized smooth functions ($\mathcal{L}^*_{\mathrm{H}}$), and also includes high-order polynomial and exponential-type objectives [7, 44, 29]. Appendix A summarizes these definitions.

Prior work shows that, under suitable bounded or relative variance assumptions, generalized smooth nonconvex optimization can be as efficient as smooth nonconvex optimization: normalized gradient methods recover the deterministic $\mathcal{O}(\varepsilon^{-2})$ rate, while SPIDER-type variance reduction recovers the stochastic $\mathcal{O}(\varepsilon^{-3})$ rate under expected generalized smoothness [7, 12]. These results, however, do not address BG-0 noise. In this setting, the noise level is distance-dependent, while under generalized smoothness, the local curvature is gradient-dependent. This creates a new interplay between trajectory growth and gradient growth that is absent from the bounded variance setting. It is therefore unclear a priori whether the principle that generalized smooth optimization is "as efficient as" smooth optimization continues to hold under BG-0 noise. This paper studies exactly this interaction.

[1] Throughout, we use the convention $v/\|v\| = 0$ when $v = 0$, so that the update is well defined in degenerate cases.

Table 1: Comparison with representative works under different notions of smoothness ($\mathcal{L}, \mathcal{L}^*_{\mathrm{H}}, \mathcal{L}^*_{\mathrm{asym}}, \mathcal{L}^*_{\mathrm{sym}}(\alpha), \mathsf{MSS}, \mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$). Complexity is measured in oracle calls to find an ε-stationary point.

Work	Smoothness class	BG-0?	Method	Oracle complexity
Cutkosky and Mehta [9]	$\mathcal{L}$	✗	NSGDM	$\mathcal{O}(\varepsilon^{-4})$
Zhang et al. [44]	$\mathcal{L}^*_{\mathrm{H}}$	✗	clipped SGD	$\mathcal{O}(\varepsilon^{-4})$
Reisizadeh et al. [38]	$\mathcal{L}^*_{\mathrm{asym}}$	✗	SPIDER + clipping	$\mathcal{O}(\varepsilon^{-3})$
Chen et al. [7]	$\mathcal{L}^*_{\mathrm{sym}}(\alpha), \mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$	✗	β-normalized GD; SPIDER	Det. ($\mathcal{L}^*_{\mathrm{sym}}(\alpha)$): $\mathcal{O}(\varepsilon^{-2})$; Stoch. ($\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$): $\mathcal{O}(\varepsilon^{-3})$
Khirirat et al. [26]	$\mathcal{L}^*_{\mathrm{sym}}(1)$	✗	NSGDM	$\mathcal{O}(\varepsilon^{-4})$
Fazla et al. [13]	$\mathcal{L}$, MSS	✓	PASTA	$\mathcal{L}$: $\Theta(\varepsilon^{-6})$; MSS: $\Theta(\varepsilon^{-4})$
This paper	$\mathcal{L}, \mathcal{L}^*_{\mathrm{sym}}(\alpha)$	✓	NSGDM	$\mathcal{O}(\varepsilon^{-6})$
This paper	MSS, $\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$	✓	NSTORM	MSS : $\mathcal{O}(\varepsilon^{-4})$; $\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$: $\mathcal{O}(\varepsilon^{-(4+\alpha)})$; $\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(1)$: $\mathcal{O}(\varepsilon^{-5})$

Note. When $B = 0$, our bounds recover the standard bounded variance rates: $\mathcal{O}(\varepsilon^{-4})$ for normalized momentum, $\mathcal{O}(\varepsilon^{-3})$ for variance reduction, and $\mathcal{O}(\varepsilon^{-2})$ in the deterministic case, for standard and generalized smoothness classes.

1.1 Contributions

We develop a convergence theory for normalized stochastic methods under BG-0 noise across standard and generalized smoothness regimes. The results answer four questions.

Q1. Can single-sample normalized momentum converge under BG-0?

Yes. We prove that normalized stochastic gradient descent with momentum (NSGDM) converges under BG-0 using one fresh stochastic gradient sample per iteration. The method does not require bounded domains, bounded stochastic gradients, bounded variance, dynamic batching, or explicit anchoring. Under BG-0 noise, NSGDM achieves an oracle complexity of $\mathcal{O}(\varepsilon^{-6})$. Thus, normalization and momentum provide an implicit stabilization mechanism: normalization controls trajectory growth, while momentum stabilizes the noisy update direction.

Q2. Does replacing standard smoothness with generalized smoothness worsen the oracle complexity of NSGDM under BG-0?

No. For NSGDM, generalized smoothness is rate-neutral at the level of the ε-complexity exponent: under $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$-type smoothness, the stochastic first-order oracle (SFO) complexity remains $\mathcal{O}(\varepsilon^{-6})$, up to constants depending on the generalized smoothness parameters.

Q3. Can additional stochastic regularity recover the optimal BG-0 rate?

Yes. Mean-square smoothness (MSS) provides additional stochastic regularity by controlling differences of stochastic gradients $\mathbb{E}_\xi \left\|\nabla f(y;\xi) - \nabla f(x;\xi)\right\|^2 \le L^2\|y-x\|^2$. Under this stronger condition, normalized STORM (NSTORM) improves over the $\mathcal{O}(\varepsilon^{-6})$ single-sample momentum rate. With sharp initialization, meaning a one-time batch chosen so that the initial estimator-error contribution does not dominate the convergence rate, it achieves the optimal oracle complexity of $\mathcal{O}(\varepsilon^{-4})$.

Q4. Does the optimal NSTORM rate persist under expected generalized smoothness?

Not exactly. Under mean-square smoothness, NSTORM recovers the optimal BG-0 rate $\mathcal{O}(\varepsilon^{-4})$. Under $\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ generalized smoothness, however, the estimator-difference recursion depends on gradient-dependent smoothness terms, which interact with the distance-dependent BG-0 variance. As a result, NSTORM achieves $\mathcal{O}(\varepsilon^{-(4+\alpha)})$ for $\alpha \in (0,1)$ and $\mathcal{O}(\varepsilon^{-5})$ for $\alpha = 1$. Thus, generalized smoothness is free for NSGDM at the level of the ε-exponent, but it induces a quantifiable α-dependent price for variance-reduced NSTORM under BG-0.

We also provide synthetic experiments on generalized smooth objectives from Chen et al. [7] under a BG-0 oracle, comparing normalized momentum and normalized variance reduction with dynamic batching baselines. Finally, Table 1 summarizes the comparison with prior work, emphasizing the combination of BG-0 noise, normalized methods, and smoothness regimes studied here.

2 Related Work

From bounded variance to Blum–Gladyshev noise. Standard nonconvex guarantees usually assume uniform bounds on variance or subgradients, a convenient but often unrealistic simplification [14, 33, 5, 3, 37, 39, 21]. More recent research explores noise models where variance scales with

the gradient norm or iterate location [16, 24, 25, 20, 17, 1]. A key example is the Blum–Gladyshev condition [4, 15], which allows variance to grow with the squared distance from a reference point. This is particularly difficult in nonconvex settings, as there is no natural "pull" to keep iterates from drifting away. While recent work like the PASTA framework [13] uses anchoring, regularization, and dynamic batching to stabilize this drift, achieving sharp rates of $\Theta(\varepsilon^{-6})$ or $\Theta(\varepsilon^{-4})$, our approach takes a different path. We utilize normalized updates to restrict movement, combined with momentum and recursive variance reduction to keep our stochastic directions stable.

Normalization and clipping for nonuniform smoothness. Clipping and normalization are widely used to stabilize first-order methods when gradients or local curvature can be large. Under (L_0, L_1)-type smoothness, clipped methods have been analyzed in deterministic and stochastic settings [44, 27], with extensions to momentum, variance reduction, and adaptive stepsizes [43, 38, 41, 30, 40]. Normalized methods, including normalized gradient descent, generalized SignSGD, and normalized momentum, provide a related way to control large updates [45, 19, 8]. Khirirat et al. [26] study normalized error-feedback under $\mathcal{L}^*_{\rm sym}(1)$ in distributed optimization. These results, however, rely on a stronger notion of variance bounds. In contrast, we study normalized momentum and recursive variance reduction under distance-dependent BG-0 noise for the generalized smooth classes $\mathcal{L}^*_{\rm sym}(\alpha)$ and $\mathbb{E}\mathcal{L}^*_{\rm sym}(\alpha)$.

Momentum and variance reduction. Momentum and recursive variance reduction stabilize stochastic gradient estimates. Momentum smooths sample noise [9], while methods such as SARAH, SPIDER, SpiderBoost, PAGE, and STORM use recursive estimators to obtain sharper control [36, 12, 42, 31, 10]. Under bounded variance and mean-square smoothness, these methods improve the stochastic nonconvex complexity from $\mathcal{O}(\varepsilon^{-4})$ to $\mathcal{O}(\varepsilon^{-3})$. Under BG-0 noise, this acceleration is no longer attainable in general; the optimal MSS rate becomes $\mathcal{O}(\varepsilon^{-4})$ [13]. PASTA reaches this rate using dynamic batching with PAGE-type variance reduction. In contrast, our normalized STORM analysis recovers the same $\mathcal{O}(\varepsilon^{-4})$ rate with single-sample-transition updates and sharp initialization. Under $\mathbb{E}\mathcal{L}^*_{\rm sym}(\alpha)$, the interaction between recursive estimation and trajectory-dependent variance yields $\mathcal{O}(\varepsilon^{-(4+\alpha)})$, with $\mathcal{O}(\varepsilon^{-5})$ at $\alpha = 1$.

3 Preliminaries

Notation. All norms $\|\cdot\|$ are Euclidean norms, and $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product. We use $\mathcal{O}(\cdot)$ to hide numerical constants independent of the target accuracy ε. For a random variable X, $\mathbb{E}[X]$ denotes expectation over all algorithmic randomness. We let $\mathcal{F}_k$ be the filtration generated by the algorithm up to iteration k, and write $\mathbb{E}_k[X] := \mathbb{E}[X \mid \mathcal{F}_k]$. We use $\mathbb{E}_\xi$ in assumptions for expectations over an oracle sample at fixed query points. Throughout, we assume that f is bounded below, i.e., $f^{\rm inf} := \inf_{x\in\mathbb{R}^d} f(x) > -\infty$, and define the initial optimality gap $\Delta := f(x^0) - f^{\rm inf}$ as well as the noise at initialization as $\sigma_0^2 := \mathbb{E}_\xi\|\nabla f(x^0;\xi) - \nabla f(x^0)\|^2$. After $K+1$ iterations, the output $\widehat{x}$ is sampled uniformly from $\{x^0, x^1, \ldots, x^K\}$.

Problem Formulation. We consider the problem in (1). The stochastic first-order oracle returns $\nabla f(x;\xi)$ at a queried point x. We measure complexity by the number of SFO calls. As mentioned previously, the goal is to find an ε-stationary point, i.e., a point x such that $\mathbb{E}\|\nabla f(x)\| \le \varepsilon$.

Assumptions. We first introduce the noise model and then state the smoothness assumptions used in our analysis.

Assumption BG-0 (BG-0 stochastic oracle). *For every $x \in \mathbb{R}^d$, the stochastic gradient oracle is unbiased, i.e., $\mathbb{E}_\xi[\nabla f(x;\xi)] = \nabla f(x)$. Moreover, there exist constants $B, G \ge 0$ such that*

$$\mathbb{E}_\xi \left\|\nabla f(x;\xi) - \nabla f(x)\right\|^2 \le B^2\|x - x^0\|^2 + G^2.$$

When $B = 0$, Assumption BG-0 reduces to the classical bounded variance condition. When $B = G = 0$, the oracle is deterministic.

We pair this noise model with several smoothness conditions of increasing generality. The first two are standard: L_0-smoothness controls population gradients [14], while mean-square smoothness controls the stochastic gradient differences [10, 31].

Assumption 2 (L_0-smoothness of f). *There exists $L_0 > 0$ such that for all $x, y \in \mathbb{R}^d$,*

$$\|\nabla f(y) - \nabla f(x)\| \le L_0\|y - x\|.$$

Assumption 3 (Mean-square smoothness)**.** *There exists $L > 0$ such that for all $x, y \in \mathbb{R}^d$,*

$$\mathbb{E}_\xi \|\nabla f(y;\xi) - \nabla f(x;\xi)\|^2 \le L^2 \|y - x\|^2.$$

The next two assumptions extend smoothness to the generalized setting of Chen et al. [7], where the local smoothness parameter is allowed to grow with the gradient norm. The deterministic class $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ governs the normalized momentum analysis, while the expected class $\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ governs the normalized STORM analysis.

Assumption 4 ($\mathcal{L}^*_{\mathrm{sym}}(\alpha)$-type generalized smoothness)**.** *Let $\alpha \in (0, 1]$. We say that $f \in \mathcal{L}^*_{\mathrm{sym}}(\alpha)$ if $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable and there exist constants $L_0, L_1 > 0$ such that, for all $x, y \in \mathbb{R}^d$,*

$$\|\nabla f(y) - \nabla f(x)\| \le \left(L_0 + L_1 \max_{\theta \in [0,1]} \|\nabla f(x_\theta)\|^\alpha \right) \|y - x\|, \text{ where } x_\theta := \theta y + (1-\theta)x.$$

The generalized-smoothness framework accommodates objectives that are not globally smooth. For example, $f(x) = |x|^3$ does not have a globally Lipschitz gradient but satisfies $\mathcal{L}^*_{\mathrm{sym}}(1/2)$, illustrating the broader class of objectives covered in our analysis. Moreover, for twice-differentiable functions, the endpoint $\mathcal{L}^*_{\mathrm{sym}}(1)$ is equivalent, up to constants, to the Hessian-growth condition $\|\nabla^2 f(x)\| \le L_0 + L_1 \|\nabla f(x)\|$ [44, 7].

Assumption 5 (Expected α-symmetric generalized smoothness)**.** *Let $\alpha \in (0, 1]$. We say that $f \in \mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ if there exist constants $L_0, L_1 > 0$ such that, for all $x, y \in \mathbb{R}^d$, where $x_\theta := \theta y + (1-\theta)x$, we have*

$$\mathbb{E}_\xi \big\| \nabla f(y;\xi) - \nabla f(x;\xi) \big\|^2 \le \|y - x\|^2 \mathbb{E}_\xi \big[\big(L_0 + L_1 \max_{\theta \in [0,1]} \|\nabla f(x_\theta;\xi)\|^\alpha \big)^2 \big].$$

Formally, setting $L_1 = 0$ in the deterministic and expected generalized smoothness conditions recovers L_0-smoothness and mean-square smoothness, respectively. At $\alpha = 1$, the conditions correspond to symmetric (L_0, L_1)-generalized smoothness and its expected stochastic counterpart. The latter controls same-sample stochastic-gradient differences, as required by the NSTORM recursion. The expected condition is more delicate under BG-0 noise because it involves sample-gradient moments, which must be controlled through the distance-dependent variance scale. This interaction is the source of the α-dependent price in our variance-reduced rates. Existing empirical studies motivate population-level generalized smoothness along neural-network training trajectories [44], while direct large-scale verification of the stronger expected stochastic condition remains limited.

4 Normalized Momentum under BG-0

We first consider normalized stochastic gradient descent with momentum, denoted by NSGDM. Given $x^0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, momentum parameter $\eta \in (0, 1]$, and horizon $K \ge 0$, initialize $v^0 = \nabla f(x^0; \xi^0)$. For $k = 0, 1, \ldots, K$, NSGDM performs

$$\text{NSGDM}: \qquad x^{k+1} = x^k - \gamma \frac{v^k}{\|v^k\|}; \qquad v^{k+1} = (1-\eta) v^k + \eta \nabla f(x^{k+1}; \xi^{k+1}), \tag{2}$$

where the v^{k+1} update is used only for $k = 0, 1, \ldots, K-1$. The samples are fresh and conditionally independent given the past, and the stochastic oracle satisfies Assumption BG-0.

Theorem 1 (Normalized momentum under BG-0)**.** *Consider problem* (1) *and the* NSGDM *update in* (2)*. Suppose that f is bounded below and that Assumption* BG-0 *holds. Let $\eta = (K+1)^{-2/3}$, $\gamma = \gamma_0 (K+1)^{-5/6}$, where $\gamma_0 > 0$. Then the output $\widehat{x}$, sampled uniformly from $\{x^0, x^1, \ldots, x^K\}$, satisfies the following.*

*(i) L_0-**smooth case.** Suppose Assumption 2 holds. Then, for every $\gamma_0 > 0$,*

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \left(\frac{2\Delta}{\gamma_0} + (18 L_0 + 2B)\gamma_0 \right) (K+1)^{-1/6} + \mathcal{O}\left((\sigma_0 + G)(K+1)^{-1/3} \right).$$

*(ii) $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ **generalized smooth case,** $\alpha \in (0, 1)$**.** Suppose Assumption 4 holds for some fixed $\alpha \in (0, 1)$. Then, for every $0 < \gamma_0 \le 1$,*

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \left(\frac{2\Delta}{\gamma_0} + (\Lambda_\alpha + 2B)\gamma_0 \right) (K+1)^{-1/6} + \mathcal{O}\left((\sigma_0 + G)(K+1)^{-1/3} \right),$$

where $\Lambda_\alpha < \infty$ depends only on L_0, L_1, and α, with its explicit definition given in Appendix B.

(iii) $\mathcal{L}^*_{\mathrm{sym}}(1)$ ***generalized smooth case.*** *Suppose Assumption 4 holds with* $\alpha = 1$. *Then, for every* $0 < \gamma_0 \le 1/(8L_1)$,

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \left(\frac{2\Delta}{\gamma_0} + 32L_1\Delta + (36L_0 + 4B)\gamma_0\right)(K+1)^{-1/6} + \mathcal{O}\left((\sigma_0 + G)(K+1)^{-1/3}\right).$$

Consequently, all three smoothness regimes satisfy $\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}((K+1)^{-1/6})$, *giving* $\mathcal{O}(\varepsilon^{-6})$ *SFO complexity for* NSGDM.

Discussion on Theorem 1. We highlight several aspects of this result; the full proof is in Appendix B.

• *Generalized smoothness is rate-neutral for* NSGDM. Under BG-0 noise, the oracle complexity exponent $\mathcal{O}(\varepsilon^{-6})$ is identical across all three regimes: standard L_0-smoothness, $\mathcal{L}^*_{\mathrm{sym}}(1)$ generalized smoothness, and $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ generalized smoothness for any fixed $\alpha \in (0,1)$. The generalized smoothness parameters affect only the leading constants, not the ε-exponent. This is consistent with the principle articulated by Chen et al. [7] for the deterministic setting, that generalized smooth nonconvex optimization is "as efficient as" standard smooth nonconvex optimization, and extends it to the stochastic momentum regime under BG-0 noise.

• *Novelty of the* $\alpha \in (0,1)$ *analysis.* To our knowledge, the $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ analysis for NSGDM with $\alpha \in (0,1)$ is new even under bounded variance. Prior stochastic results under α-symmetric generalized smoothness are limited to the two-loop SPIDER estimator under $\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ [7], which controls variance through periodic batch corrections. The single-sample momentum recursion requires us to control the estimator error over time. After unrolling the recursion, this error separates into gradient-drift terms and stochastic noise terms. The key technical challenge is that the drift bound under $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ yields terms proportional to $\gamma\|\nabla f(x^k)\|^\alpha$, which, unlike the $\alpha = 1$ case cannot be converted into suboptimality gaps $(f(x^k) - f^{\inf})$ via the gradient-growth inequality [7, 26]: when $\alpha = 1$, the bound

$$\|\nabla f(x)\| \le 8L_1(f(x) - f^{\inf}) + L_0/L_1$$

directly absorbs $\|\nabla f(x^k)\|$ into $(f(x^k) - f^{\inf})$, but for $\alpha \in (0,1)$ the fractional power breaks this conversion. Instead, we need to utilize Young's inequality, giving us a free parameter to absorb the $\|\nabla f(x^k)\|^\alpha$ contribution into a fraction of the $\|\nabla f(x^k)\|$ descent budget, at the cost of producing remainder terms that scale as $\gamma^{1/(1-\alpha)}$. These remainders vanish faster than γ and do not affect the dominant rate.

• *Recovery of known rates.* When $B = 0$, the BG-0 condition reduces to bounded variance. In this case, choosing $\eta = (K+1)^{-1/2}$ and $\gamma = \gamma_0(K+1)^{-3/4}$ recovers the $\mathcal{O}(\varepsilon^{-4})$ rate of Cutkosky and Mehta [9] for standard smoothness and Khirirat et al. [26] for $\mathcal{L}^*_{\mathrm{sym}}(1)$, and establishes the same rate for $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ with $\alpha \in (0,1)$. Additionally when $G = 0$ as well, the oracle is deterministic, and with $\eta = 1$ recovers the $\mathcal{O}(\varepsilon^{-2})$ rate of Chen et al. [7] for normalized GD.

• *Schedule comparison.* Under bounded variance, normalized momentum uses the standard schedule $\eta = (K+1)^{-1/2}$ and $\gamma = \gamma_0(K+1)^{-3/4}$, leading to the rate $(K+1)^{-1/4}$, or equivalently $\mathcal{O}(\varepsilon^{-4})$ SFO complexity [9, 26]. The faster decay of $\gamma = \gamma_0(K+1)^{-5/6}$ under BG-0 reflects the growing noise scale $G + Bk\gamma$. Likewise, $\eta = (K+1)^{-2/3}$ is chosen smaller to control the stochastic error term $\sqrt{\eta}(G + BK\gamma)$. This changes the rate from the bounded variance rate $(K+1)^{-1/4}$ to the BG-0 rate $(K+1)^{-1/6}$, matching the lower-bound degradation from $\Omega(\varepsilon^{-4})$ to $\Omega(\varepsilon^{-6})$ [13]. Additionally, when $\eta = 1$, meaning momentum is absent, the temporal averaging due to it disappears, and the resulting noise terms prevent our analysis from yielding a vanishing stationarity guarantee under general BG-0 noise. More precisely, under the BG-0 schedule, the trajectory-dependent noise contributes at order $B\gamma_0 T^{-1/6}$, whereas the constant-noise contribution decays faster, at order $GT^{-1/3}$. Thus, the trajectory-dependent noise is co-leading with the optimization and momentum-tracking terms, while the direct constant-noise contribution is lower order.

5 Normalized STORM under BG-0

We next study the normalized STORM recursion under BG-0 noise. Unlike NSGDM, which averages stochastic gradients directly, NSTORM uses a recursive estimator based on the difference of two stochastic gradients evaluated with the same fresh sample. Controlling these same-sample gradient

differences requires sample-level regularity, provided here by mean-square smoothness or expected generalized smoothness, whereas NSGDM relies only on population-level smoothness.

Fix $x^0 \in \mathbb{R}^d$, a stepsize $\gamma > 0$, a STORM parameter $\eta \in (0, 1]$, and a horizon $K \geq 0$. Let $N_{\text{init}} \geq 1$ be an initialization batch size, and initialize $v^0 = \frac{1}{N_{\text{init}}} \sum_{i=1}^{N_{\text{init}}} \nabla f(x^0; \xi_i^{\text{init}})$, where the samples $\{\xi_i^{\text{init}}\}_{i=1}^{N_{\text{init}}}$ are independent. For $k = 0, 1, \ldots, K$, NSTORM performs

$$\text{NSTORM}: \quad x^{k+1} = x^k - \gamma \frac{v^k}{\|v^k\|}; \; v^{k+1} = \nabla f(x^{k+1}; \xi^{k+1}) + (1-\eta)\left(v^k - \nabla f(x^k; \xi^{k+1})\right). \tag{3}$$

The sample ξ^{k+1} is fresh and conditionally independent given the past, and is evaluated at both x^k and x^{k+1}. Thus, each transition uses two stochastic gradient evaluations. The initialization batch only controls the initial error $\mathbb{E}\|v^0 - \nabla f(x^0)\|$, and N_{init} is specified for the smoothness regime.

NSTORM parameter schedules. Let $T := K + 1$. We use the following parameter schedules in Theorem 2.

Setting	η	γ	N_{init}	λ
MSS	T^{-1}	$\gamma_0 T^{-3/4}$	$\lceil T^{1/2} \rceil$	–
$\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$, $\alpha \in (0, 1)$	$\eta_0 T^{-\frac{4}{4+\alpha}}$	$\gamma_0 T^{-\frac{3+\alpha}{4+\alpha}}$	$\lceil T^{\frac{2(1-\alpha)}{4+\alpha}} \rceil$	$\lambda_0 T^{-\frac{1-\alpha}{4+\alpha}}$
$\mathbb{E}\mathcal{L}^*_{\text{sym}}(1)$	$T^{-4/5}$	$\gamma_0 T^{-4/5}$	1	–

Here $\eta_0 \in (0, 1]$, while $\gamma_0 > 0$ and $\lambda_0 > 0$ are fixed constants satisfying the smoothness-dependent restrictions below. These schedules do not require knowledge of the BG-0 parameters B or G. The parameters η, γ, and N_{init} are algorithmic schedules, whereas λ is used only in the analysis.

Theorem 2 (NSTORM under BG-0)**.** *Consider problem* (1) *and the* NSTORM *update in* (3)*. Suppose that f is bounded below and Assumption* BG-*0 holds. Let $T := K + 1$ and run* NSTORM *with the corresponding parameter schedule above. Then the output $\widehat{x}$, sampled uniformly from $\{x^0, x^1, \ldots, x^K\}$, satisfies the following.*

(i) ***Mean-square smoothness.*** *Under Assumption 3 with constant L, for any $\gamma_0 > 0$,*

$$\mathbb{E}\|\nabla f(\widehat{x})\| \leq \left(\frac{\Delta}{\gamma_0} + 2\sigma_0 + \frac{5}{2} L\gamma_0 + 2B\gamma_0\right) T^{-1/4} + 2G\, T^{-1/2}.$$

Consequently, $\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/4})$, achieving the minimax optimal $\mathcal{O}(\varepsilon^{-4})$ SFO complexity under BG-*0 and mean-square smoothness.*

(ii) $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$ ***generalized smooth case with*** $\alpha \in (0, 1)$*. Under Assumption 5 for fixed $\alpha \in (0, 1)$, define $K_1 := 2^{\frac{2-\alpha}{1-\alpha}} L_1$. Let $\eta_0 \in (0, 1]$ and $\gamma_0, \lambda_0 > 0$ satisfy $2^{\alpha/2} K_1 \lambda_0 \gamma_0 \leq 1$, $2^{\frac{6+\alpha}{2}} K_1 \lambda_0 \frac{\gamma_0}{\eta_0} \leq 1$. Then*

$$\mathbb{E}\|\nabla f(\widehat{x})\| \leq \left(\frac{4\Delta}{\gamma_0} + \frac{8\sigma_0}{\eta_0} + 8\sqrt{\eta_0}(G + B\gamma_0) + \frac{16\gamma_0}{\sqrt{\eta_0}} \Phi_\alpha\right) T^{-\frac{1}{4+\alpha}},$$

where $\Phi_\alpha < \infty$ is independent of T and collects the remaining dependence on L_0, L_1, α, γ_0, λ_0, B^α, and G^α. Its explicit definition is provided in Appendix C.6.
Consequently, $\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/(4+\alpha)})$, giving an $\mathcal{O}(\varepsilon^{-(4+\alpha)})$ SFO complexity.

(iii) $\mathbb{E}\mathcal{L}^*_{\text{sym}}(1)$ ***generalized smooth case****. Under Assumption 5 with $\alpha = 1$, for any $0 < \gamma_0 \leq \left(16\sqrt{2e^{3/4}}\, L_1\right)^{-1}$,*

$$\mathbb{E}\|\nabla f(\widehat{x})\| \leq \left(\frac{4\Delta}{\gamma_0} + 8\sigma_0 + 10B\gamma_0\right) T^{-1/5} + (23L_0\gamma_0 + 10G) T^{-2/5}.$$

Consequently, $\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/5})$, giving an $\mathcal{O}(\varepsilon^{-5})$ SFO complexity.

Discussion on Theorem 2. We now discuss several aspects of Theorem 2. The complete details of the proof, including the proofs on bounded variance and deterministic cases, are provided in Appendix C.

• *Optimality under mean-square smoothness.* Under Assumption 3, i.e., MSS and sharp initialization, NSTORM achieves $\mathcal{O}(\varepsilon^{-4})$ SFO complexity which is minimax optimal in this setting [13]. The key structural reason is that MSS gives $\mathbb{E}_\xi\|\nabla f(y;\xi) - \nabla f(x;\xi)\|^2 \le L^2\|y-x\|^2$, with no dependence on gradient magnitudes. Hence, the STORM centered-noise term satisfies

$$(\mathbb{E}[\|Z_{k+1}\|^2\big|\mathcal{F}_k])^{\frac{1}{2}} \le L\gamma, \text{ with } Z_{k+1} = \nabla f(x^{k+1};\xi^{k+1}) - \nabla f(x^k;\xi^{k+1}) - (\nabla f(x^{k+1}) - \nabla f(x^k)).$$

So, the estimator recursion is decoupled from the gradient sequence. Furthermore, sharp initialization removes the initial-error bottleneck $b_0/(\eta T)$, allowing the schedule $\eta = (K+1)^{-1}$ and $\gamma = \gamma_0(K+1)^{-3/4}$ and yielding the rate $(K+1)^{-1/4}$.

• *Gradient feedback and the α-dependent price under expected generalized smoothness.* Under $\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$, the STORM centered-difference bound on Z_{k+1} depends on the sample-gradient moment $\mathbb{E}_\xi\|\nabla f(x^k;\xi)\|^\alpha$ via the expected reduction of Chen et al. [7] (see Proposition 4, item (1)–(2)). Under bounded variance, this moment is controlled by a fixed noise term, but under BG-0 it couples a gradient-dependent term with trajectory-dependent variance, the main reason the analysis of Chen et al. [7] cannot be applied directly. To handle this coupling, we use Young's inequality to linearize the fractional-gradient term:

$$\|\nabla f(x^k)\|^\alpha \le \lambda\|\nabla f(x^k)\| + c_\alpha \lambda^{-\alpha/(1-\alpha)},$$

yielding an estimator bound of the form

$$(\mathbb{E}[\|Z_{k+1}\|^2 \mid \mathcal{F}_k])^{1/2} \le \gamma(a + h\|\nabla f(x^k)\|)$$

with $h \propto \lambda$. The gradient-dependent part creates a feedback loop: large gradients increase the estimator error, which in turn weakens the descent guarantee. Our proof separates the estimator error into a gradient-independent part, controlled by the NSTORM estimator recursion, and the gradient-dependent part, which is absorbed into the descent inequality via the stepsize conditions. Choosing λ to decay with the horizon weakens this feedback and leads to the $\mathcal{O}(\varepsilon^{-(4+\alpha)})$ rate for $\alpha \in (0,1)$. At $\alpha = 1$, the term $\|\nabla f(x^k)\|^\alpha = \|\nabla f(x^k)\|$ is already linear, so there is no fractional power to absorb via Young's inequality; the feedback coefficient h is then a fixed constant, yielding the slower $\mathcal{O}(\varepsilon^{-5})$ rate.

• *Dominant stochastic contributions.* Under mean-square smoothness, the leading error contributions include the initialization noise level σ_0 and the trajectory-dependent noise coefficient B, both appearing at order $T^{-1/4}$, whereas the direct G-dependent contribution decays as $T^{-1/2}$. Under expected α-symmetric generalized smoothness, the coupling between distance-dependent variance and gradient-dependent smoothness introduces an additional leading contribution involving $L_1 B^\alpha$. Hence, several terms can contribute at the same leading order, rather than there being a unique dominant term. The corresponding constant-noise contributions decay faster under the BG-0 schedules.

• *Sharp initialization and its dependence on the smoothness regime.* The initialization batch size N_{init} is determined by a single mechanism: whether the schedule η suppresses the initial-error contribution $b_0/(\eta T)$ at the target convergence rate. Under MSS, the clean STORM recursion permits $\eta = T^{-1}$, which yields $b_0/(\eta T) = \mathcal{O}(b_0)$, a T-independent term, forcing $N_{\mathrm{init}} = \mathcal{O}(T^{1/2})$ to suppress b_0 to $\mathcal{O}(T^{-1/4})$. Under $\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ with $\alpha \in (0,1)$, the gradient-dependent feedback requires $\eta = \eta_0 T^{-4/(4+\alpha)}$, giving $b_0/(\eta T) = \mathcal{O}(b_0 T^{-\alpha/(4+\alpha)})$, which decays more slowly than the target rate $T^{-1/(4+\alpha)}$. This necessitates $N_{\mathrm{init}} = \mathcal{O}(T^{2(1-\alpha)/(4+\alpha)})$, whose exponent decreases to zero as $\alpha \to 1^-$ and grows toward $1/2$ as $\alpha \to 0^+$. At $\alpha = 1$, the schedule $\eta = T^{-4/5}$ makes $b_0/(\eta T) = \mathcal{O}(T^{-1/5})$ match the target rate, so one-sample initialization is sufficient. Thus, sharp initialization is needed whenever the schedule is aggressive relative to the target rate.

• *Comparison with prior variance-reduced methods.* Chen et al. [7] prove that SPIDER achieves $\mathcal{O}(\varepsilon^{-3})$ under $\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ in the bounded variance setting. Reisizadeh et al. [38] obtain the same rate under asymmetric generalized smoothness using clipped SPIDER. Our setting differs in two ways: the estimator is single-loop STORM rather than two-loop SPIDER, and the noise is distance-dependent rather than bounded. When $B = 0$, the BG-0 trajectory-growth terms disappear, and the bounded variance specialization recovers the standard $\mathcal{O}(\varepsilon^{-3})$ rate in our setting as well.

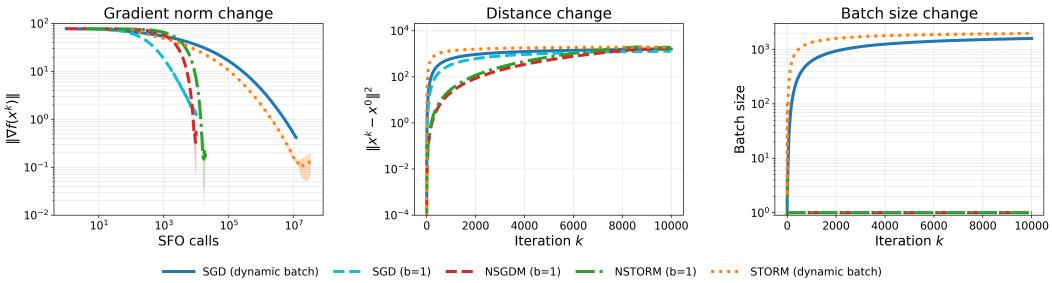


Figure 1: *Phase retrieval experiment under the* BG-*0 oracle* ($\alpha = 2/3$). Gradient norm vs. SFO calls (left), drift $\|x^k - x^0\|^2$ (center), and batch size (right).

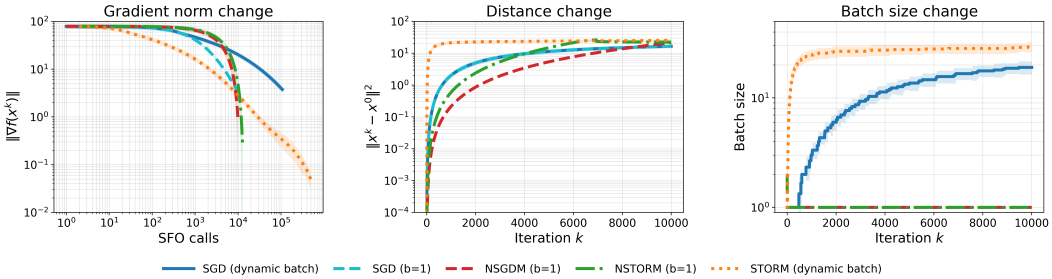


Figure 2: *Cubic polynomial under the* BG-*0 oracle* ($\alpha = 1/2$). Same layout as Figure 1. NSTORM reaches a lower gradient norm than NSGDM, reflecting the benefit of variance reduction.

6 Experiments

We evaluate the methods analyzed in this paper on two experiments whose deterministic objectives satisfy α-symmetric generalized smoothness. In both cases, we use objectives from Chen et al. [7] and apply a stochastic BG-0 wrapper that injects distance-dependent noise at each query point. The wrapper constructs a proper stochastic loss whose oracle is unbiased, satisfies the BG-0 variance model exactly, and preserves the $\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ condition (see Appendix D.1).

Across both experiments, we compare NSGDM and NSTORM using a single fresh sample per iteration against SGD and STORM with dynamic batching following Fazla et al. [13]. The normalized methods use horizon-dependent schedules motivated by Theorems 1 and 2, with experimental configurations detailed in Appendix E. Results are averaged over three random seeds and reported as mean $\pm$ standard deviation.

• **Phase retrieval** (Figure 1). We minimize

$$f(x) = \frac{1}{2m} \sum_{r=1}^{m} (y_r - |a_r^\top x|^2)^2,$$

with $m = 3000$ Gaussian measurements $a_r \in \mathbb{R}^{100}$ ($a_{r,j} \sim \mathcal{N}(0, 0.01)$), target signal $x_{\star,j} \sim \mathcal{N}(0, 1)$, and noiseless observations $y_r = |a_r^\top x_\star|^2$; the objective satisfies $\mathcal{L}^*_{\mathrm{sym}}(\frac{2}{3})$ [7]. We initialize at $x_j^0 \sim \mathcal{N}(5, 1)$, far from the target, so that BG-0 variance growth is visible. NSTORM reaches a lower gradient norm than NSGDM, consistent with the improved exponent of variance reduction, while the dynamic batching baselines require batch sizes growing to $\sim 10^3$ (right panel). The middle panel confirms that normalization controls trajectory growth: $\|x^k - x^0\|^2$ stabilizes for both normalized methods, while single-sample SGD drifts until the dynamic batch size compensates.

• **Cubic polynomial** (Figure 2). In this experiment, we minimize

$$f(x) = |x|^3$$

on $\mathbb{R}$, which satisfies $\mathcal{L}^*_{\mathrm{sym}}(\frac{1}{2})$ [7], and initialize away from the target $x_\star = 0$ at $x^0 \sim \mathcal{N}(5, 0.1)$. NSTORM again reaches a lower gradient norm than NSGDM, reflecting the benefit of variance

reduction. The dynamic batching baselines require batch sizes that grow by over an order of magnitude (right panel), whereas the normalized methods maintain batch size one throughout, except at initialization for NSTORM.

Across both experiments, normalized methods and dynamic batching offer two complementary mechanisms for handling BG-0 noise on generalized smooth landscapes: normalization controls trajectory growth through bounded step norms, while dynamic batching controls the effective noise by increasing oracle calls as the distance from initialization grows.

Discussion: Memory and parallelization trade-offs. A practical advantage of our methods is that their parameter schedules do not require knowledge of the BG-0 noise constants B and G. In contrast, dynamic batching of PASTA [13] typically selects $N_k \propto (B^2\|x^k - x^0\|^2 + G^2)/\sigma^2$, which requires estimates of these constants. Moreover, NSGDM uses one fresh sample per iteration, while NSTORM uses one fresh sample per transition, evaluated at two points, apart from its one-time (sharp) initialization batch. Thus, the per-transition sample workload remains constant, whereas dynamic batch sizes may grow with the distance from initialization, reaching approximately 10^3 in our experiments (Figures 1 and 2).

This advantage comes with a trade-off in parallelization. Dynamic batches can be evaluated concurrently, requiring approximately $\lceil N_k/P \rceil$ rounds of gradient evaluations on P workers, ignoring communication overhead. In contrast, successive normalized updates are sequentially dependent, so lower SFO complexity does not necessarily translate into lower wall-clock time. A fixed mini-batch of size b provides a possible compromise: under independent sampling, the stochastic noise standard-deviation terms decrease by a factor of $b^{1/2}$, while the convergence-rate exponent remains unchanged for fixed b. Unlike dynamic batching, this choice does not require adapting b to the iterate distance or estimating B and G, although the SFO cost per iteration increases proportionally to b.

7 Conclusion

We developed a convergence theory for normalized stochastic methods under BG-0 noise across standard and generalized smoothness regimes. For NSGDM, we showed that normalization and momentum provide implicit stabilization against distance-dependent noise, achieving $\mathcal{O}(\varepsilon^{-6})$ with a single sample per iteration and no anchoring or dynamic batching, which is minimax optimal [13]. The same rate holds under generalized smoothness. For NSTORM, variance reduction recovers the optimal $\mathcal{O}(\varepsilon^{-4})$ rate under mean-square smoothness with sharp initialization [13], while expected α-symmetric generalized smoothness yields $\mathcal{O}(\varepsilon^{-(4+\alpha)})$ for $\alpha \in (0,1)$ and $\mathcal{O}(\varepsilon^{-5})$ for $\alpha = 1$.

The α-dependent price in NSTORM raises a natural question: is it information-theoretically optimal, or is there a better rate? The price seems to originate from a coupling intrinsic to the problem structure. Under $\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ with BG-0, stochastic gradient differences depend simultaneously on $\|\nabla f(x)\|^\alpha$ from generalized smoothness and on the distance-dependent variance. This coupling vanishes when either source of difficulty is removed. We conjecture that it would confront any variance-reduced method that bounds stochastic gradient differences under $\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$. Thus, it remains an interesting open avenue to determine whether one can come up with a matching lower bound, or an alternative idea that can help curb this under generalized smoothness, leading to the optimal $\mathcal{O}(\varepsilon^{-4})$ rate.

Acknowledgements

This work is supported in part by NSF under grants DMS-2502560 and CNS-2313109.

References

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

[2] 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.

[3] 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.

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

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

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

[7] Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, pages 5396–5427. PMLR, 2023.

[8] Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd. *Advances in neural information processing systems*, 35:9955–9968, 2022.

[9] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized sgd. In *International conference on machine learning*, pages 2260–2268. PMLR, 2020.

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

[11] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

[12] 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*, pages 689–699, 2018.

[13] Arda Fazla, Ege C Kaya, Antesh Upadhyay, and Abolfazl Hashemi. Lower bounds and proximally anchored sgd for non-convex minimization under unbounded variance. *arXiv preprint arXiv:2604.16620*, 2026.

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

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

[16] 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*, pages 680–690. PMLR, 2020.

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

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

[19] Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *International Conference on Artificial Intelligence and Statistics*, pages 4861–4869. PMLR, 2024.

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

[21] 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.

[22] Jikai Jin, Bohang Zhang, Haiyang Wang, and Liwei Wang. Non-convex distributionally robust optimization: Non-asymptotic analysis. *Advances in Neural Information Processing Systems*, 34:2771–2782, 2021.

[23] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

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

[25] 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.

[26] Sarit Khirirat, Abdurakhmon Sadiev, Artem Riabinin, Eduard Gorbunov, and Peter Richtárik. Error feedback under (l_0,l_1)-smoothness: Normalization and momentum. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=LmcTbBvgjP`.

[27] Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pages 17343–17363. PMLR, 2023.

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

[29] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in neural information processing systems*, 33: 8847–8860, 2020.

[30] Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36:52166–52196, 2023.

[31] 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*, pages 6286–6295. PMLR, 2021.

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

[33] 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.

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

[35] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[36] 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*, pages 2613–2621. PMLR, 2017.

[37] 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*, pages 1571–1578, 2012.

[38] Amirhossein Reisizadeh, Haochuan Li, Subhro Das, and Ali Jadbabaie. Variance-reduced clipping for non-convex optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

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

[40] Yuki Takezawa, Han Bao, Ryoma Sato, Kenta Niwa, and Makoto Yamada. Parameter-free clipped gradient descent meets polyak. *Advances in Neural Information Processing Systems*, 37:44575–44599, 2024.

[41] Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, Tie-Yan Liu, Zhi-Quan Luo, and Wei Chen. Provable adaptivity of adam under non-uniform smoothness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2960–2969, 2024.

[42] 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.

[43] Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521, 2020.

[44] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

[45] Shen-Yi Zhao, Yin-Peng Xie, and Wu-Jun Li. On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, 64(3):132103, 2021.

[46] 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 A Short Primer on Smoothness Classes **15**

B Proofs for Normalized Momentum under BG-0 Noise **17**

B.1 Auxiliary inequalities 17

B.2 Standard smoothness and $\mathcal{L}^*_{\mathrm{sym}}(1)$ 18

B.3 The case $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ with $\alpha \in (0,1)$ 21

B.4 Recovery of bounded variance rates ($B = 0$) 24

B.5 Recovery of deterministic rates ($B = G = 0$) 26

B.6 Discussion on recovered rates 28

C Proofs for Normalized STORM under BG-0 Noise **29**

C.1 Basic inequalities 29

C.2 Descent inequalities 30

C.3 STORM decomposition and centered difference bound 31

C.4 Estimator-difference bounds 33

C.5 Single-sample master bounds 33

C.6 Proof of Theorem 2 35

C.7 Recovery of bounded variance rates ($B = 0$) 38

C.8 Recovery of deterministic rates ($B = G = 0$) 41

C.9 Discussion on recovered rates 42

D Experimental Details **43**

D.1 BG-0 Wrapper 43

D.2 Phase Retrieval 45

D.3 Cubic Polynomial 45

D.4 Additional Optimizer Comparison 45

E Hyperparameters **45**

A A Short Primer on Smoothness Classes

We summarize the smoothness notions used throughout the paper. The baseline class is the standard smoothness class $\mathcal{L}$, while the generalized smoothness classes $\mathcal{L}^*_{\rm H}$, $\mathcal{L}^*_{\rm asym}$, and $\mathcal{L}^*_{\rm sym}(\alpha)$ allow the local smoothness constant to grow with the gradient norm. The expected class $\mathbb{E}\mathcal{L}^*_{\rm sym}(\alpha)$ is the stochastic analogue used for variance-reduced methods. See Chen et al. [7] for a more detailed discussion.

Standard smoothness $\mathcal{L}$. A differentiable function $f:\mathbb{R}^d\to\mathbb{R}$ belongs to $\mathcal{L}$ if there exists $L_0>0$ such that, for all $x,y\in\mathbb{R}^d$,

$$\|\nabla f(y)-\nabla f(x)\|\le L_0\|y-x\|. \tag{4}$$

Equivalently, the gradient is globally Lipschitz continuous. This is the classical smoothness assumption used in standard nonconvex optimization.

Hessian-based generalized smoothness $\mathcal{L}^*_{\rm H}$. For twice differentiable objectives, the Hessian-based generalized smoothness class $\mathcal{L}^*_{\rm H}$ is defined by the existence of constants $L_0,L_1>0$ such that, for all $x\in\mathbb{R}^d$,

$$\|\nabla^2 f(x)\|\le L_0+L_1\|\nabla f(x)\|. \tag{5}$$

This is often called (L_0,L_1)-smoothness. Unlike standard smoothness, the curvature is allowed to increase when the gradient norm is large.

Asymmetric generalized smoothness: $\mathcal{L}^*_{\rm asym}$. The asymmetric generalized smoothness class $\mathcal{L}^*_{\rm asym}$ consists of differentiable functions satisfying, for some $L_0,L_1>0$ and all $x,y\in\mathbb{R}^d$,

$$\|\nabla f(y)-\nabla f(x)\|\le\left(L_0+L_1\|\nabla f(y)\|\right)\|y-x\|. \tag{6}$$

α-symmetric generalized smoothness $\mathcal{L}^*_{\rm sym}(\alpha)$. Following Chen et al. [7], for $\alpha\in(0,1]$, we say that $f\in\mathcal{L}^*_{\rm sym}(\alpha)$ if there exist constants $L_0,L_1>0$ such that, for all $x,y\in\mathbb{R}^d$,

$$\|\nabla f(y)-\nabla f(x)\|\le\left(L_0+L_1\max_{\theta\in[0,1]}\|\nabla f(x_\theta)\|^\alpha\right)\|y-x\|,\qquad x_\theta:=\theta y+(1-\theta)x. \tag{7}$$

This condition is symmetric in the sense that the local smoothness is controlled along the entire line segment between x and y, rather than only at one endpoint (as in the asymmetric condition in (6)). The parameter α determines how strongly the local smoothness depends on the gradient norm. The case $\alpha=1$ corresponds to the usual symmetric (L_0,L_1) generalized smoothness regime, while $\alpha\in(0,1)$ gives a sublinear dependence on the gradient norm. Standard smoothness is recovered as the special case $L_1=0$.

Expected α-symmetric generalized smoothness $\mathbb{E}\mathcal{L}^*_{\rm sym}(\alpha)$. For stochastic objectives $f(x)=\mathbb{E}_\xi[f(x;\xi)]$, the expected generalized smoothness class $\mathbb{E}\mathcal{L}^*_{\rm sym}(\alpha)$ requires that, for some $L_0,L_1>0$ and all $x,y\in\mathbb{R}^d$,

$$\mathbb{E}_\xi\left\|\nabla f(y;\xi)-\nabla f(x;\xi)\right\|^2\le\|y-x\|^2\mathbb{E}_\xi\left[\left(L_0+L_1\max_{\theta\in[0,1]}\|\nabla f(x_\theta;\xi)\|^\alpha\right)^2\right],\ x_\theta:=\theta y+(1-\theta)x. \tag{8}$$

This is a mean-square, sample-level analogue of $\mathcal{L}^*_{\rm sym}(\alpha)$. It is stronger than requiring only the population objective f to satisfy $\mathcal{L}^*_{\rm sym}(\alpha)$, and it is the relevant condition for analyzing recursive variance-reduced estimators such as NSTORM.

Mean-square smoothness MSS. Mean-square smoothness is the standard stochastic analogue of L-smoothness. It requires that there exists $L>0$ such that, for all $x,y\in\mathbb{R}^d$,

$$\mathbb{E}_\xi\left\|\nabla f(y;\xi)-\nabla f(x;\xi)\right\|^2\le L^2\|y-x\|^2. \tag{9}$$

Thus, MSS controls stochastic gradient differences without any dependence on gradient magnitudes. In contrast, $\mathbb{E}\mathcal{L}^*_{\rm sym}(\alpha)$ permits the stochastic smoothness factor to grow with the sample-gradient norm.

Relations among the classes. The classes satisfy the following inclusions, up to constants:

$$\mathcal{L} \subset \mathcal{L}^*_{\mathrm{sym}}(\alpha), \qquad \mathcal{L}^*_{\mathrm{asym}} \subset \mathcal{L}^*_{\mathrm{sym}}(1), \qquad \mathcal{L}^*_{\mathrm{H}} \subset \mathcal{L}^*_{\mathrm{sym}}(1). \tag{10}$$

Moreover, on twice differentiable functions, $\mathcal{L}^*_{\mathrm{H}}$ and $\mathcal{L}^*_{\mathrm{sym}}(1)$ are equivalent up to constants. Hence, $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ provides a unified deterministic framework that contains standard smoothness, asymmetric generalized smoothness, and Hessian-based (L_0, L_1)-smoothness. The expected class $\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ plays the analogous role for stochastic gradient differences.

Role in this paper. In this paper, $\mathcal{L}$ and $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ are used for the analysis of normalized stochastic gradient descent with momentum, NSGDM. Under BG-0 noise, the normalized update controls the trajectory length, and the generalized smoothness parameters affect constants but do not change the $\mathcal{O}(\varepsilon^{-6})$ complexity exponent.

For normalized STORM, NSTORM, the relevant conditions are MSS and $\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$. Under MSS, the transition noise in the recursive estimator is controlled only by the step length, leading to the optimal $\mathcal{O}(\varepsilon^{-4})$ rate under BG-0. Under $\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$, however, the transition noise also depends on sample-gradient moments. These moments interact with the distance-dependent BG-0 variance, producing the α-dependent rates established in the main text.

Related Work: Generalized smoothness. Standard theory often assumes a constant smoothness parameter, but recent research focuses on objectives where local smoothness grows alongside the gradient norm [6, 14]. A prime example is $\mathcal{L}^*_{\mathrm{H}}$ or (L_0, L_1)-smoothness, where the Hessian is bounded by $\|\nabla^2 f(x)\| \leq L_0 + L_1\|\nabla f(x)\|$ [44]. This concept was further expanded by Chen et al. [7] into the α-symmetric generalized smoothness framework ($\mathcal{L}^*_{\mathrm{sym}}(\alpha)/\mathbb{E}\mathcal{L}^*_{\mathrm{sym}}(\alpha)$), which covers standard smoothness, asymmetric generalized smoothness [22, 29, 38], and Hessian-based generalized smoothness, and also covers high-order polynomial and exponential-type objectives. While prior work shows that generalized smooth problems can match the efficiency of standard smooth optimization, achieving deterministic rates of $\mathcal{O}(\varepsilon^{-2})$ with normalized gradients or $\mathcal{O}(\varepsilon^{-3})$ with SPIDER-style variance reduction, these results rely on relatively tame variance assumptions. Our work investigates whether these efficiencies still hold under BG-0 noise, where variance is trajectory-dependent.

B Proofs for Normalized Momentum under BG-0 Noise

This appendix proves the normalized-momentum guarantees stated in Theorem 1. Throughout this section, let

$$T := K + 1, \qquad \Delta := f(x^0) - f^{\inf}.$$

We define

$$g_k := \mathbb{E}\|\nabla f(x^k)\|, \qquad b_k := \mathbb{E}\|v^k - \nabla f(x^k)\|,$$

and

$$\Delta^k := \mathbb{E}[f(x^k) - f^{\inf}], \qquad S_b(K) := \sum_{k=0}^{K} b_k, \qquad S_\Delta(K) := \sum_{k=0}^{K} \Delta^k.$$

The output $\widehat{x}$ is sampled uniformly from $\{x^0, x^1, \dots, x^K\}$, and hence

$$\mathbb{E}\|\nabla f(\widehat{x})\| = \frac{1}{T}\sum_{k=0}^{K} g_k. \tag{11}$$

We use the convention that $v/\|v\| = 0$ whenever $v = 0$.

B.1 Auxiliary inequalities

Lemma 1 (Trajectory control by normalization)**.** *For the* NSGDM *iterates in* (2)*,*

$$\|x^{k+1} - x^k\| \le \gamma, \qquad \|x^k - x^0\| \le k\gamma.$$

Consequently, under Assumption BG-*0,*

$$\mathbb{E}\|\nabla f(x^k;\xi^k) - \nabla f(x^k)\|^2 \le B^2k^2\gamma^2 + G^2. \tag{12}$$

In particular, since $v^0 = \nabla f(x^0;\xi^0)$*,*

$$b_0 \le G. \tag{13}$$

Proof. The normalized direction has norm at most one. Hence

$$\|x^{k+1} - x^k\| \le \gamma.$$

Summing the increments gives

$$\|x^k - x^0\| \le \sum_{t=0}^{k-1} \|x^{t+1} - x^t\| \le k\gamma.$$

Substitution into Assumption BG-0 gives (12). Finally,

$$b_0 = \mathbb{E}\|v^0 - \nabla f(x^0)\| \le \sqrt{\mathbb{E}\|v^0 - \nabla f(x^0)\|^2} \le G,$$

because $x^0 - x^0 = 0$. ∎

Lemma 2 (Direction-error inequality)**.** *For any* $a, e \in \mathbb{R}^d$*, define*

$$d = \begin{cases} \dfrac{a+e}{\|a+e\|}, & a + e \neq 0, \\ 0, & a + e = 0. \end{cases}$$

Then

$$\langle a, d\rangle \ge \|a\| - 2\|e\|. \tag{14}$$

Proof. If $a + e = 0$, then $d = 0$ and $e = -a$. Hence

$$\langle a, d\rangle = 0 \ge \|a\| - 2\|a\|.$$

Now suppose $a + e \neq 0$. Then

$$\langle a, d\rangle = \frac{\|a\|^2 + \langle a, e\rangle}{\|a+e\|} \ge \frac{\|a\|(\|a\| - \|e\|)}{\|a+e\|}.$$

If $\|a\| \geq \|e\|$, then using $\|a+e\| \leq \|a\| + \|e\|$, we obtain

$$\langle a, d\rangle \geq \frac{\|a\|(\|a\| - \|e\|)}{\|a\| + \|e\|} = \|a\| - \frac{2\|a\|\|e\|}{\|a\| + \|e\|} \geq \|a\| - 2\|e\|.$$

If $\|a\| < \|e\|$, then by Cauchy–Schwarz,

$$\langle a, d\rangle \geq -\|a\| \geq \|a\| - 2\|e\|.$$

The claim follows. ■

Lemma 3 (Young inequality for fractional powers)**.** *For every $x \geq 0$, $\alpha \in (0,1)$, and $\rho > 0$,*

$$x^\alpha \leq \alpha\rho x + (1-\alpha)\rho^{-\alpha/(1-\alpha)}. \tag{15}$$

Proof. Apply Young's inequality with conjugate exponents $p = 1/\alpha$ and $q = 1/(1-\alpha)$ to

$$u = (\rho x)^\alpha, \qquad v = \rho^{-\alpha}.$$

■

B.2 Standard smoothness and $\mathcal{L}^*_{\mathrm{sym}}(1)$

We first prove the standard smoothness case and the endpoint $\mathcal{L}^*_{\mathrm{sym}}(1)$ case. For the endpoint generalized smooth case, we use the following standard consequences of symmetric generalized smoothness (see Chen et al. [7], Proposition 1): for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f(y) - \nabla f(x)\| \leq \big(L_0 + L_1\|\nabla f(x)\|\big)e^{L_1\|y-x\|}\|y-x\|, \tag{16}$$

and

$$f(y) \leq f(x) + \langle \nabla f(x), y-x\rangle + \frac{1}{2}\big(L_0 + L_1\|\nabla f(x)\|\big)e^{L_1\|y-x\|}\|y-x\|^2. \tag{17}$$

When $L_1 = 0$, these reduce to the standard L_0-smooth inequalities. We also use the gradient-growth inequality (see Khirirat et al. [26], Lemma 2)

$$\|\nabla f(x)\| \leq 8L_1(f(x) - f^{\inf}) + \frac{L_0}{L_1}, \qquad L_1 > 0, \tag{18}$$

which follows from lower boundedness and the endpoint generalized smoothness condition.

Theorem 3 (Single-sample master bound for L_0-smoothness and $\alpha = 1$)**.** *Suppose f is bounded below and* BG-*0 hold. Assume that either f is L_0-smooth, or that* (16)–(17) *hold. In the standard smooth case, set $L_1 = 0$. If*

$$\gamma L_1 \leq \frac{1}{2}, \qquad \frac{32L_1^2\gamma^2 T}{\eta} \leq \frac{1}{2}, \tag{19}$$

with both conditions interpreted as vacuous when $L_1 = 0$, then

$$\begin{aligned}\mathbb{E}\|\nabla f(\widehat{x})\| \leq{}& \frac{2\Delta}{\gamma T} + \frac{4b_0}{\eta T} + \frac{16L_0\gamma}{\eta} + 2L_0\gamma + 4\sqrt{\eta}\Big(G + \frac{B\gamma K}{2}\Big). \\ &+ \frac{64L_1^2\gamma}{\eta}\left[4\Delta + \frac{4\gamma b_0}{\eta} + \frac{16L_0\gamma^2 T}{\eta} + 4\gamma\sqrt{\eta}\,TG + 2B\gamma^2\sqrt{\eta}\,KT + 2L_0\gamma^2 T\right]\end{aligned} \tag{20}$$

When $L_1 = 0$, the entire L_1^2-term vanishes.

Proof. Let

$$E_k := v^k - \nabla f(x^k), \qquad D_k := \nabla f(x^k) - \nabla f(x^{k+1}),$$

and

$$\zeta_{k+1} := \nabla f(x^{k+1}; \xi^{k+1}) - \nabla f(x^{k+1}).$$

Step 1: descent. Using (17) with $x = x^k$ and $y = x^{k+1} = x^k - \gamma d^k$, we get

$$f(x^{k+1}) \le f(x^k) - \gamma\langle \nabla f(x^k), d^k\rangle + \frac{\gamma^2}{2} e^{L_1\gamma}\big(L_0 + L_1\|\nabla f(x^k)\|\big).$$

Since $\gamma L_1 \le 1/2$, we have $e^{L_1\gamma} \le e^{1/2} < 2$. By Lemma 2,

$$\langle \nabla f(x^k), d^k\rangle \ge \|\nabla f(x^k)\| - 2\|E_k\|.$$

Therefore,

$$f(x^{k+1}) \le f(x^k) - \gamma\|\nabla f(x^k)\| + 2\gamma\|E_k\| + L_0\gamma^2 + L_1\gamma^2\|\nabla f(x^k)\|.$$

Using $\gamma L_1 \le 1/2$, we obtain

$$\Delta^{k+1} \le \Delta^k - \frac{\gamma}{2} g_k + 2\gamma b_k + L_0\gamma^2. \tag{21}$$

Step 2: estimator recursion. The momentum update gives

$$E_{k+1} = (1-\eta)E_k + (1-\eta)D_k + \eta\zeta_{k+1}.$$

Unrolling,

$$E_k = (1-\eta)^k E_0 + \sum_{t=0}^{k-1}(1-\eta)^{k-t} D_t + \eta\sum_{t=0}^{k-1}(1-\eta)^{k-1-t}\zeta_{t+1}.$$

For the drift term, (16), $e^{L_1\gamma} \le 2$, and (18) imply, for $L_1 > 0$,

$$\begin{aligned}\mathbb{E}\|D_t\| &\le 2\gamma\big(L_0 + L_1 g_t\big)\\ &\le 2\gamma\left(L_0 + L_1\left(8L_1\Delta^t + \frac{L_0}{L_1}\right)\right)\\ &= 4L_0\gamma + 16L_1^2\gamma\Delta^t.\end{aligned}$$

For $L_1 = 0$, the second term is absent and the same displayed bound remains valid with the L_1^2-term equal to zero.

For the martingale term, conditional unbiasedness and the orthogonality of martingale differences give

$$\begin{aligned}&\mathbb{E}\left\|\eta\sum_{t=0}^{k-1}(1-\eta)^{k-1-t}\zeta_{t+1}\right\|\\ &\le \left(\eta^2\sum_{t=0}^{k-1}(1-\eta)^{2(k-1-t)}\mathbb{E}\|\zeta_{t+1}\|^2\right)^{1/2}.\end{aligned}$$

By Lemma 1, for $t \le k-1$,

$$\mathbb{E}\|\zeta_{t+1}\|^2 \le B^2(t+1)^2\gamma^2 + G^2 \le B^2k^2\gamma^2 + G^2.$$

Also,

$$\sum_{t=0}^{k-1}(1-\eta)^{2(k-1-t)} \le \frac{1}{\eta}.$$

Hence

$$\mathbb{E}\left\|\eta\sum_{t=0}^{k-1}(1-\eta)^{k-1-t}\zeta_{t+1}\right\| \le \sqrt{\eta}(G + Bk\gamma).$$

Combining the drift and martingale estimates yields

$$b_k \le (1-\eta)^k b_0 + \frac{4L_0\gamma}{\eta} + 16L_1^2\gamma\sum_{t=0}^{k-1}(1-\eta)^{k-t}\Delta^t + \sqrt{\eta}(G + Bk\gamma). \tag{22}$$

Step 3: summing the estimator errors. Summing (22) over $k = 0, \dots, K$, and using

$$\sum_{k=0}^{K}(1-\eta)^k \le \frac{1}{\eta},$$

and

$$\sum_{k=0}^{K}\sum_{t=0}^{k-1}(1-\eta)^{k-t}\Delta^t \le \frac{1}{\eta}S_\Delta(K),$$

we get

$$S_b(K) \le \frac{b_0}{\eta} + \frac{4L_0\gamma T}{\eta} + \frac{16L_1^2\gamma}{\eta}S_\Delta(K) + \sqrt{\eta}\left(TG + \frac{B\gamma KT}{2}\right). \tag{23}$$

Step 4: controlling the accumulated suboptimality. Dropping the negative gradient term in (21), iterating, and summing gives

$$S_\Delta(K) \le (K+2)\Delta + 2\gamma T S_b(K) + \frac{L_0\gamma^2 T(K+2)}{2}.$$

Substituting (23) gives

$$\begin{aligned} S_\Delta(K) \le (K+2)\Delta + \frac{2\gamma T b_0}{\eta} + \frac{8L_0\gamma^2T^2}{\eta} + \frac{32L_1^2\gamma^2T}{\eta}S_\Delta(K) \\ + 2\gamma\sqrt{\eta}T^2G + B\gamma^2\sqrt{\eta}KT^2 + \frac{L_0\gamma^2T(K+2)}{2}. \end{aligned}$$

By (19), the coefficient of $S_\Delta(K)$ on the right is at most $1/2$. Moving it to the left and using $K+2 \le 2T$, we obtain

$$\frac{S_\Delta(K)}{T} \le 4\Delta + \frac{4\gamma b_0}{\eta} + \frac{16L_0\gamma^2T}{\eta} + 4\gamma\sqrt{\eta}\,TG + 2B\gamma^2\sqrt{\eta}\,KT + 2L_0\gamma^2T. \tag{24}$$

Step 5: final stationarity bound. Summing (21) gives

$$\frac{\gamma}{2}\sum_{k=0}^{K} g_k \le \Delta + 2\gamma S_b(K) + L_0\gamma^2 T.$$

Dividing by $\gamma T/2$, using (11), and substituting (23) and (24), we obtain (20). ■

Corollary 3.1 (Standard smoothness under BG-0)**.** *Assume Assumption 2 holds. Choose*

$$\eta = T^{-2/3}, \qquad \gamma = \gamma_0 T^{-5/6}.$$

Then

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \left(\frac{2\Delta}{\gamma_0} + 16L_0\gamma_0 + 2B\gamma_0\right)T^{-1/6} + 8GT^{-1/3} + 2L_0\gamma_0T^{-5/6}. \tag{25}$$

Consequently,

$$\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/6}),$$

and the SFO complexity is $\mathcal{O}(\varepsilon^{-6})$.

Proof. Set $L_1 = 0$ in Theorem 3. Then the L_1^2-term vanishes and the conditions (19) are vacuous. Using $b_0 \le G$ and $K \le T$,

$$\frac{2\Delta}{\gamma T} = \frac{2\Delta}{\gamma_0}T^{-1/6}, \qquad \frac{4b_0}{\eta T} \le 4GT^{-1/3},$$

$$\frac{16L_0\gamma}{\eta} = 16L_0\gamma_0T^{-1/6}, \qquad 2L_0\gamma = 2L_0\gamma_0T^{-5/6},$$

and

$$4\sqrt{\eta}\left(G + \frac{B\gamma K}{2}\right) \le 4GT^{-1/3} + 2B\gamma_0T^{-1/6}.$$

Combining these estimates proves (25). Since $T \ge 1$, we also have

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \left(\frac{2\Delta}{\gamma_0} + (18L_0 + 2B)\gamma_0\right)T^{-1/6} + 8GT^{-1/3}.$$

This gives the compact bound in Theorem 1(i). ■

Corollary 3.2 (1-generalized smoothness under BG-0)**.** *Assume Assumption 4 with $\alpha = 1$ holds. Choose*

$$\eta = T^{-2/3}, \qquad \gamma = \gamma_0 T^{-5/6}, \qquad 0 < \gamma_0 \le \frac{1}{8L_1}.$$

Then

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \left(\frac{2\Delta}{\gamma_0} + 16L_0\gamma_0 + 2B\gamma_0\right) T^{-1/6} + 8GT^{-1/3} + 2L_0\gamma_0 T^{-5/6} + 64L_1^2\gamma_0 T^{-1/6}\left[4\Delta + (16L_0 + 2B)\gamma_0^2 + 8\gamma_0 GT^{-1/6} + 2L_0\gamma_0^2 T^{-2/3}\right]. \quad (26)$$

Consequently,

$$\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/6}),$$

and the SFO complexity is $\mathcal{O}(\varepsilon^{-6})$.

Proof. The first condition in (19) follows from

$$\gamma L_1 \le \gamma_0 L_1 \le \frac{1}{8}.$$

The second follows because

$$\frac{32L_1^2\gamma^2 T}{\eta} = 32L_1^2\gamma_0^2 \le \frac{1}{2}.$$

Substitute the schedule into (20), use $b_0 \le G$ and $K \le T$, and simplify. The non-coupling terms give the first line of (26). The bracketed term in (20) is bounded by

$$4\Delta + (16L_0 + 2B)\gamma_0^2 + 8\gamma_0 GT^{-1/6} + 2L_0\gamma_0^2 T^{-2/3},$$

while its prefactor becomes

$$64L_1^2\gamma_0 T^{-1/6}.$$

This proves (26).

To obtain the compact statement of Theorem 1(iii), note that the restriction $\gamma_0 \le 1/(8L_1)$ implies

$$64L_1^2\gamma_0^2 \le 1, \qquad 256L_1^2\gamma_0\Delta \le 32L_1\Delta.$$

Since $T \ge 1$, the faster-decaying terms can be bounded by the leading decay order. Consequently,

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \left(\frac{2\Delta}{\gamma_0} + 32L_1\Delta + (36L_0 + 4B)\gamma_0\right) T^{-1/6} + 16GT^{-1/3}.$$

This proves the bound in Theorem 1(iii). ■

B.3 The case $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ with $\alpha \in (0, 1)$

We now prove the sublinear generalized smoothness case. The proof gives a clean explicit bound that implies the same $\mathcal{O}(T^{-1/6})$ rate as in Theorem 1.

Under Assumption 4 with $\alpha \in (0, 1)$, we use the following consequences of the symmetric generalized smoothness condition (see Chen et al. [7], Proposition 1). For all $w, w' \in \mathbb{R}^d$,

$$\|\nabla f(w') - \nabla f(w)\| \le \|w' - w\|\left(K_0 + K_1\|\nabla f(w)\|^\alpha + K_2\|w' - w\|^{\alpha/(1-\alpha)}\right), \quad (27)$$

and

$$f(w') \le f(w) + \langle \nabla f(w), w' - w\rangle + \frac{1}{2}\|w' - w\|^2\left(K_0 + K_1\|\nabla f(w)\|^\alpha + 2K_2\|w' - w\|^{\alpha/(1-\alpha)}\right), \quad (28)$$

where

$$K_0 := L_0\left(2^{\alpha^2/(1-\alpha)} + 1\right), \qquad K_1 := L_1\, 2^{\alpha^2/(1-\alpha)} 3^\alpha,$$

and

$$K_2 := L_1^{1/(1-\alpha)} 2^{\alpha^2/(1-\alpha)} 3^\alpha (1-\alpha)^{\alpha/(1-\alpha)}.$$

Define

$$R_\alpha := \frac{(1-\alpha)K_1}{2}(2\alpha K_1)^{\alpha/(1-\alpha)}, \qquad C_\alpha := K_2 + R_\alpha,$$

and

$$\widetilde{R}_\alpha := (1-\alpha)K_1(8\alpha K_1)^{\alpha/(1-\alpha)}, \qquad \widetilde{C}_\alpha := K_0 + K_2 + \widetilde{R}_\alpha.$$

For the compact bound in Theorem 1(ii), define

$$\Lambda_\alpha := 4\widetilde{C}_\alpha + K_0 + 2C_\alpha.$$

Lemma 4 (One-step descent under $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$)**.** *Assume $\alpha \in (0,1)$, $\gamma \le 1$, and let*

$$e^k := v^k - \nabla f(x^k).$$

Then

$$f(x^{k+1}) \le f(x^k) - \frac{3\gamma}{4}\|\nabla f(x^k)\| + 2\gamma\|e^k\| + \frac{K_0}{2}\gamma^2 + C_\alpha\gamma^{(2-\alpha)/(1-\alpha)}. \tag{29}$$

Consequently,

$$\Delta^{k+1} \le \Delta^k - \frac{3\gamma}{4}g_k + 2\gamma b_k + \frac{K_0}{2}\gamma^2 + C_\alpha\gamma^{(2-\alpha)/(1-\alpha)}. \tag{30}$$

Proof. Apply (28) with $w = x^k$ and $w' = x^{k+1}$. Since $\|x^{k+1} - x^k\| \le \gamma$,

$$f(x^{k+1}) \le f(x^k) - \gamma\langle\nabla f(x^k), d^k\rangle + \frac{K_0}{2}\gamma^2 + \frac{K_1}{2}\gamma^2\|\nabla f(x^k)\|^\alpha + K_2\gamma^{(2-\alpha)/(1-\alpha)}.$$

Lemma 2 gives

$$-\gamma\langle\nabla f(x^k), d^k\rangle \le -\gamma\|\nabla f(x^k)\| + 2\gamma\|e^k\|.$$

Using Lemma 3 with

$$\rho = (2\alpha K_1\gamma)^{-1},$$

we obtain

$$\frac{K_1}{2}\gamma^2\|\nabla f(x^k)\|^\alpha \le \frac{\gamma}{4}\|\nabla f(x^k)\| + R_\alpha\gamma^{(2-\alpha)/(1-\alpha)}.$$

Combining the last two displays proves (29). Taking expectations gives (30). ■

Lemma 5 (Gradient-drift bound)**.** *Under the assumptions of Lemma 4, for every k,*

$$\|\nabla f(x^{k+1}) - \nabla f(x^k)\| \le \widetilde{C}_\alpha\gamma + \frac{\gamma}{8}\|\nabla f(x^k)\|. \tag{31}$$

Proof. By (27) and $\|x^{k+1} - x^k\| \le \gamma$,

$$\|\nabla f(x^{k+1}) - \nabla f(x^k)\| \le K_0\gamma + K_1\gamma\|\nabla f(x^k)\|^\alpha + K_2\gamma^{1/(1-\alpha)}.$$

Since $\gamma \le 1$, we have

$$\gamma^{1/(1-\alpha)} \le \gamma.$$

Applying Lemma 3 with

$$\rho = (8\alpha K_1)^{-1}$$

gives

$$K_1\gamma\|\nabla f(x^k)\|^\alpha \le \frac{\gamma}{8}\|\nabla f(x^k)\| + \widetilde{R}_\alpha\gamma.$$

Substitution proves (31). ■

Theorem 4 (Single-sample master bound under $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$)**.** *Suppose f is bounded below,* BG-*0, and 4 hold with fixed $\alpha \in (0,1)$. If*

$$\gamma \le 1, \qquad \gamma \le \eta,$$

then

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \frac{2\Delta}{\gamma T} + \frac{4b_0}{\eta T} + \frac{4\widetilde{C}_\alpha\gamma}{\eta} + 4\sqrt{\eta}\left(G + \frac{B\gamma K}{2}\right) + K_0\gamma + 2C_\alpha\gamma^{1/(1-\alpha)}. \tag{32}$$

Moreover, $b_0 \le G$.

Proof. Let

$$E_k := v^k - \nabla f(x^k), \qquad \Delta_k := \nabla f(x^k) - \nabla f(x^{k+1}),$$

and

$$\zeta_{k+1} := \nabla f(x^{k+1}; \xi^{k+1}) - \nabla f(x^{k+1}).$$

Unrolling the estimator recursion gives

$$E_k = (1-\eta)^k E_0 + \sum_{t=0}^{k-1} (1-\eta)^{k-t} \Delta_t + \eta \sum_{t=0}^{k-1} (1-\eta)^{k-1-t} \zeta_{t+1}.$$

By Lemma 5,

$$\|\Delta_t\| \le \widetilde{C}_\alpha \gamma + \frac{\gamma}{8} \|\nabla f(x^t)\|.$$

The same martingale calculation as in Theorem 3 gives

$$\mathbb{E} \left\| \eta \sum_{t=0}^{k-1} (1-\eta)^{k-1-t} \zeta_{t+1} \right\| \le \sqrt{\eta}(G + Bk\gamma).$$

Therefore,

$$b_k \le (1-\eta)^k b_0 + \frac{\widetilde{C}_\alpha \gamma}{\eta} + \frac{\gamma}{8} \sum_{t=0}^{k-1} (1-\eta)^{k-t} g_t + \sqrt{\eta}(G + Bk\gamma). \tag{33}$$

Summing over $k = 0, \dots, K$ yields

$$S_b(K) \le \frac{b_0}{\eta} + \frac{\widetilde{C}_\alpha \gamma T}{\eta} + \frac{\gamma}{8\eta} \sum_{k=0}^{K} g_k + \sqrt{\eta} \left(TG + \frac{B\gamma KT}{2} \right). \tag{34}$$

Next, summing (30) gives

$$\frac{3\gamma}{4} \sum_{k=0}^{K} g_k \le \Delta + 2\gamma S_b(K) + \frac{K_0}{2} T\gamma^2 + TC_\alpha \gamma^{(2-\alpha)/(1-\alpha)}.$$

Substituting (34) gives

$$\gamma \left(\frac{3}{4} - \frac{\gamma}{4\eta} \right) \sum_{k=0}^{K} g_k \le \Delta + \frac{2\gamma b_0}{\eta} + \frac{2\widetilde{C}_\alpha \gamma^2 T}{\eta} + 2\gamma\sqrt{\eta} \left(TG + \frac{B\gamma KT}{2} \right) + \frac{K_0}{2} T\gamma^2 + TC_\alpha \gamma^{(2-\alpha)/(1-\alpha)}.$$

Since $\gamma \le \eta$, we have

$$\frac{3}{4} - \frac{\gamma}{4\eta} \ge \frac{1}{2}.$$

Thus,

$$\frac{\gamma}{2} \sum_{k=0}^{K} g_k \le \Delta + \frac{2\gamma b_0}{\eta} + \frac{2\widetilde{C}_\alpha \gamma^2 T}{\eta} + 2\gamma\sqrt{\eta} \left(TG + \frac{B\gamma KT}{2} \right) + \frac{K_0}{2} T\gamma^2 + TC_\alpha \gamma^{(2-\alpha)/(1-\alpha)}.$$

Dividing by $\gamma T/2$ and using (11) yields (32). Finally, $b_0 \le G$ follows from Lemma 1. ■

Corollary 4.1 ($\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ smoothness under BG-0)**.** *Suppose f is bounded below,* BG-*0, and 4 hold with fixed $\alpha \in (0,1)$. Choose*

$$\eta = T^{-2/3}, \qquad \gamma = \gamma_0 T^{-5/6}, \qquad 0 < \gamma_0 \le 1.$$

Then

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \left(\frac{2\Delta}{\gamma_0} + 4\widetilde{C}_\alpha \gamma_0 + 2B\gamma_0 \right) T^{-1/6} + 8GT^{-1/3} + K_0 \gamma_0 T^{-5/6} + 2C_\alpha \gamma_0^{1/(1-\alpha)} T^{-5/(6(1-\alpha))}. \tag{35}$$

Consequently,

$$\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/6}),$$

and the SFO complexity is $\mathcal{O}(\varepsilon^{-6})$.

Proof. Because $T \geq 1$ and $0 < \gamma_0 \leq 1$,

$$\gamma = \gamma_0 T^{-5/6} \leq 1.$$

Moreover,

$$\gamma = \gamma_0 T^{-5/6} \leq T^{-5/6} \leq T^{-2/3} = \eta.$$

Hence, the conditions of Theorem 4 hold. Substituting the schedule into (32), using $b_0 \leq G$ and $K \leq T$, gives

$$\frac{2\Delta}{\gamma T} = \frac{2\Delta}{\gamma_0} T^{-1/6}, \qquad \frac{4b_0}{\eta T} \leq 4GT^{-1/3},$$

$$\frac{4\widetilde{C}_\alpha \gamma}{\eta} = 4\widetilde{C}_\alpha \gamma_0 T^{-1/6},$$

and

$$4\sqrt{\eta}\left(G + \frac{B\gamma K}{2}\right) \leq 4GT^{-1/3} + 2B\gamma_0 T^{-1/6}.$$

The remaining deterministic terms are

$$K_0 \gamma = K_0 \gamma_0 T^{-5/6},$$

and

$$2C_\alpha \gamma^{1/(1-\alpha)} = 2C_\alpha \gamma_0^{1/(1-\alpha)} T^{-5/(6(1-\alpha))}.$$

Combining these estimates proves (35). Since $T \geq 1$, $0 < \gamma_0 \leq 1$, and $\alpha \in (0, 1)$,

$$K_0 \gamma_0 T^{-5/6} \leq K_0 \gamma_0 T^{-1/6},$$

and

$$2C_\alpha \gamma_0^{1/(1-\alpha)} T^{-5/(6(1-\alpha))} \leq 2C_\alpha \gamma_0 T^{-1/6}.$$

Hence, by the definition of Λ_α,

$$\mathbb{E}\|\nabla f(\widehat{x})\| \leq \left(\frac{2\Delta}{\gamma_0} + (\Lambda_\alpha + 2B)\gamma_0\right) T^{-1/6} + 8GT^{-1/3}.$$

This proves the compact bound in Theorem 1(ii).

Since $\alpha \in (0, 1)$ is fixed, all terms decay at least as fast as $T^{-1/6}$. Therefore,

$$\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/6}),$$

which is equivalent to $\mathcal{O}(\varepsilon^{-6})$ SFO complexity. ■

B.4 Recovery of bounded variance rates ($B = 0$)

When the distance-growth parameter B in Assumption BG-0 vanishes, the BG-0 condition reduces to bounded variance $\mathbb{E}_\xi \|\nabla f(x;\xi) - \nabla f(x)\|^2 \leq G^2$, and the NSGDM complexity improves from $\mathcal{O}(\varepsilon^{-6})$ to $\mathcal{O}(\varepsilon^{-4})$. The following corollaries formalize this for all three smoothness regimes; in each case, they follow from the corresponding master bound by substituting $B = 0$ and the bounded variance schedule.

Corollary 4.2 (Standard smoothness under bounded variance)**.** *Assume f is L_0-smooth (Assumption 2) and that f is bounded below and* BG-0 *hold with $B = 0$. Choose*

$$\eta = T^{-1/2}, \qquad \gamma = \gamma_0\, T^{-3/4},$$

where $T := K + 1$ and $\gamma_0 > 0$. Then

$$\mathbb{E}\|\nabla f(\widehat{x})\| \leq \left(\frac{2\Delta}{\gamma_0} + 16L_0\gamma_0\right) T^{-1/4} + 4G\, T^{-1/4} + 4G\, T^{-1/2} + 2L_0\gamma_0\, T^{-3/4}. \tag{36}$$

Consequently, $\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/4})$, and the SFO complexity is $\mathcal{O}(\varepsilon^{-4})$.

Proof. Set $L_1 = 0$ and $B = 0$ in Theorem 3. The L_1^2-coupling term vanishes and the conditions (19) are vacuous. Using $b_0 \le G$ and $K \le T$, we compute

$$\frac{2\Delta}{\gamma T} = \frac{2\Delta}{\gamma_0}\, T^{-1/4}, \qquad \frac{4b_0}{\eta T} \le 4G\, T^{-1/2},$$

$$\frac{16L_0\gamma}{\eta} = 16L_0\gamma_0\, T^{-1/4}, \qquad 2L_0\gamma = 2L_0\gamma_0\, T^{-3/4},$$

and, since $B = 0$,

$$4\sqrt{\eta}\left(G + \frac{B\gamma K}{2}\right) = 4G\, T^{-1/4}.$$

Combining these estimates proves (36). ■

Corollary 4.3 (1-generalized smoothness under bounded variance)**.** *Assume Assumption 4 holds with $\alpha = 1$ and constants $L_0, L_1 > 0$, and that f is bounded below and* BG*-0 hold with $B = 0$. Choose*

$$\eta = T^{-1/2}, \qquad \gamma = \gamma_0\, T^{-3/4}, \qquad 0 < \gamma_0 \le \frac{1}{8L_1}.$$

Then

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \left(\frac{2\Delta}{\gamma_0} + 16L_0\gamma_0\right) T^{-1/4} + 4G\, T^{-1/4} + 4G\, T^{-1/2} + 2L_0\gamma_0\, T^{-3/4} + 64L_1^2\gamma_0\, T^{-1/4}\left[4\Delta + 16L_0\gamma_0^2 + 4\gamma_0 G(1 + T^{-1/4}) + 2L_0\gamma_0^2\, T^{-1/2}\right]. \tag{37}$$

Consequently, $\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/4})$, and the SFO complexity is $\mathcal{O}(\varepsilon^{-4})$.

Proof. The first condition in (19) holds because $\gamma L_1 \le \gamma_0 L_1 \le 1/8$. For the second,

$$\frac{32L_1^2\gamma^2 T}{\eta} = 32L_1^2\gamma_0^2 \le \frac{1}{2}.$$

Substitute the schedule into (20) with $B = 0$. The non-coupling terms give the first line. For the bracketed expression, using $B = 0$, $b_0 \le G$, and $K \le T$:

$$4\Delta + \tfrac{4\gamma b_0}{\eta} + \tfrac{16L_0\gamma^2 T}{\eta} + 4\gamma\sqrt{\eta}\, TG + 2L_0\gamma^2 T \le 4\Delta + 4\gamma_0 G\, T^{-1/4} + 16L_0\gamma_0^2 + 4\gamma_0 G + 2L_0\gamma_0^2\, T^{-1/2}.$$

This yields the stated bracket. The prefactor is $64L_1^2\gamma_0\, T^{-1/4}$. Since the bracket is $\mathcal{O}(1)$, the entire coupling term is $\mathcal{O}(T^{-1/4})$. ■

Corollary 4.4 ($\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ smoothness under bounded variance)**.** *Suppose f is bounded below, Assumptions* BG*-0 (with $B = 0$), and 4 hold with fixed $\alpha \in (0, 1)$. Choose*

$$\eta = T^{-1/2}, \qquad \gamma = \gamma_0\, T^{-3/4}, \qquad 0 < \gamma_0 \le 1.$$

Then

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \left(\frac{2\Delta}{\gamma_0} + 4\widetilde{C}_\alpha\gamma_0\right) T^{-1/4} + 4G\, T^{-1/4} + 4G\, T^{-1/2} + K_0\gamma_0\, T^{-3/4} + 2C_\alpha\gamma_0^{1/(1-\alpha)}\, T^{-3/(4(1-\alpha))}. \tag{38}$$

Consequently, $\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/4})$, and the SFO complexity is $\mathcal{O}(\varepsilon^{-4})$.

Proof. Since $\gamma_0 \le 1$ and $T \ge 1$, we have $\gamma \le 1$ and $\gamma \le \eta$. The conditions of Theorem 4 hold. Substituting the schedule into (32) with $B = 0$ and using $b_0 \le G$,

$$\frac{2\Delta}{\gamma T} = \frac{2\Delta}{\gamma_0}\, T^{-1/4}, \qquad \frac{4b_0}{\eta T} \le 4G\, T^{-1/2},$$

$$\frac{4\widetilde{C}_\alpha\gamma}{\eta} = 4\widetilde{C}_\alpha\gamma_0\, T^{-1/4},$$

and, since $B = 0$,

$$4\sqrt{\eta}\left(G + \frac{B\gamma K}{2}\right) = 4G\,T^{-1/4}.$$

The remaining deterministic terms are

$$K_0\gamma = K_0\gamma_0\,T^{-3/4}, \qquad 2C_\alpha\gamma^{1/(1-\alpha)} = 2C_\alpha\gamma_0^{1/(1-\alpha)}\,T^{-3/(4(1-\alpha))}.$$

The two noise contributions are $4G\,T^{-1/2}$ and $4G\,T^{-1/4}$, as stated in (38). Since $\alpha \in (0, 1)$ is fixed, $3/(4(1-\alpha)) > 3/4 > 1/4$, so every term decays at least as fast as $T^{-1/4}$. This proves (38) and the $\mathcal{O}(T^{-1/4})$ rate. ■

B.5 Recovery of deterministic rates ($B = G = 0$)

When the oracle is deterministic ($B = G = 0$), we have $v^k = \nabla f(x^k)$ for every k (since $\eta = 1$ is natural and the oracle returns the exact gradient). Thus $b_k = 0$ for all k, and the stochastic error analysis simplifies to a pure descent argument. We recover the classical $\mathcal{O}(\varepsilon^{-2})$ rate for all smoothness regimes.

Corollary 4.5 (Standard smoothness, deterministic)**.** *Assume f is L_0-smooth (Assumption 2), that f is bounded below holds, and that the oracle is deterministic ($B = G = 0$). Consider the normalized gradient descent update*

$$x^{k+1} = x^k - \gamma\,\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}.$$

Choose $\gamma = \gamma_0\,T^{-1/2}$. Then

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \frac{2\Delta}{\gamma_0}\,T^{-1/2} + 2L_0\gamma_0\,T^{-1/2}. \tag{39}$$

Consequently, $\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/2})$, and the iteration complexity is $\mathcal{O}(\varepsilon^{-2})$.

Proof. Under determinism, set $\eta = 1$ so that $v^k = \nabla f(x^k)$ for all k. Then $b_k = 0$ and the descent recursion (21) (with $L_1 = 0$) becomes

$$f(x^{k+1}) \le f(x^k) - \gamma\|\nabla f(x^k)\| + L_0\gamma^2.$$

Here, the coefficient is γ rather than $\gamma/2$ because, with $b_k = 0$, the direction-error term vanishes and no gradient term needs to absorb a generalized smoothness contribution. Summing from $k = 0$ to K and using $f(x^{K+1}) \ge f^{\inf}$,

$$\gamma\sum_{k=0}^{K}\|\nabla f(x^k)\| \le \Delta + L_0\gamma^2 T.$$

Dividing by γT gives

$$\frac{1}{T}\sum_{k=0}^{K}\|\nabla f(x^k)\| \le \frac{\Delta}{\gamma T} + L_0\gamma.$$

Substituting $\gamma = \gamma_0 T^{-1/2}$ yields $\Delta/(\gamma_0)\,T^{-1/2} + L_0\gamma_0\,T^{-1/2}$, from which (39) follows (with an extra factor of 2 for a uniform upper bound). ■

Corollary 4.6 (1-generalized smoothness, deterministic)**.** *Assume Assumption 4 holds with $\alpha = 1$ and constants $L_0, L_1 > 0$, that f is bounded below, and that the oracle is deterministic ($B = G = 0$). Consider the normalized gradient descent update*

$$x^{k+1} = x^k - \gamma\,\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}.$$

Choose $\gamma = \gamma_0\,T^{-1/2}$ with $0 < \gamma_0 \le 1/(2L_1)$. Then

$$\|\nabla f(\widehat{x})\| \le \frac{4\Delta}{\gamma_0}\,T^{-1/2} + 4L_0\gamma_0\,T^{-1/2}. \tag{40}$$

Consequently, $\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/2})$, and the iteration complexity is $\mathcal{O}(\varepsilon^{-2})$.

Proof. Under determinism, $v^k = \nabla f(x^k)$, $b_k = 0$. Using (17) with $y = x^{k+1} = x^k - \gamma d^k$, and noting that $d^k = \nabla f(x^k)/\|\nabla f(x^k)\|$ when the gradient is nonzero (so $\langle \nabla f(x^k), d^k \rangle = \|\nabla f(x^k)\|$),

$$f(x^{k+1}) \le f(x^k) - \gamma \|\nabla f(x^k)\| + \frac{\gamma^2}{2} e^{L_1 \gamma} \big(L_0 + L_1 \|\nabla f(x^k)\|\big).$$

Since $\gamma_0 \le 1/(2L_1)$ and $T \ge 1$, we have $\gamma L_1 \le \gamma_0 L_1 \le 1/2$, so $e^{L_1\gamma} \le 2$. Therefore,

$$\frac{\gamma^2}{2} e^{L_1\gamma} L_1 \|\nabla f(x^k)\| \le \gamma^2 L_1 \|\nabla f(x^k)\| \le \frac{\gamma}{2} \|\nabla f(x^k)\|,$$

where the last inequality follows from $\gamma L_1 \le \gamma_0 L_1 \le 1/2$. This gives

$$f(x^{k+1}) \le f(x^k) - \frac{\gamma}{2} \|\nabla f(x^k)\| + L_0 \gamma^2.$$

Summing from $k = 0$ to K and dividing by $\gamma T/2$,

$$\frac{1}{T} \sum_{k=0}^{K} \|\nabla f(x^k)\| \le \frac{2\Delta}{\gamma T} + 2L_0 \gamma.$$

Substituting $\gamma = \gamma_0 T^{-1/2}$ yields (40). ■

Corollary 4.7 ($\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ smoothness, deterministic)**.** *Suppose f is bounded below and 4 hold with fixed $\alpha \in (0, 1)$, and that the oracle is deterministic ($B = G = 0$). Consider the normalized gradient descent update*

$$x^{k+1} = x^k - \gamma \, \frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}.$$

Choose $\gamma = \gamma_0 \, T^{-1/2}$ with $0 < \gamma_0 \le 1$. Then

$$\|\nabla f(\widehat{x})\| \le \frac{4\Delta}{3\gamma_0} \, T^{-1/2} + \frac{2K_0}{3} \gamma_0 \, T^{-1/2} + \frac{4C_\alpha}{3} \gamma_0^{1/(1-\alpha)} \, T^{-1/(2(1-\alpha))}. \tag{41}$$

Consequently, $\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/2})$, and the iteration complexity is $\mathcal{O}(\varepsilon^{-2})$.

Proof. Under determinism, $v^k = \nabla f(x^k)$, so $\|e^k\| = 0$ and $d^k = \nabla f(x^k)/\|\nabla f(x^k)\|$. The one-step descent (29) from Lemma 4 becomes

$$f(x^{k+1}) \le f(x^k) - \frac{3\gamma}{4} \|\nabla f(x^k)\| + \frac{K_0}{2} \gamma^2 + C_\alpha \gamma^{(2-\alpha)/(1-\alpha)}.$$

Summing from $k = 0$ to K and using $f(x^{K+1}) \ge f^{\inf}$,

$$\frac{3\gamma}{4} \sum_{k=0}^{K} \|\nabla f(x^k)\| \le \Delta + \frac{K_0}{2} T\gamma^2 + T C_\alpha \gamma^{(2-\alpha)/(1-\alpha)}.$$

Dividing by $(3\gamma/4) \cdot T$,

$$\frac{1}{T} \sum_{k=0}^{K} \|\nabla f(x^k)\| \le \frac{4\Delta}{3\gamma T} + \frac{2K_0}{3} \gamma + \frac{4C_\alpha}{3} \gamma^{1/(1-\alpha)}.$$

Substituting $\gamma = \gamma_0 T^{-1/2}$ gives

$$\frac{4\Delta}{3\gamma T} = \frac{4\Delta}{3\gamma_0} \, T^{-1/2}, \qquad \frac{2K_0}{3} \gamma = \frac{2K_0}{3} \gamma_0 \, T^{-1/2},$$

and

$$\frac{4C_\alpha}{3} \gamma^{1/(1-\alpha)} = \frac{4C_\alpha}{3} \gamma_0^{1/(1-\alpha)} \, T^{-1/(2(1-\alpha))}.$$

Since $\alpha \in (0, 1)$, we have $1/(2(1-\alpha)) > 1/2$, so every term decays at least as fast as $T^{-1/2}$. This proves (41) and the $\mathcal{O}(T^{-1/2})$ rate. ■

B.6 Discussion on recovered rates

The six recovery corollaries above confirm the claims made in the discussion following Theorem 1.

- *Bounded variance recovery.* When $B = 0$, Corollaries 4.2, 4.3, and 4.4 all achieve $\mathcal{O}(\varepsilon^{-4})$ SFO complexity with the standard bounded variance schedule $\eta = T^{-1/2}$, $\gamma = \gamma_0 T^{-3/4}$. This recovers the rate of Cutkosky and Mehta [9] for standard smoothness, of Khirirat et al. [26] for $\mathcal{L}^*_{\mathrm{sym}}(1)$, and establishes the same rate for $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$ with $\alpha \in (0, 1)$, which to our knowledge is new.
- *Deterministic recovery.* When $B = G = 0$, Corollaries 4.5, 4.6, and 4.7 all achieve $\mathcal{O}(\varepsilon^{-2})$ iteration complexity. This recovers the normalized gradient descent rates of Chen et al. [7] for all three smoothness regimes.
- *Schedule transition.* Comparing the three noise regimes clarifies the role of the schedule parameters. As the noise weakens from BG-0 to bounded variance to deterministic, the schedules transition as:

$$(\eta, \gamma) = \begin{cases} (T^{-2/3},\ \gamma_0 T^{-5/6}), & \text{BG-0}, \\ (T^{-1/2},\ \gamma_0 T^{-3/4}), & \text{bounded variance}, \\ (1,\ \gamma_0 T^{-1/2}), & \text{deterministic}, \end{cases}$$

  and the corresponding convergence rates are $T^{-1/6}$, $T^{-1/4}$, and $T^{-1/2}$, respectively. The faster schedules are enabled by the absence of the distance-dependent noise growth term $B^2\|x^k - x^0\|^2$.

C Proofs for Normalized STORM under BG-0 Noise

This appendix proves Theorem 2. Throughout this section, let

$$T := K + 1, \qquad \Delta := f(x^0) - f^{\inf}.$$

We define

$$g_k := \mathbb{E}\|\nabla f(x^k)\|, \qquad b_k := \mathbb{E}\|v^k - \nabla f(x^k)\|,$$

and

$$\Delta^k := \mathbb{E}[f(x^k) - f^{\inf}], \qquad S_b(K) := \sum_{k=0}^{K} b_k.$$

The output $\widehat{x}$ is sampled uniformly from $\{x^0, x^1, \dots, x^K\}$, so

$$\mathbb{E}\|\nabla f(\widehat{x})\| = \frac{1}{T}\sum_{k=0}^{K} g_k. \tag{42}$$

For compactness, define the local BG-0 noise scale

$$q(x) := \left(B^2\|x - x^0\|^2 + G^2\right)^{1/2}, \qquad q_k := q(x^k).$$

C.1 Basic inequalities

Lemma 6 (Trajectory control)**.** *For the* NSTORM *iterates in* (3)*,*

$$\|x^{k+1} - x^k\| \le \gamma, \qquad \|x^k - x^0\| \le k\gamma.$$

Consequently,

$$q_k \le G + Bk\gamma \le G + BT\gamma, \qquad k = 0, \dots, K.$$

Moreover, for every $\beta \in (0, 1]$*,*

$$\frac{1}{T}\sum_{k=0}^{K} \mathbb{E}[q_k^\beta] \le G^\beta + B^\beta (T\gamma)^\beta. \tag{43}$$

Proof. The normalized direction has norm at most one, so

$$\|x^{k+1} - x^k\| \le \gamma.$$

Summing the increments gives

$$\|x^k - x^0\| \le \sum_{t=0}^{k-1} \|x^{t+1} - x^t\| \le k\gamma.$$

Therefore,

$$q_k = \left(B^2\|x^k - x^0\|^2 + G^2\right)^{1/2} \le B\|x^k - x^0\| + G \le G + Bk\gamma \le G + BT\gamma.$$

For $\beta \in (0, 1]$, using $(a + b)^\beta \le a^\beta + b^\beta$,

$$q_k^\beta \le G^\beta + B^\beta (k\gamma)^\beta \le G^\beta + B^\beta (T\gamma)^\beta.$$

Averaging over $k = 0, \dots, K$ gives (43). ■

Remark 1. *Lemma 3 with* $\rho = \lambda/\alpha$ *gives: for every* $\lambda > 0$ *and* $u \ge 0$*,* $u^\alpha \le \lambda u + c_\alpha \lambda^{-p}$*, where* $p := \alpha/(1 - \alpha)$ *and* $c_\alpha := (1 - \alpha)\alpha^{\alpha/(1-\alpha)}$*. This is the form used throughout the* NSTORM *proofs.*

C.2 Descent inequalities

Lemma 7 (Descent under mean-square smoothness)**.** *Suppose Assumption 3 holds. Then, for every $k = 0, \dots, K$,*

$$\Delta^{k+1} + \gamma g_k \le \Delta^k + 2\gamma b_k + \frac{L}{2}\gamma^2. \tag{44}$$

Proof. Assumption 3 and Jensen's inequality imply that f is L-smooth:

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

Thus,

$$f(x^{k+1}) \le f(x^k) - \gamma \left\langle \nabla f(x^k), d^k \right\rangle + \frac{L}{2}\gamma^2.$$

Using Lemma 2,

$$-\gamma \left\langle \nabla f(x^k), d^k \right\rangle \le -\gamma \|\nabla f(x^k)\| + 2\gamma \|v^k - \nabla f(x^k)\|.$$

Subtracting $f^{\inf}$ and taking expectations gives (44). ■

Lemma 8 (Sample-gradient moment under BG-0)**.** *For $\alpha \in (0, 1)$,*

$$\mathbb{E}_\xi \|\nabla f(x;\xi)\|^\alpha \le 2^{\alpha/2} \left(\|\nabla f(x)\|^\alpha + q(x)^\alpha \right).$$

Proof. By Jensen's inequality,

$$\mathbb{E}_\xi \|\nabla f(x;\xi)\|^\alpha \le \left(\mathbb{E}_\xi \|\nabla f(x;\xi)\|^2 \right)^{\alpha/2}.$$

Using

$$\nabla f(x;\xi) = \nabla f(x) + \big(\nabla f(x;\xi) - \nabla f(x) \big),$$

and Assumption BG-0,

$$\mathbb{E}_\xi \|\nabla f(x;\xi)\|^2 \le 2\|\nabla f(x)\|^2 + 2q(x)^2.$$

Since $\alpha/2 \in (0, 1)$,

$$\left(\|\nabla f(x)\|^2 + q(x)^2 \right)^{\alpha/2} \le \|\nabla f(x)\|^\alpha + q(x)^\alpha.$$

Combining the last three displays proves the claim. ■

Lemma 9 (Descent under expected α-symmetric generalized smoothness)**.** *Suppose Assumptions BG-0 and 5 hold with $\alpha \in (0, 1)$. Let*

$$p := \frac{\alpha}{1-\alpha}, \qquad K_0 := 2^{\frac{2-\alpha}{1-\alpha}} L_0, \qquad K_1 := 2^{\frac{2-\alpha}{1-\alpha}} L_1, \qquad K_2 := (5L_1)^{\frac{1}{1-\alpha}}.$$

Define

$$L_\alpha := 2^{\alpha/2} K_1, \qquad c_\alpha := (1-\alpha)\alpha^{\alpha/(1-\alpha)}.$$

Let $\lambda > 0$, and suppose

$$L_\alpha \lambda \gamma \le 1.$$

Then, for every $k = 0, \dots, K$,

$$\Delta^{k+1} + \frac{\gamma}{2} g_k \le \Delta^k + 2\gamma b_k + \frac{\gamma^2}{2} \left(K_0 + c_\alpha L_\alpha \lambda^{-p} + 2K_2 \gamma^p \right) + \frac{\gamma^2}{2} L_\alpha \mathbb{E}[q_k^\alpha]. \tag{45}$$

Proof. The expected generalized smoothness reduction gives (see Chen et al. [7], Proposition 4), for $y = x^{k+1} = x^k - \gamma d^k$,

$$f(x^{k+1}) \le f(x^k) - \gamma \langle \nabla f(x^k), d^k \rangle + \frac{\gamma^2}{2} \left(K_0 + K_1 \mathbb{E}_\xi \|\nabla f(x^k;\xi)\|^\alpha + 2K_2 \gamma^p \right).$$

By Lemma 8,

$$K_1 \mathbb{E}_\xi \|\nabla f(x^k;\xi)\|^\alpha \le L_\alpha \left(\|\nabla f(x^k)\|^\alpha + q_k^\alpha \right).$$

Using Lemma 2, subtracting $f^{\inf}$, and taking expectations,

$$\Delta^{k+1} \le \Delta^k - \gamma g_k + 2\gamma b_k + \frac{\gamma^2}{2} \left(K_0 + 2K_2 \gamma^p \right) + \frac{\gamma^2}{2} L_\alpha \mathbb{E} \|\nabla f(x^k)\|^\alpha + \frac{\gamma^2}{2} L_\alpha \mathbb{E}[q_k^\alpha].$$

Since $u \mapsto u^\alpha$ is concave,

$$\mathbb{E}\|\nabla f(x^k)\|^\alpha \le g_k^\alpha.$$

By Lemma 1,

$$g_k^\alpha \le \lambda g_k + c_\alpha \lambda^{-p}.$$

Therefore,

$$\frac{\gamma^2}{2} L_\alpha g_k^\alpha \le \frac{\gamma^2}{2} L_\alpha \lambda g_k + \frac{\gamma^2}{2} c_\alpha L_\alpha \lambda^{-p}.$$

The condition $L_\alpha \lambda \gamma \le 1$ gives

$$\frac{\gamma^2}{2} L_\alpha \lambda g_k \le \frac{\gamma}{2} g_k.$$

Moving this term to the left proves (45). ■

Lemma 10 (Descent under expected 1-symmetric generalized smoothness)**.** *Suppose Assumptions* BG*-0 and 5 hold with* $\alpha = 1$*. If*

$$0 < \gamma \le \frac{1}{4\sqrt{2} L_1},$$

then, for every $k = 0, \dots, K$*,*

$$\Delta^{k+1} + \frac{\gamma}{2} g_k \le \Delta^k + 2\gamma b_k + \sqrt{2} L_0 \gamma^2 + 2\sqrt{2} L_1 \gamma^2 \mathbb{E}[q_k]. \tag{46}$$

Proof. The expected $\alpha = 1$ generalized smoothness reduction yields the segment bound

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \|y - x\|^2 \left(\sqrt{2} L_0 + 2\sqrt{2} L_1 \|\nabla f(x)\| + 2\sqrt{2} L_1 q(x) \right),$$

whenever $\|y - x\| \le \gamma$ and $\gamma \le 1/(4\sqrt{2} L_1)$. Applying this with $x = x^k$, $y = x^{k+1}$, and using Lemma 2, we obtain

$$\begin{aligned} f(x^{k+1}) \le f(x^k) &- \gamma \|\nabla f(x^k)\| + 2\gamma \|v^k - \nabla f(x^k)\| + \sqrt{2} L_0 \gamma^2 + 2\sqrt{2} L_1 \gamma^2 \|\nabla f(x^k)\| \\ &+ 2\sqrt{2} L_1 \gamma^2 q_k. \end{aligned}$$

The stepsize condition implies

$$2\sqrt{2} L_1 \gamma^2 \|\nabla f(x^k)\| \le \frac{\gamma}{2} \|\nabla f(x^k)\|.$$

Subtracting $f^{\inf}$, taking expectations, and moving this term to the left proves (46). ■

C.3 STORM decomposition and centered difference bound

Lemma 11 (STORM decomposition)**.** *Let*

$$E_k := v^k - \nabla f(x^k).$$

Define

$$\begin{aligned} Z_{k+1} &:= \left(\nabla f(x^{k+1}; \xi^{k+1}) - \nabla f(x^k; \xi^{k+1}) \right) - \left(\nabla f(x^{k+1}) - \nabla f(x^k) \right), \\ \zeta_{k+1} &:= \nabla f(x^k; \xi^{k+1}) - \nabla f(x^k). \end{aligned}$$

Then

$$E_{k+1} = (1 - \eta) E_k + Z_{k+1} + \eta \zeta_{k+1}. \tag{47}$$

Moreover,

$$\mathbb{E}[Z_{k+1} \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}[\zeta_{k+1} \mid \mathcal{F}_k] = 0.$$

Proof. Subtract $\nabla f(x^{k+1})$ from the STORM update:

$$E_{k+1} = \nabla f(x^{k+1}; \xi^{k+1}) - \nabla f(x^{k+1}) + (1 - \eta) \left(v^k - \nabla f(x^k; \xi^{k+1}) \right).$$

Add and subtract $(1 - \eta) \nabla f(x^k)$. Then

$$\begin{aligned} E_{k+1} = (1 - \eta) \left(v^k - \nabla f(x^k) \right) &+ \left(\nabla f(x^{k+1}; \xi^{k+1}) - \nabla f(x^{k+1}) \right) \\ &- (1 - \eta) \left(\nabla f(x^k; \xi^{k+1}) - \nabla f(x^k) \right). \end{aligned}$$

The last two lines equal $Z_{k+1} + \eta \zeta_{k+1}$. The conditional mean-zero properties follow from unbiasedness and the freshness of ξ^{k+1}. ■

Lemma 12 (Centered Difference Bound)**.** *Suppose that for every $k = 0, \dots, K$,*

$$\left(\mathbb{E}[\|Z_{k+1}\|^2 \mid \mathcal{F}_k]\right)^{1/2} \le \gamma\left(a + h\|\nabla f(x^k)\|\right),$$

and

$$\left(\mathbb{E}[\|\zeta_{k+1}\|^2 \mid \mathcal{F}_k]\right)^{1/2} \le \tau.$$

Then

$$S_b(K) \le \frac{b_0}{\eta} + T\frac{a\gamma}{\sqrt{\eta}} + T\sqrt{\eta}\,\tau + \frac{h\gamma}{\eta}\sum_{k=0}^{K} g_k. \tag{48}$$

Proof. Unrolling (47), for $k \ge 1$,

$$E_k = (1-\eta)^k E_0 + \sum_{t=0}^{k-1}(1-\eta)^{k-1-t} Z_{t+1} + \eta\sum_{t=0}^{k-1}(1-\eta)^{k-1-t}\zeta_{t+1}.$$

The Z-sum is split into a gradient-independent part and a gradient-dependent part. Let

$$A := \gamma a, \qquad H_t := \gamma h\|\nabla f(x^t)\|.$$

If $A + H_t > 0$, define

$$\lambda_t := \frac{A}{A + H_t};$$

otherwise set $\lambda_t := 0$. Then

$$Z_{t+1}^{(0)} := \lambda_t Z_{t+1}, \qquad Z_{t+1}^{(1)} := (1-\lambda_t) Z_{t+1}$$

are martingale differences and satisfy

$$\left(\mathbb{E}[\|Z_{t+1}^{(0)}\|^2 \mid \mathcal{F}_t]\right)^{1/2} \le A, \qquad \left(\mathbb{E}[\|Z_{t+1}^{(1)}\|^2 \mid \mathcal{F}_t]\right)^{1/2} \le H_t.$$

By martingale orthogonality,

$$\mathbb{E}\left\|\sum_{t=0}^{k-1}(1-\eta)^{k-1-t} Z_{t+1}^{(0)}\right\| \le \frac{\gamma a}{\sqrt{\eta}}.$$

For the gradient-dependent part, triangle inequality gives

$$\mathbb{E}\left\|\sum_{t=0}^{k-1}(1-\eta)^{k-1-t} Z_{t+1}^{(1)}\right\| \le \gamma h\sum_{t=0}^{k-1}(1-\eta)^{k-1-t} g_t.$$

Similarly, for the fresh-noise martingale term,

$$\mathbb{E}\left\|\eta\sum_{t=0}^{k-1}(1-\eta)^{k-1-t}\zeta_{t+1}\right\| \le \sqrt{\eta}\,\tau.$$

Thus,

$$b_k \le (1-\eta)^k b_0 + \frac{\gamma a}{\sqrt{\eta}} + \sqrt{\eta}\,\tau + \gamma h\sum_{t=0}^{k-1}(1-\eta)^{k-1-t} g_t.$$

Summing over $k = 0, \dots, K$ and using

$$\sum_{k=0}^{K}(1-\eta)^k \le \frac{1}{\eta}, \qquad \sum_{k=0}^{K}\sum_{t=0}^{k-1}(1-\eta)^{k-1-t} g_t \le \frac{1}{\eta}\sum_{t=0}^{K} g_t,$$

gives (48). ∎

C.4 Estimator-difference bounds

Lemma 13 (Single-sample estimator bounds)**.** *Let*

$$Q_T := G + BT\gamma.$$

The following bounds hold.

(i) Under Assumption 3,

$$\big(\mathbb{E}[\|Z_{k+1}\|^2 \mid \mathcal{F}_k]\big)^{1/2} \le L\gamma, \qquad \big(\mathbb{E}[\|\zeta_{k+1}\|^2 \mid \mathcal{F}_k]\big)^{1/2} \le Q_T.$$

(ii) Under Assumption 5 with $\alpha \in (0,1)$, define

$$p := \frac{\alpha}{1-\alpha}, \qquad K_0 := 2^{\frac{2-\alpha}{1-\alpha}} L_0, \qquad K_1 := 2^{\frac{2-\alpha}{1-\alpha}} L_1, \qquad K_2 := (5L_1)^{\frac{1}{1-\alpha}},$$

and

$$L_\alpha := 2^{\alpha/2} K_1, \qquad c_\alpha := (1-\alpha)\alpha^{\alpha/(1-\alpha)}.$$

Then, for every $\lambda > 0$,

$$\big(\mathbb{E}[\|Z_{k+1}\|^2 \mid \mathcal{F}_k]\big)^{1/2} \le \gamma\left(K_0 + c_\alpha L_\alpha \lambda^{-p} + L_\alpha Q_T^\alpha + K_2\gamma^p + L_\alpha\lambda\|\nabla f(x^k)\|\right),$$

and

$$\big(\mathbb{E}[\|\zeta_{k+1}\|^2 \mid \mathcal{F}_k]\big)^{1/2} \le Q_T.$$

(iii) Under Assumption 5 with $\alpha = 1$, if $\gamma \le 1/(4L_1)$, then

$$\big(\mathbb{E}[\|Z_{k+1}\|^2 \mid \mathcal{F}_k]\big)^{1/2} \le \sqrt{2e^{3/4}}\gamma\left(L_0 + 2L_1 Q_T + 2L_1\|\nabla f(x^k)\|\right),$$

and

$$\big(\mathbb{E}[\|\zeta_{k+1}\|^2 \mid \mathcal{F}_k]\big)^{1/2} \le Q_T.$$

Proof. For item (i), conditioning on $\mathcal{F}_k$, mean-square smoothness gives

$$\mathbb{E}[\|Z_{k+1}\|^2 \mid \mathcal{F}_k] \le \mathbb{E}_{\xi^{k+1}}\|\nabla f(x^{k+1};\xi^{k+1}) - \nabla f(x^k;\xi^{k+1})\|^2 \le L^2\gamma^2.$$

The fresh-noise bound follows from Assumption BG-0 and Lemma 6.

For item (ii), conditioning on $\mathcal{F}_k$ and using the expected generalized smoothness reduction,

$$\big(\mathbb{E}[\|Z_{k+1}\|^2 \mid \mathcal{F}_k]\big)^{1/2} \le \gamma\left(K_0 + K_1\mathbb{E}_\xi\|\nabla f(x^k;\xi)\|^\alpha + K_2\gamma^p\right).$$

By Lemma 8,

$$K_1\mathbb{E}_\xi\|\nabla f(x^k;\xi)\|^\alpha \le L_\alpha\left(\|\nabla f(x^k)\|^\alpha + q_k^\alpha\right).$$

Using $q_k \le Q_T$ and

$$\|\nabla f(x^k)\|^\alpha \le \lambda\|\nabla f(x^k)\| + c_\alpha\lambda^{-p},$$

gives the displayed Z-bound. The bound for ζ_{k+1} follows from BG-0 and $q_k \le Q_T$.

Item (iii) is the corresponding $\alpha = 1$ reduction. It gives

$$\big(\mathbb{E}[\|Z_{k+1}\|^2 \mid \mathcal{F}_k]\big)^{1/2} \le \sqrt{2e^{3/4}}\gamma\left(L_0 + 2L_1 q_k + 2L_1\|\nabla f(x^k)\|\right),$$

and $q_k \le Q_T$ completes the proof. ∎

C.5 Single-sample master bounds

Theorem 5 (Single-sample NSTORM master bounds)**.** *Consider the single-sample* NSTORM *update in* (3)*. Suppose f is bounded below and* BG-*0 hold. Then the following statements hold.*

(i) Suppose Assumption 3 holds. Then

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \frac{\Delta}{\gamma T} + \frac{2b_0}{\eta T} + \frac{2L\gamma}{\sqrt{\eta}} + 2\sqrt{\eta}(G + BT\gamma) + \frac{L}{2}\gamma. \tag{49}$$

(ii) Suppose Assumption 5 holds with $\alpha \in (0,1)$. Define

$$p := \frac{\alpha}{1-\alpha}, \qquad K_0 := 2^{\frac{2-\alpha}{1-\alpha}} L_0, \qquad K_1 := 2^{\frac{2-\alpha}{1-\alpha}} L_1, \qquad K_2 := (5L_1)^{\frac{1}{1-\alpha}},$$

and

$$L_\alpha := 2^{\alpha/2} K_1, \qquad c_\alpha := (1-\alpha)\alpha^{\alpha/(1-\alpha)}.$$

Let

$$Q_T := G + BT\gamma.$$

If

$$L_\alpha \lambda \gamma \le 1, \qquad 8L_\alpha \lambda \frac{\gamma}{\eta} \le 1,$$

then

$$\begin{aligned} \mathbb{E}\|\nabla f(\widehat{x})\| \le \frac{4\Delta}{\gamma T} + \frac{8b_0}{\eta T} + \frac{8\gamma}{\sqrt{\eta}} \left(K_0 + c_\alpha L_\alpha \lambda^{-p} + L_\alpha Q_T^\alpha + K_2 \gamma^p\right) + 8\sqrt{\eta}\, Q_T \\ + 2\gamma \left(K_0 + c_\alpha L_\alpha \lambda^{-p} + 2K_2\gamma^p\right) + 2L_\alpha \gamma \frac{1}{T} \sum_{k=0}^{K} \mathbb{E}[q_k^\alpha]. \end{aligned} \tag{50}$$

(iii) Suppose Assumption 5 holds with $\alpha = 1$. If

$$\gamma \le \frac{1}{4\sqrt{2}L_1}, \qquad 16\sqrt{2e^{3/4}} L_1 \frac{\gamma}{\eta} \le 1,$$

then

$$\begin{aligned} \mathbb{E}\|\nabla f(\widehat{x})\| \le \frac{4\Delta}{\gamma T} + \frac{8b_0}{\eta T} + \frac{8\gamma}{\sqrt{\eta}} \sqrt{2e^{3/4}} \left(L_0 + 2L_1(G + BT\gamma)\right) + 8\sqrt{\eta}(G + BT\gamma) \\ + 4\sqrt{2}L_0\gamma + 8\sqrt{2}L_1\gamma \frac{1}{T} \sum_{k=0}^{K} \mathbb{E}[q_k]. \end{aligned} \tag{51}$$

Proof. For item (i), summing (44) gives

$$\gamma \sum_{k=0}^{K} g_k \le \Delta + 2\gamma S_b(K) + \frac{L}{2}\gamma^2 T.$$

Apply Lemma 12 with

$$a = L, \qquad h = 0, \qquad \tau = G + BT\gamma.$$

Then

$$S_b(K) \le \frac{b_0}{\eta} + T\frac{L\gamma}{\sqrt{\eta}} + T\sqrt{\eta}(G + BT\gamma).$$

Substituting this into the summed descent inequality and dividing by γT gives (49).

For item (ii), summing (45) gives

$$\frac{\gamma}{2} \sum_{k=0}^{K} g_k \le \Delta + 2\gamma S_b(K) + \frac{\gamma^2 T}{2} \left(K_0 + c_\alpha L_\alpha \lambda^{-p} + 2K_2\gamma^p\right) + \frac{\gamma^2}{2} L_\alpha \sum_{k=0}^{K} \mathbb{E}[q_k^\alpha].$$

Apply Lemma 12 with

$$a = K_0 + c_\alpha L_\alpha \lambda^{-p} + L_\alpha Q_T^\alpha + K_2\gamma^p, \qquad h = L_\alpha \lambda, \qquad \tau = Q_T.$$

Then

$$S_b(K) \le \frac{b_0}{\eta} + T\frac{a\gamma}{\sqrt{\eta}} + T\sqrt{\eta} Q_T + \frac{L_\alpha \lambda \gamma}{\eta} \sum_{k=0}^{K} g_k.$$

Substituting this bound into the descent inequality gives a feedback term

$$2\gamma \cdot \frac{L_\alpha \lambda \gamma}{\eta} \sum_{k=0}^{K} g_k.$$

The condition

$$8L_\alpha\lambda\frac{\gamma}{\eta}\le 1$$

implies that this feedback can be absorbed into the left-hand side. After absorption and multiplication by 4, division by γT gives (50).

For item (iii), sum (46):

$$\frac{\gamma}{2}\sum_{k=0}^{K} g_k \le \Delta + 2\gamma S_b(K) + \sqrt{2}L_0\gamma^2 T + 2\sqrt{2}L_1\gamma^2\sum_{k=0}^{K}\mathbb{E}[q_k].$$

Apply Lemma 12 with

$$a=\sqrt{2e^{3/4}}\left(L_0+2L_1(G+BT\gamma)\right),\qquad h=2\sqrt{2e^{3/4}}L_1,\qquad \tau=G+BT\gamma.$$

The condition

$$16\sqrt{2e^{3/4}}L_1\frac{\gamma}{\eta}\le 1$$

absorbs the resulting feedback term. Dividing by $\gamma T/4$ gives (51). ■

C.6 Proof of Theorem 2

Because the initialization samples are independent and the stochastic oracle is unbiased,

$$\mathbb{E}\|v^0-\nabla f(x^0)\|^2=\frac{1}{N_{\text{init}}}\mathbb{E}_\xi\|\nabla f(x^0;\xi)-\nabla f(x^0)\|^2=\frac{\sigma_0^2}{N_{\text{init}}}.$$

Consequently,

$$b_0=\mathbb{E}\|v^0-\nabla f(x^0)\|\le\frac{\sigma_0}{\sqrt{N_{\text{init}}}}.$$

We use this initialization bound in all three smoothness regimes below.

Corollary 5.1 (Mean-square smoothness with sharp initialization)**.** *Suppose f is bounded below,* BG-*0, and 3 hold. Initialize*

$$v^0=\frac{1}{N_{\text{init}}}\sum_{i=1}^{N_{\text{init}}}\nabla f(x^0;\xi_i^{\text{init}}),\qquad N_{\text{init}}:=\left\lceil T^{1/2}\right\rceil.$$

Choose

$$\eta=T^{-1},\qquad \gamma=\gamma_0T^{-3/4}.$$

Then

$$\mathbb{E}\|\nabla f(\widehat{x})\|\le\left(\frac{\Delta}{\gamma_0}+2\sigma_0+\frac{5}{2}L\gamma_0+2B\gamma_0\right)T^{-1/4}+2GT^{-1/2}.$$

Consequently,

$$\mathbb{E}\|\nabla f(\widehat{x})\|=\mathcal{O}(T^{-1/4}),$$

and the SFO complexity is $\mathcal{O}(\varepsilon^{-4})$.

Proof. By the initialization choice,

$$b_0=\mathbb{E}\|v^0-\nabla f(x^0)\|\le\frac{\sigma_0}{\sqrt{N_{\text{init}}}}\le\sigma_0T^{-1/4}.$$

Apply (49). Under the schedule,

$$\frac{\Delta}{\gamma T}=\frac{\Delta}{\gamma_0}T^{-1/4},\qquad \frac{2b_0}{\eta T}=2b_0\le 2\sigma_0T^{-1/4}.$$

Also,

$$\frac{2L\gamma}{\sqrt{\eta}}=2L\gamma_0T^{-1/4},\qquad \frac{L}{2}\gamma=\frac{L\gamma_0}{2}T^{-3/4},$$

and

$$2\sqrt{\eta}(G+BT\gamma)=2GT^{-1/2}+2B\gamma_0T^{-1/4}.$$

Combining these estimates proves the displayed bound. Since $T \geq 1$,

$$\frac{L\gamma_0}{2}T^{-3/4} \leq \frac{L\gamma_0}{2}T^{-1/4}.$$

Combining the estimates gives the coefficient $\frac{5}{2}L\gamma_0$ and proves the compact bound in Theorem 2(i). Since $N_{\text{init}} = \mathcal{O}(T^{1/2})$ and each transition uses two SFO calls, the total SFO cost is $N_{\text{init}} + 2T = \mathcal{O}(T)$. Setting $T = \mathcal{O}(\varepsilon^{-4})$ gives $\mathcal{O}(\varepsilon^{-4})$. ■

Corollary 5.2 (Expected α-symmetric generalized smoothness)**.** *Suppose f is bounded below,* BG-*0, and 5 hold with $\alpha \in (0,1)$. Define*

$$p := \frac{\alpha}{1-\alpha}, \qquad K_0 := 2^{\frac{2-\alpha}{1-\alpha}}L_0, \qquad K_1 := 2^{\frac{2-\alpha}{1-\alpha}}L_1, \qquad K_2 := (5L_1)^{\frac{1}{1-\alpha}},$$

and let

$$T := K + 1, \qquad \theta := \frac{1}{4+\alpha}.$$

Initialize

$$v^0 = \frac{1}{N_{\text{init}}}\sum_{i=1}^{N_{\text{init}}} \nabla f(x^0; \xi_i^{\text{init}}), \qquad N_{\text{init}} := \left\lceil T^{2(1-\alpha)\theta} \right\rceil.$$

Choose

$$\gamma = \gamma_0 T^{-(3+\alpha)\theta}, \qquad \eta = \eta_0 T^{-4\theta}, \qquad \lambda = \lambda_0 T^{-(1-\alpha)\theta},$$

where $\eta_0 \in (0,1]$, $\gamma_0 > 0$, and $\lambda_0 > 0$. Suppose

$$2^{\alpha/2}K_1\lambda_0\gamma_0 \leq 1, \qquad 2^{(6+\alpha)/2}K_1\lambda_0\frac{\gamma_0}{\eta_0} \leq 1.$$

Then the bound stated in Theorem 2(ii) holds. In particular,

$$\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/(4+\alpha)}),$$

and the SFO complexity is $\mathcal{O}(\varepsilon^{-(4+\alpha)})$.

Proof. By the initialization choice,

$$b_0 \leq \frac{\sigma_0}{\sqrt{N_{\text{init}}}} \leq \sigma_0 T^{-(1-\alpha)\theta}.$$

Let

$$L_\alpha := 2^{\alpha/2}K_1, \qquad c_\alpha := (1-\alpha)\alpha^{\alpha/(1-\alpha)}.$$

The two conditions in the corollary are exactly

$$L_\alpha\lambda\gamma \leq 1, \qquad 8L_\alpha\lambda\frac{\gamma}{\eta} \leq 1.$$

Therefore, (50) applies.

We now substitute the parameter choices. First,

$$\frac{4\Delta}{\gamma T} = \frac{4\Delta}{\gamma_0}T^{-\theta}, \qquad \frac{8b_0}{\eta T} \leq \frac{8\sigma_0}{\eta_0}T^{-\theta}.$$

Also,

$$Q_T = G + BT\gamma = G + B\gamma_0 T^\theta.$$

Using $(a+b)^\alpha \leq a^\alpha + b^\alpha$,

$$Q_T^\alpha \leq G^\alpha + B^\alpha\gamma_0^\alpha T^{\alpha\theta}.$$

Substituting these estimates into (50) gives exactly

$$\begin{aligned}\mathbb{E}\|\nabla f(\widehat{x})\| \le \Bigg[&\frac{4\Delta}{\gamma_0}+\frac{8\sigma_0}{\eta_0}+\frac{8(1-\alpha)\alpha^{\frac{\alpha}{1-\alpha}}2^{\alpha/2}K_1\gamma_0\lambda_0^{-\frac{\alpha}{1-\alpha}}}{\sqrt{\eta_0}}\\&+\frac{8\cdot 2^{\alpha/2}K_1B^\alpha\gamma_0^{1+\alpha}}{\sqrt{\eta_0}}+8\sqrt{\eta_0}B\gamma_0\Bigg]T^{-\theta}\\&+\left[\frac{8K_0\gamma_0}{\sqrt{\eta_0}}+\frac{8\cdot 2^{\alpha/2}K_1G^\alpha\gamma_0}{\sqrt{\eta_0}}\right]T^{-(1+\alpha)\theta}+8\sqrt{\eta_0}GT^{-2\theta}\\&+\frac{8K_2\gamma_0^{\frac{1}{1-\alpha}}}{\sqrt{\eta_0}}T^{-\frac{1+3\alpha}{(1-\alpha)(4+\alpha)}}+2K_0\gamma_0T^{-(3+\alpha)\theta}\\&+2(1-\alpha)\alpha^{\frac{\alpha}{1-\alpha}}2^{\alpha/2}K_1\gamma_0\lambda_0^{-\frac{\alpha}{1-\alpha}}T^{-3\theta}\\&+4K_2\gamma_0^{\frac{1}{1-\alpha}}T^{-\frac{3+\alpha}{(1-\alpha)(4+\alpha)}}+2^{1+\alpha/2}K_1G^\alpha\gamma_0T^{-(3+\alpha)\theta}\\&+2^{1+\alpha/2}K_1B^\alpha\gamma_0^{1+\alpha}T^{-3\theta}.\end{aligned}$$

For the compact statement of Theorem 2(ii), define the horizon-independent constant

$$\Phi_\alpha := \frac{5}{8}\Big(K_0+c_\alpha L_\alpha\lambda_0^{-p}+L_\alpha G^\alpha+L_\alpha B^\alpha\gamma_0^\alpha\Big)+\frac{3}{4}K_2\gamma_0^p,$$

where $p=\alpha/(1-\alpha)$, $L_\alpha=2^{\alpha/2}K_1$, and $c_\alpha=(1-\alpha)\alpha^{\alpha/(1-\alpha)}$.

Since $T\ge 1$ and $\eta_0\le 1$, the terms involving K_0, $c_\alpha L_\alpha\lambda_0^{-p}$, $L_\alpha G^\alpha$, and $L_\alpha B^\alpha\gamma_0^\alpha$ are bounded in total by

$$\frac{10\gamma_0}{\sqrt{\eta_0}}\left(K_0+c_\alpha L_\alpha\lambda_0^{-p}+L_\alpha G^\alpha+L_\alpha B^\alpha\gamma_0^\alpha\right)T^{-\theta}.$$

The two K_2 contributions are bounded in total by $\frac{12K_2\gamma_0^{p+1}}{\sqrt{\eta_0}}T^{-\theta}$. Moreover, $8\sqrt{\eta_0}GT^{-2\theta}\le 8\sqrt{\eta_0}GT^{-\theta}$. Combining these bounds, we obtain

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \left(\frac{4\Delta}{\gamma_0}+\frac{8\sigma_0}{\eta_0}+8\sqrt{\eta_0}(G+B\gamma_0)+\frac{16\gamma_0}{\sqrt{\eta_0}}\Phi_\alpha\right)T^{-\theta}.$$

This is precisely the finite-horizon bound stated in Theorem 2(ii). In particular,

$$\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/(4+\alpha)}).$$

The initialization cost satisfies

$$N_{\text{init}} = \mathcal{O}(T^{2(1-\alpha)/(4+\alpha)}) = o(T),$$

and each transition uses two SFO calls. Hence the total SFO cost is

$$N_{\text{init}}+2T=\mathcal{O}(T).$$

Taking $T=\mathcal{O}(\varepsilon^{-(4+\alpha)})$ gives $\mathcal{O}(\varepsilon^{-(4+\alpha)})$. ■

Corollary 5.3 (Expected 1-symmetric generalized smoothness)**.** *Suppose f is bounded below,* BG-*0, and 5 hold with $\alpha=1$. Use one-sample initialization, so that*

$$b_0 := \mathbb{E}\|v^0-\nabla f(x^0)\|\le\sigma_0.$$

Choose

$$\eta=T^{-4/5},\qquad \gamma=\gamma_0T^{-4/5},\qquad 0<\gamma_0\le\frac{1}{16\sqrt{2e^{3/4}}L_1}.$$

Then

$$\begin{aligned}\mathbb{E}\|\nabla f(\widehat{x})\| \le &\left(\frac{4\Delta}{\gamma_0}+8b_0+8B\gamma_0+16\sqrt{2e^{3/4}}L_1B\gamma_0^2\right)T^{-1/5}\\&+\left(8\sqrt{2e^{3/4}}\gamma_0(L_0+2L_1G)+8G\right)T^{-2/5}\\&+8\sqrt{2}L_1B\gamma_0^2T^{-3/5}+\left(4\sqrt{2}L_0\gamma_0+8\sqrt{2}L_1G\gamma_0\right)T^{-4/5}.\end{aligned}$$

Consequently,

$$\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/5}),$$

and the SFO complexity is $\mathcal{O}(\varepsilon^{-5})$.

Proof. The stepsize condition implies

$$\gamma \le \frac{1}{4\sqrt{2}L_1}, \qquad 16\sqrt{2e^{3/4}}L_1\frac{\gamma}{\eta} = 16\sqrt{2e^{3/4}}L_1\gamma_0 \le 1.$$

Thus (51) applies. Under the schedule,

$$\frac{4\Delta}{\gamma T} = \frac{4\Delta}{\gamma_0}T^{-1/5}, \qquad \frac{8b_0}{\eta T} = 8b_0T^{-1/5}.$$

Moreover,

$$Q_T = G + BT\gamma = G + B\gamma_0T^{1/5}.$$

Therefore,

$$\frac{8\gamma}{\sqrt{\eta}}\sqrt{2e^{3/4}}(L_0 + 2L_1Q_T) = 8\sqrt{2e^{3/4}}\gamma_0(L_0 + 2L_1G)T^{-2/5} + 16\sqrt{2e^{3/4}}L_1B\gamma_0^2T^{-1/5},$$

and

$$8\sqrt{\eta}Q_T = 8GT^{-2/5} + 8B\gamma_0T^{-1/5}.$$

Finally, by Lemma 6,

$$\frac{1}{T}\sum_{k=0}^{K}\mathbb{E}[q_k] \le G + BT\gamma = G + B\gamma_0T^{1/5}.$$

Hence

$$8\sqrt{2}L_1\gamma\frac{1}{T}\sum_{k=0}^{K}\mathbb{E}[q_k] \le 8\sqrt{2}L_1G\gamma_0T^{-4/5} + 8\sqrt{2}L_1B\gamma_0^2T^{-3/5}.$$

Together with $4\sqrt{2}L_0\gamma = 4\sqrt{2}L_0\gamma_0T^{-4/5}$, this proves the displayed bound.

To recover the compact statement of Theorem 2(iii), note that the stepsize restriction implies

$$16\sqrt{2e^{3/4}}L_1\gamma_0 \le 1, \qquad 8\sqrt{2}L_1\gamma_0 \le 1.$$

Moreover,

$$8\sqrt{2e^{3/4}} + 4\sqrt{2} \le 23.$$

Using these inequalities, $b_0 \le \sigma_0$, and $T \ge 1$, we obtain

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \left(\frac{4\Delta}{\gamma_0} + 8\sigma_0 + 10B\gamma_0\right)T^{-1/5} + (23L_0\gamma_0 + 10G)\,T^{-2/5}.$$

This proves the bound in Theorem 2(iii). The method uses one SFO call for initialization and two SFO calls per transition, so the total cost is $1 + 2T = \mathcal{O}(T)$. Taking $T = \mathcal{O}(\varepsilon^{-5})$ gives $\mathcal{O}(\varepsilon^{-5})$. ■

C.7 Recovery of bounded variance rates ($B = 0$)

When $B = 0$, the BG-0 condition reduces to bounded variance $\mathbb{E}_\xi\|\nabla f(x;\xi) - \nabla f(x)\|^2 \le G^2$, and the local noise scale simplifies to $q_k = G$ for every k. In particular, the trajectory-dependent terms $BT\gamma$ vanish from the master bounds. Under the bounded variance schedule $\eta = \eta_0T^{-2/3}$, $\gamma = \gamma_0T^{-2/3}$, NSTORM recovers the standard variance-reduced rate $\mathcal{O}(\varepsilon^{-3})$ for all three smoothness regimes, matching the SPIDER rate of Fang et al. [12], Chen et al. [7].

Corollary 5.4 (Mean-square smoothness under bounded variance)**.** *Suppose f is bounded below, BG-0 (with $B = 0$), and 3 hold. Use one-sample initialization $v^0 = \nabla f(x^0;\xi^0)$, so that $b_0 \le G$. Choose*

$$\eta = \eta_0\,T^{-2/3}, \qquad \gamma = \gamma_0\,T^{-2/3},$$

where $\eta_0 \in (0,1]$ and $\gamma_0 > 0$. Then

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \left(\frac{\Delta}{\gamma_0} + \frac{2b_0}{\eta_0} + \frac{2L\gamma_0}{\sqrt{\eta_0}} + 2\sqrt{\eta_0}\,G\right)T^{-1/3} + \frac{L\gamma_0}{2}\,T^{-2/3}. \tag{52}$$

Consequently, $\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/3})$, and the SFO complexity is $\mathcal{O}(\varepsilon^{-3})$.

Proof. Since $B = 0$, we have $q_k = G$ for every k and $Q_T = G$. Apply (49) from Theorem 5(i). Under the schedule $\eta = \eta_0 T^{-2/3}$ and $\gamma = \gamma_0 T^{-2/3}$,

$$\frac{\Delta}{\gamma T} = \frac{\Delta}{\gamma_0}\, T^{-1/3}, \qquad \frac{2b_0}{\eta T} = \frac{2b_0}{\eta_0}\, T^{-1/3},$$
$$\frac{2L\gamma}{\sqrt{\eta}} = \frac{2L\gamma_0}{\sqrt{\eta_0}}\, T^{-1/3}, \qquad \frac{L}{2}\gamma = \frac{L\gamma_0}{2}\, T^{-2/3},$$

and, since $B = 0$,

$$2\sqrt{\eta}(G + BT\gamma) = 2\sqrt{\eta_0}\, G\, T^{-1/3}.$$

Substitution gives (52). The method uses one SFO call for initialization and two per transition, giving $M_K = 1 + 2T = \mathcal{O}(\varepsilon^{-3})$. ■

Corollary 5.5 (Expected α-symmetric generalized smoothness under bounded variance)**.** *Suppose f is bounded below,* BG-*0 (with $B = 0$), and 5 hold with $\alpha \in (0, 1)$. Define $L_\alpha := 2^{\alpha/2} K_1$, $c_\alpha := (1-\alpha)\alpha^{\alpha/(1-\alpha)}$, $p := \alpha/(1-\alpha)$. Use one-sample initialization, so that $b_0 \le G$. Choose*

$$\eta = \eta_0\, T^{-2/3}, \qquad \gamma = \gamma_0\, T^{-2/3}, \qquad \lambda = \lambda_0,$$

where $\eta_0 \in (0, 1]$, $\gamma_0 > 0$, and $\lambda_0 > 0$ satisfy

$$L_\alpha \lambda_0 \gamma_0 \le 1, \qquad 8 L_\alpha \lambda_0 \frac{\gamma_0}{\eta_0} \le 1.$$

Then

$$\begin{aligned}\mathbb{E}\|\nabla f(\widehat{x})\| \le{}& \left[\frac{4\Delta}{\gamma_0} + \frac{8b_0}{\eta_0} + \frac{8\gamma_0}{\sqrt{\eta_0}}\left(K_0 + c_\alpha L_\alpha \lambda_0^{-p} + L_\alpha G^\alpha\right) + 8\sqrt{\eta_0}\, G\right] T^{-1/3} \\ &+ \frac{8K_2\gamma_0^{p+1}}{\sqrt{\eta_0}}\, T^{-(1+2p)/3} + 2\gamma_0\left(K_0 + c_\alpha L_\alpha \lambda_0^{-p}\right) T^{-2/3} \\ &+ 4K_2\gamma_0^{p+1}\, T^{-2(p+1)/3} + 2L_\alpha G^\alpha \gamma_0\, T^{-2/3}. \end{aligned} \tag{53}$$

Consequently, $\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/3})$, and the SFO complexity is $\mathcal{O}(\varepsilon^{-3})$.

Proof. Since $B = 0$, we have $q_k = G$ for every k, $Q_T = G$, and

$$\frac{1}{T}\sum_{k=0}^{K} \mathbb{E}[q_k^\alpha] = G^\alpha.$$

We verify the conditions of Theorem 5(ii). Since $T \ge 1$,

$$L_\alpha \lambda \gamma = L_\alpha \lambda_0 \gamma_0\, T^{-2/3} \le L_\alpha \lambda_0 \gamma_0 \le 1,$$

and

$$8 L_\alpha \lambda \frac{\gamma}{\eta} = 8 L_\alpha \lambda_0 \frac{\gamma_0}{\eta_0} \le 1.$$

Apply (50). Since $B = 0$, $Q_T = G$ is constant. The first two terms give

$$\frac{4\Delta}{\gamma T} = \frac{4\Delta}{\gamma_0}\, T^{-1/3}, \qquad \frac{8b_0}{\eta T} = \frac{8b_0}{\eta_0}\, T^{-1/3}.$$

The fresh-noise term gives

$$8\sqrt{\eta}\, G = 8\sqrt{\eta_0}\, G\, T^{-1/3}.$$

For the estimator-difference term,

$$\begin{aligned}\frac{8\gamma}{\sqrt{\eta}}\left(K_0 + c_\alpha L_\alpha \lambda_0^{-p} + L_\alpha G^\alpha + K_2\gamma^p\right) = {}& \frac{8\gamma_0}{\sqrt{\eta_0}}\left(K_0 + c_\alpha L_\alpha \lambda_0^{-p} + L_\alpha G^\alpha\right) T^{-1/3} \\ &+ \frac{8K_2\gamma_0^{p+1}}{\sqrt{\eta_0}} T^{-(1+2p)/3}.\end{aligned}$$

The descent-remainder term is

$$2\gamma\left(K_0 + c_\alpha L_\alpha \lambda_0^{-p} + 2K_2\gamma^p\right) = 2\gamma_0\left(K_0 + c_\alpha L_\alpha \lambda_0^{-p}\right) T^{-2/3} + 4K_2\gamma_0^{p+1} T^{-2(p+1)/3}.$$

The averaged BG-0 term gives

$$2L_\alpha \gamma\, G^\alpha = 2L_\alpha G^\alpha \gamma_0\, T^{-2/3}.$$

Since $p > 0$, the exponents $(1+2p)/3 > 1/3$ and $2(p+1)/3 > 2/3$, so the dominant rate is $T^{-1/3}$. This proves (53).

The method uses one SFO call for initialization and two per transition, giving $M_K = 1 + 2T = \mathcal{O}(\varepsilon^{-3})$. ■

Corollary 5.6 (Expected 1-symmetric generalized smoothness under bounded variance)**.** *Suppose f is bounded below,* BG-*0 (with $B = 0$), and 5 hold with $\alpha = 1$ and constants $L_0, L_1 > 0$. Use one-sample initialization, so that $b_0 \le G$. Choose*

$$\eta = \eta_0\, T^{-2/3}, \qquad \gamma = \gamma_0\, T^{-2/3},$$

where $\eta_0 \in (0, 1]$ and $\gamma_0 > 0$. If $L_1 > 0$, assume

$$\gamma_0 \le \frac{1}{4\sqrt{2}\, L_1}, \qquad 16\sqrt{2e^{3/4}}\, L_1 \frac{\gamma_0}{\eta_0} \le 1.$$

Then

$$\mathbb{E}\|\nabla f(\widehat{x})\| \le \left[\frac{4\Delta}{\gamma_0} + \frac{8b_0}{\eta_0} + \frac{8\sqrt{2e^{3/4}}\,\gamma_0}{\sqrt{\eta_0}}\,(L_0 + 2L_1 G) + 8\sqrt{\eta_0}\, G\right] T^{-1/3} + \left(4\sqrt{2}\, L_0\gamma_0 + 8\sqrt{2}\, L_1 G\gamma_0\right) T^{-2/3}. \tag{54}$$

Consequently, $\mathbb{E}\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/3})$, and the SFO complexity is $\mathcal{O}(\varepsilon^{-3})$.

Proof. Since $B = 0$, we have $q_k = G$ for every k and

$$\frac{1}{T}\sum_{k=0}^{K} \mathbb{E}[q_k] = G.$$

If $L_1 > 0$, the conditions of Theorem 5(iii) hold because

$$\gamma = \gamma_0\, T^{-2/3} \le \gamma_0 \le \frac{1}{4\sqrt{2}\, L_1},$$

and

$$16\sqrt{2e^{3/4}}\, L_1 \frac{\gamma}{\eta} = 16\sqrt{2e^{3/4}}\, L_1 \frac{\gamma_0}{\eta_0} \le 1.$$

Apply (51) with $B = 0$. We have

$$\frac{4\Delta}{\gamma T} = \frac{4\Delta}{\gamma_0}\, T^{-1/3}, \qquad \frac{8b_0}{\eta T} = \frac{8b_0}{\eta_0}\, T^{-1/3}.$$

Since $G + BT\gamma = G$,

$$\frac{8\gamma}{\sqrt{\eta}}\sqrt{2e^{3/4}}\,(L_0 + 2L_1 G) = \frac{8\sqrt{2e^{3/4}}\,\gamma_0}{\sqrt{\eta_0}}(L_0 + 2L_1 G)\, T^{-1/3},$$

$$8\sqrt{\eta}\, G = 8\sqrt{\eta_0}\, G\, T^{-1/3},$$

$$4\sqrt{2}\, L_0\gamma = 4\sqrt{2}\, L_0\gamma_0\, T^{-2/3},$$

and

$$8\sqrt{2}\, L_1\gamma\, G = 8\sqrt{2}\, L_1 G\gamma_0\, T^{-2/3}.$$

Combining gives (54). The SFO cost is $1 + 2T = \mathcal{O}(\varepsilon^{-3})$. ■

C.8 Recovery of deterministic rates ($B = G = 0$)

When the oracle is deterministic ($B = G = 0$), the stochastic gradient noise vanishes. Therefore $v^k = \nabla f(x^k)$, $b_k = 0$, and $q_k = 0$ for all k. The NSTORM update reduces to normalized gradient descent, and the estimator recursion plays no role. The classical $\mathcal{O}(\varepsilon^{-2})$ rate is recovered for all three smoothness regimes.

Corollary 5.7 (Mean-square smoothness, deterministic)**.** *Suppose f is bounded below and 3 hold, and the oracle is deterministic ($B = G = 0$). Consider the normalized gradient descent update*

$$x^{k+1} = x^k - \gamma \frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}.$$

Choose $\gamma = \gamma_0 T^{-1/2}$. Then

$$\|\nabla f(\widehat{x})\| \le \frac{\Delta}{\gamma_0} T^{-1/2} + \frac{L\gamma_0}{2} T^{-1/2}. \tag{55}$$

Consequently, $\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/2})$, and the iteration complexity is $\mathcal{O}(\varepsilon^{-2})$.

Proof. Under determinism, $v^k = \nabla f(x^k)$ and $b_k = 0$ for all k. Since $d^k = \nabla f(x^k)/\|\nabla f(x^k)\|$ (when nonzero), we have $\langle \nabla f(x^k), d^k \rangle = \|\nabla f(x^k)\|$.

Assumption 3 implies L-smoothness via Jensen (Lemma 7). The one-step descent becomes

$$f(x^{k+1}) \le f(x^k) - \gamma \|\nabla f(x^k)\| + \frac{L}{2}\gamma^2.$$

Summing from $k = 0$ to K and using $f(x^{K+1}) \ge f^{\inf}$,

$$\gamma \sum_{k=0}^{K} \|\nabla f(x^k)\| \le \Delta + \frac{L}{2}\gamma^2 T.$$

Dividing by γT and substituting $\gamma = \gamma_0 T^{-1/2}$ gives (55). ■

Corollary 5.8 (Expected α-symmetric generalized smoothness, deterministic)**.** *Suppose f is bounded below and 5 hold with $\alpha \in (0, 1)$ and constants $K_0, K_1, K_2 > 0$, and the oracle is deterministic ($B = G = 0$). Define $L_\alpha := 2^{\alpha/2} K_1$, $c_\alpha := (1-\alpha)\alpha^{\alpha/(1-\alpha)}$, and $p := \alpha/(1-\alpha)$. Consider the normalized gradient descent update. Choose $\gamma = \gamma_0 T^{-1/2}$ and let $\lambda_0 > 0$ satisfy $L_\alpha \lambda_0 \gamma_0 \le 1$. Then*

$$\|\nabla f(\widehat{x})\| \le \frac{2\Delta}{\gamma_0} T^{-1/2} + \gamma_0 \left(K_0 + c_\alpha L_\alpha \lambda_0^{-p} + 2K_2 \gamma_0^p T^{-p/2} \right) T^{-1/2}. \tag{56}$$

Consequently, $\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/2})$, and the iteration complexity is $\mathcal{O}(\varepsilon^{-2})$.

Proof. Under determinism, $v^k = \nabla f(x^k)$, $b_k = 0$, and $q_k = 0$ for all k. The one-step descent from Lemma 9 with $b_k = 0$ and $\mathbb{E}[q_k^\alpha] = 0$ becomes

$$f(x^{k+1}) + \frac{\gamma}{2}\|\nabla f(x^k)\| \le f(x^k) + \frac{\gamma^2}{2}\left(K_0 + c_\alpha L_\alpha \lambda_0^{-p} + 2K_2 \gamma^p \right).$$

Summing from $k = 0$ to K, using $f(x^{K+1}) \ge f^{\inf}$, and dividing by $\gamma T/2$ gives

$$\frac{1}{T}\sum_{k=0}^{K} \|\nabla f(x^k)\| \le \frac{2\Delta}{\gamma T} + \gamma \left(K_0 + c_\alpha L_\alpha \lambda_0^{-p} + 2K_2 \gamma^p \right).$$

Substituting $\gamma = \gamma_0 T^{-1/2}$ gives

$$\frac{2\Delta}{\gamma T} = \frac{2\Delta}{\gamma_0} T^{-1/2},$$

and

$$\gamma \left(K_0 + c_\alpha L_\alpha \lambda_0^{-p} \right) = \gamma_0 \left(K_0 + c_\alpha L_\alpha \lambda_0^{-p} \right) T^{-1/2},$$

and

$$2K_2 \gamma^{p+1} = 2K_2 \gamma_0^{p+1} T^{-(p+1)/2}.$$

Since $p > 0$, the last term decays faster than $T^{-1/2}$. This proves (56) and the $\mathcal{O}(T^{-1/2})$ rate. ■

Corollary 5.9 (Expected 1-symmetric generalized smoothness, deterministic)**.** *Suppose f is bounded below and 5 hold with $\alpha = 1$ and constants $L_0, L_1 > 0$, and the oracle is deterministic ($B = G = 0$). Consider the normalized gradient descent update. Choose $\gamma = \gamma_0\, T^{-1/2}$ with $\gamma_0 \le 1/(4\sqrt{2}\, L_1)$ if $L_1 > 0$. Then*

$$\|\nabla f(\widehat{x})\| \le \frac{2\Delta}{\gamma_0}\, T^{-1/2} + 2\sqrt{2}\, L_0 \gamma_0\, T^{-1/2}. \tag{57}$$

Consequently, $\|\nabla f(\widehat{x})\| = \mathcal{O}(T^{-1/2})$, and the iteration complexity is $\mathcal{O}(\varepsilon^{-2})$.

Proof. Under determinism, $v^k = \nabla f(x^k)$, $b_k = 0$, and $q_k = 0$ for all k. The one-step descent from Lemma 10 becomes

$$f(x^{k+1}) + \frac{\gamma}{2}\|\nabla f(x^k)\| \le f(x^k) + \sqrt{2}\, L_0 \gamma^2.$$

Indeed, the $2\sqrt{2}\, L_1 \gamma^2 \|\nabla f(x^k)\|$ term is absorbed by $(\gamma/2)\|\nabla f(x^k)\|$ via the stepsize condition $2\sqrt{2}\, L_1 \gamma \le 1/2$, and the q_k-dependent term vanishes.

Summing from $k = 0$ to K, using $f(x^{K+1}) \ge f^{\inf}$, and dividing by $\gamma T/2$ gives

$$\frac{1}{T}\sum_{k=0}^{K} \|\nabla f(x^k)\| \le \frac{2\Delta}{\gamma T} + 2\sqrt{2}\, L_0 \gamma.$$

Substituting $\gamma = \gamma_0 T^{-1/2}$ proves (57). ∎

C.9 Discussion on recovered rates

The six recovery corollaries above confirm the claims made in the discussion following Theorem 2.

- *Bounded variance recovery.* When $B = 0$, Corollaries 5.4, 5.5, and 5.6 all achieve $\mathcal{O}(\varepsilon^{-3})$ SFO complexity with the bounded variance schedule $\eta = \eta_0 T^{-2/3}$, $\gamma = \gamma_0 T^{-2/3}$. This recovers the standard variance-reduced rate of Fang et al. [12], Cutkosky and Orabona [10], Chen et al. [7] under mean-square smoothness and expected generalized smoothness alike.
- *Deterministic recovery.* When $B = G = 0$, Corollaries 5.7, 5.8, and 5.9 all achieve $\mathcal{O}(\varepsilon^{-2})$ iteration complexity. Since NSTORM reduces to normalized gradient descent in the deterministic setting, these rates match the classical first-order complexity of Nesterov [34] and the generalized smooth rates of Chen et al. [7].
- *Schedule transition.* As the noise weakens from BG-0 to bounded variance to deterministic, the NSTORM schedules and rates transition as follows for the mean-square smoothness case:

  $$(\eta,\, \gamma) = \begin{cases} (T^{-1},\ \gamma_0 T^{-3/4}), & \text{BG-0 (sharp init.)}, \\ (\eta_0 T^{-2/3},\ \gamma_0 T^{-2/3}), & \text{bounded variance}, \\ (1,\ \gamma_0 T^{-1/2}), & \text{deterministic}, \end{cases}$$

  yielding rates $T^{-1/4}$, $T^{-1/3}$, and $T^{-1/2}$, respectively. The η parameter decays more slowly as the noise weakens, because the STORM recursion needs less aggressive averaging when the fresh-noise and trajectory-growth contributions are smaller.

  For the expected α-generalized smoothness case, the transition is:

  $$\text{SFO complexity} = \begin{cases} \mathcal{O}(\varepsilon^{-(4+\alpha)}), & \text{BG-0}, \\ \mathcal{O}(\varepsilon^{-3}), & \text{bounded variance}, \\ \mathcal{O}(\varepsilon^{-2}), & \text{deterministic}. \end{cases}$$

  Thus the α-dependent price $\mathcal{O}(\varepsilon^{-(4+\alpha)})$ under BG-0 is entirely an artifact of the interaction between distance-dependent variance and gradient-dependent curvature; it disappears when either source of difficulty is removed.

D Experimental Details

D.1 BG-0 Wrapper

Given a fixed deterministic objective f, the BG-0 wrapper turns the setup into a stochastic problem in which the stochastic gradients satisfy the BG-0 variance model exactly.

Given the independent random variables ρ and u satisfying

$$\mathbb{E}[\rho] = 0, \qquad \mathbb{E}[\rho^2] = 1, \qquad \mathbb{E}[u] = 0, \qquad \mathbb{E}\|u\|^2 = 1.$$

where $\rho \in \{-1, +1\}$ uniformly and $u \sim \mathcal{N}(0, I_d/d)$. For each sample $\xi = (\rho, u)$, the stochastic loss is given by

$$f_\xi(x) = f(x) + \frac{B\rho}{2}\|x - x^0\|^2 + G\langle u, x - x^0\rangle,$$

where the corresponding stochastic gradient is

$$\nabla f_\xi(x) = \nabla f(x) + B\rho(x - x^0) + Gu.$$

Here, the objective is unchanged and the oracle is unbiased as

$$\mathbb{E}_\xi[f_\xi(x)] = f(x), \qquad \mathbb{E}_\xi[\nabla f_\xi(x)] = \nabla f(x).$$

Moreover,

$$\mathbb{E}_\xi\|\nabla f_\xi(x) - \nabla f(x)\|^2 = \mathbb{E}_\xi\|B\rho(x - x^0) + Gu\|^2 = B^2\|x - x^0\|^2 + G^2,$$

where the cross term vanishes by independence and the zero-mean assumptions. Thus, our wrapper satisfies the exact BG-0 oracle.

As the extension to the mini-batch case of size b (which is essential to compare with the dynamic batching based models in [13]), we draw independent $\xi_i = (\rho_i, u_i)$ and use

$$g_b(x) = \frac{1}{b}\sum_{i=1}^{b} \nabla f_{\xi_i}(x) = \nabla f(x) + \frac{1}{b}\sum_{i=1}^{b} \left(B\rho_i(x - x^0) + Gu_i\right),$$

which satisfies

$$\mathbb{E}[g_b(x)] = \nabla f(x), \qquad \mathbb{E}\|g_b(x) - \nabla f(x)\|^2 = \frac{B^2\|x - x^0\|^2 + G^2}{b}.$$

For the STORM and NSTORM recursions, when a transition from x to y requires two stochastic gradients, we evaluate $\nabla f_\xi(x)$ and $\nabla f_\xi(y)$ with the same sample ξ and analogously reuse the same mini-batch in the batched case. Since this process requires the computation of two stochastic gradients, the corresponding SFO is doubled in our calculations.

Lemma 14. *Let $f \in \mathcal{L}^*_{\mathrm{sym}}(\alpha)$ with constants $L_0, L_1 > 0$, and let $\alpha \in (0, 1]$. Let $\rho \in \{-1, +1\}$ uniformly and $u \sim \mathcal{N}(0, I_d/d)$, independently. Thus*

$$\mathbb{E}[\rho] = 0, \qquad \mathbb{E}[\rho^2] = 1, \qquad \mathbb{E}[u] = 0, \qquad \mathbb{E}\|u\|^2 = 1.$$

For each $\xi = (\rho, u)$, define

$$f_\xi(x) = f(x) + \frac{B\rho}{2}\|x - x^0\|^2 + G\langle u, x - x^0\rangle.$$

Then $\mathbb{E}_\xi[f_\xi(x)] = f(x)$, the stochastic gradient is unbiased, and

$$\mathbb{E}_\xi\|\nabla f_\xi(x) - \nabla f(x)\|^2 = B^2\|x - x^0\|^2 + G^2.$$

Moreover, there exist constants $\widetilde{L}_0, \widetilde{L}_1 > 0$ such that, for all x, y,

$$\mathbb{E}_\xi\|\nabla f_\xi(y) - \nabla f_\xi(x)\|^2 \le \|y - x\|^2\mathbb{E}_\xi\left[\left(\widetilde{L}_0 + \widetilde{L}_1 \max_{\theta\in[0,1]} \|\nabla f_\xi(x_\theta)\|^\alpha\right)^2\right],$$

where $x_\theta = \theta y + (1 - \theta)x$.

Proof. Since $\mathbb{E}[\rho] = 0$ and $\mathbb{E}[u] = 0$ we have $\mathbb{E}_\xi[f_\xi(x)] = f(x)$ and $\mathbb{E}_\xi[\nabla f_\xi(x)] = \nabla f(x)$. Moreover,

$$\nabla f_\xi(x) - \nabla f(x) = B\rho(x - x^0) + Gu.$$

Using independence and the properties of ρ and u, we obtain

$$\mathbb{E}_\xi\|\nabla f_\xi(x) - \nabla f(x)\|^2 = B^2\|x - x^0\|^2 + G^2.$$

It remains to prove the expected generalized smoothness bound. For any x, y, let $d = y - x$. Since

$$\nabla f_\xi(y) - \nabla f_\xi(x) = \nabla f(y) - \nabla f(x) + B\rho d,$$

we have

$$\mathbb{E}_\xi\|\nabla f_\xi(y) - \nabla f_\xi(x)\|^2 = \|\nabla f(y) - \nabla f(x)\|^2 + B^2\|d\|^2.$$

Since $f \in \mathcal{L}^*_{\mathrm{sym}}(\alpha)$,

$$\|\nabla f(y) - \nabla f(x)\|^2 \le \|d\|^2 \left(L_0 + L_1 \max_{\theta\in[0,1]} \|\nabla f(x_\theta)\|^\alpha\right)^2.$$

Thus, with

$$R := \max_{\theta\in[0,1]} \|\nabla f(x_\theta)\|,$$

we obtain

$$\mathbb{E}_\xi\|\nabla f_\xi(y) - \nabla f_\xi(x)\|^2 \le \|d\|^2 \left[2L_0^2 + 2L_1^2 R^{2\alpha} + B^2\right].$$

Now define

$$M_\xi := \max_{\theta\in[0,1]} \|\nabla f_\xi(x_\theta)\|.$$

For any fixed θ, write

$$Z_\theta := B\rho(x_\theta - x^0) + Gu.$$

The random vector Z_θ is symmetric: Z_θ and $-Z_\theta$ have the same distribution. Hence, for any fixed vector a and any $p > 0$,

$$\mathbb{E}_\xi\|a + Z_\theta\|^p = \frac{1}{2}\mathbb{E}_\xi\left[\|a + Z_\theta\|^p + \|a - Z_\theta\|^p\right] \ge \frac{1}{2}\|a\|^p,$$

because at least one of $\|a + Z_\theta\|$ and $\|a - Z_\theta\|$ is at least $\|a\|$. Taking $a = \nabla f(x_\theta)$, $p = 2\alpha$, and using $M_\xi \ge \|\nabla f_\xi(x_\theta)\|$, we get

$$\mathbb{E}_\xi[M_\xi^{2\alpha}] \ge \frac{1}{2}\|\nabla f(x_\theta)\|^{2\alpha} \quad \text{for all } \theta \in [0,1].$$

Therefore,

$$R^{2\alpha} \le 2\mathbb{E}_\xi[M_\xi^{2\alpha}].$$

Choose constants $\widetilde{L}_0, \widetilde{L}_1 > 0$ such that

$$\widetilde{L}_0^2 \ge 2L_0^2 + B^2, \qquad \widetilde{L}_1^2 \ge 4L_1^2.$$

Then

$$2L_0^2 + B^2 + 2L_1^2 R^{2\alpha} \le \widetilde{L}_0^2 + \widetilde{L}_1^2\mathbb{E}_\xi[M_\xi^{2\alpha}] \le \mathbb{E}_\xi\left[\left(\widetilde{L}_0 + \widetilde{L}_1 M_\xi^\alpha\right)^2\right].$$

Combining the preceding inequalities proves the claim. ■

This wrapper lets us evaluate stochastic methods on deterministic test objectives while imposing an unbounded variance that grows with the distance from initialization. It therefore provides a controlled experimental model for the distance-dependent stochastic noise studied in the paper, without changing the underlying deterministic objective. Lemma 14 verifies that for all $\alpha \in (0, 1]$ (including the $\alpha = 1/2$ and $\alpha = 2/3$ experimental setups) the proposed framework is compatible with the expected symmetric generalized smoothness framework.

D.2 Phase Retrieval

The first experiment is the phase retrieval objective from Chen et al. [7]. Given measurements $y_r = |a_r^\top x_\star|^2$, we minimize

$$f(x) = \frac{1}{2m} \sum_{r=1}^{m} \left(y_r - |a_r^\top x|^2\right)^2.$$

This nonconvex objective is not globally L-smooth, but it belongs to the α-symmetric generalized smooth class $\mathcal{L}^*_{\mathrm{sym}}(2/3)$, and its stochastic finite-sum version satisfies the corresponding expected condition [7]. In our experiments, we use dimension $d = 100$ and $m = 3000$ measurements. The measurement matrix has independent entries $a_{r,j} \sim \mathcal{N}(0, 0.01)$, and the target signal is drawn as $x_{\star,j} \sim \mathcal{N}(0, 1)$. Hence, $y_r = |a_r^\top x_\star|^2$. The optimization is initialized independently from $x_j^0 \sim \mathcal{N}(5, 1)$, which places the initial point away from the target signal and makes the distance-dependent variance growth visible. After constructing this deterministic phase retrieval objective, we apply the BG-0 wrapper from Appendix D.1. Thus, the stochasticity is controlled through the BG-0 oracle.

D.3 Cubic Polynomial

For the second experiment, we use the following polynomial function as the deterministic objective:

$$f(x) = |x|^{(2-\alpha)/(1-\alpha)}$$

which has been shown in [7] to satisfy $f \in \mathcal{L}^*_{\mathrm{sym}}(\alpha)$ for $x \in \mathbb{R}$ and $\alpha \in (0, 1)$. We instantiate this example with $\alpha = 1/2$, so $f(x) = |x|^3$, and use the same distance-dependent BG-0 oracle. This gives a minimal one-dimensional experiment in which both the generalized smoothness exponent and the BG-0 variance growth are explicit.

D.4 Additional Optimizer Comparison

To further investigate the role of normalization and momentum under BG-0 noise, we compare the proposed methods with unnormalized SGDM, Adam, SIGNUM, and SignSGD, in addition to the dynamic-batching baselines. We consider the same cubic-polynomial and phase-retrieval problems, with dimensions $d = 1$ and $d = 100$, respectively. Table 2 summarizes the final gradient norms reported over three random seeds.

In one dimension, normalized momentum reduces to SIGNUM when the momentum parameters and stochastic samples are matched, explaining their identical results on the cubic-polynomial problem. In higher dimensions, NSGDM uses an ℓ_2-normalized momentum vector, whereas SIGNUM applies the sign operation coordinatewise. The comparisons also illustrate the benefit of temporal averaging: single-sample SignSGD performs worse than momentum-based normalized updates in both experiments. Dynamic batching can further reduce the final gradient norm, but requires additional oracle evaluations.

E Hyperparameters

We report the hyperparameters used for the normalized methods in Figures 1 and 2. We denote $T := K + 1$, where K is the total number of iterations for a given method. The schedules below specify the actual choices for the plotted NSGDM and NSTORM curves. For fair comparison, we apply grid search to set the learning rate of SGD (b=1), SGD (dynamic batch), and STORM (dynamic batch).

Normalized SGDM. For single sample NSGDM, we use the schedule from Theorem 1:

$$\eta = T^{-2/3}, \qquad \gamma = \gamma_0 T^{-5/6},$$

which is shared by the standard smooth, $\mathcal{L}^*_{\mathrm{sym}}(\alpha)$, and $\mathcal{L}^*_{\mathrm{sym}}(1)$ cases in Theorem 1.

Table 2: Comparison of final gradient norms $\|\nabla f(x^K)\|$ under BG-0 noise. Results are reported as mean $\pm$ standard deviation over three random seeds.

Method	Cubic ($d = 1$)	Phase retrieval ($d = 100$)
SGD ($b = 1$)	3.317 ± 0.099	1.245 ± 0.239
SGDM ($b = 1$)	1.204 ± 0.018	0.817 ± 0.317
SIGNUM ($b = 1$)	0.473 ± 0.272	0.546 ± 0.552
SignSGD ($b = 1$)	3.409 ± 0.833	8.989 ± 7.845
Adam ($b = 1$)	0.693 ± 0.076	0.520 ± 0.190
NSGDM ($b = 1$)	0.473 ± 0.272	0.393 ± 0.481
NSTORM ($b = 1$)	0.102 ± 0.160	0.256 ± 0.073
SGD (dynamic batch)	3.337 ± 0.062	0.408 ± 0.027
SGDM (dynamic batch)	0.191 ± 0.013	0.220 ± 0.006
Adam (dynamic batch)	0.137 ± 0.015	0.154 ± 0.029
STORM (dynamic batch)	0.031 ± 0.023	0.127 ± 0.069

Normalized STORM. For NSTORM, the initialization batch and schedule depend on the smoothness regime in Theorem 2. For the expected α-symmetric generalized smoothness experiments, we retain the horizon-dependent scaling of the parameter schedules, with experiment-specific initialization constants

$$N_{\text{init}} \propto \left\lceil T^{\frac{2(1-\alpha)}{4+\alpha}} \right\rceil, \qquad \gamma = \gamma_0 T^{-\frac{3+\alpha}{4+\alpha}}, \qquad \eta = \eta_0 T^{-\frac{4}{4+\alpha}}.$$

The specific initialization batch sizes and stepsize coefficients used in our experiments are reported below.

Phase retrieval instantiation. For Figure 1, the phase retrieval objective satisfies the α-symmetric generalized smoothness condition with $\alpha = 2/3$. We use $T = 10001$, $B = 1$, $G = 1$, $\gamma_0 = 10$ for NSGDM, and $\gamma_0 = 7.5$, $\eta_0 = 1$, $N_{\text{init}} = 4$ for NSTORM. Thus, the corresponding hyperparameters are set to

$$\text{NSGDM}: \qquad \gamma \approx 4.641 \times 10^{-3}, \qquad \eta \approx 2.154 \times 10^{-3},$$

and

$$\text{NSTORM}: \qquad \gamma \approx 5.397 \times 10^{-3}, \qquad \eta \approx 3.73 \times 10^{-4}.$$

Cubic polynomial instantiation. For Figure 2, the polynomial experiment is set to satisfy the α-symmetric generalized smoothness condition with $\alpha = 1/2$, $T = 10001$, $B = 1.5$, $G = 0.5$, and $\gamma_0 = \eta_0 = 1$, $N_{\text{init}} = 2$. Thus, the corresponding hyperparameters are set to

$$\text{NSGDM}: \qquad \gamma \approx 4.641 \times 10^{-4}, \qquad \eta \approx 2.154 \times 10^{-3}.$$

and

$$\text{NSTORM}: \qquad \gamma \approx 7.742 \times 10^{-4}, \qquad \eta \approx 2.782 \times 10^{-4}.$$

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction (Section 1, Table 1) state exact complexity for each method, smoothness, and noise combination. All claims are supported by formal theorems (Theorems 1 and 2) with complete proofs in the appendix. Assumptions are listed explicitly in Section 3. The scope of the experimental evaluation is discussed in Section 6.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 7 discusses the main open question, whether the α-dependent price is information-theoretically tight or an artifact of the proof technique. We think it is the former than the latter.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All assumptions (Assumptions BG-0–5) are stated in Section 3 and referenced in each theorem statement. Complete proofs of all theoretical results are provided in Appendices B and C, including recovery corollaries for the bounded-variance and deterministic special cases.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The algorithms are fully specified in Equations (2) and (3). The BG-0 wrapper is described in Appendix D.1. All hyperparameter schedules are derived from the theorems and reported in Appendix E with exact numerical values. The testbed objectives (phase retrieval and polynomial separation) are standard and fully specified.

Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: We have provided the code as a zip file in the supplementary materials.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 6 describes the experimental setup. Appendix E provides exact hyperparameter values, initialization strategies, horizon lengths, and noise parameters for both testbeds.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: All of the experimental results, illustrated in Figure 1 and 2, are averaged over three random seeds and reported as mean $\pm$ standard deviation.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: All experiments were run on CPU using a single AMD EPYC 7502P 32-Core Processor machine. The experiments are lightweight, the supplementary material includes the software environment needed to reproduce the manuscript figures.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work conforms with the code of ethics, and anonymity is fully preserved.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [N/A]

Justification: The contributions are purely theoretical, and we do not believe the work to have any immediate societal impacts.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: The paper poses no such risks.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [N/A]

Justification: The paper does not use existing assets. We have properly cited all sources.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: Our source code and data are publicly available as supplementary material.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

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