When and Why SignSGD Outperforms SGD: A Theoretical Study Based on ℓ_1-norm Lower Bounds

Hongyi Tao* **Dingzhi Yu*** **Lijun Zhang†**
State Key Laboratory of Novel Software Technology,
Nanjing University, Nanjing 210023, China
School of Artificial Intelligence, Nanjing University, Nanjing 210023, China
{taohy,yudz,zhanglj}@lamda.nju.edu.cn

Abstract

Sign-based optimization algorithms, such as SignSGD and Muon, have garnered significant attention for their remarkable performance in training large foundation models. Despite this empirical success, we still lack a theoretical understanding of when and why these sign-based methods outperform vanilla SGD. The core obstacle is that under standard smoothness and finite variance conditions, SGD is known to be minimax optimal for finding stationary points measured by ℓ_2-norms, thereby fundamentally precluding any complexity gains for sign-based methods in standard settings. To overcome this barrier, we analyze sign-based optimizers leveraging ℓ_1-norm stationarity, ℓ_∞-smoothness, and a separable noise model, which can better capture the coordinate-wise nature of signed updates. Under this distinct problem geometry, we derive matched upper and lower bounds for SignSGD and explicitly characterize the problem class in which SignSGD provably dominates SGD. Specifically, we compare the *upper bound of SignSGD* with the *lower bound of SGD*, illustrating that SignSGD effectively reduces the complexity by a factor of d under *sparse noise*, where d is the problem dimension. Furthermore, we elevate this framework to the matrix domain, providing an equivalent optimal lower bound for the Muon optimizer, proving that extending the sign operator to matrices preserves this optimal scaling with dimensionality. Finally, we bridge our theoretical bounds to practice, demonstrating that the theoretical superiority of SignSGD accurately predicts its faster convergence during the pretraining of nanoGPT and LLaMA, with Muon also outperforming SGD in these experiments. Code is available at https://github.com/Dingzhen230/SignSGD_Outperforms_SGD.

1 Introduction

Efficient and scalable stochastic optimization algorithms play an indispensable role in the huge success of large foundation models [Devlin et al., 2019, Brown et al., 2020, Achiam et al., 2023, Touvron et al., 2023, Team et al., 2023]. Among these, sign-based optimization algorithms, such as SignSGD [Bernstein et al., 2018] and Muon [Jordan et al., 2024], have attracted increasing focus due to substantial empirical edges over stochastic gradient descent (SGD) [Robbins and Monro, 1951]. SignSGD leverages 1-bit signed gradients to update the model, which naturally enjoys many favorable empirical properties in distributed environments [Bernstein et al., 2019] and low-precision training regimes [Yu et al., 2026c]. Muon optimizes using the matrix sign operator, implemented via Newton–Schulz iterations [Kovarik, 1970, Björck and Bowie, 1971]. Numerous empirical studies [Liu et al., 2025, Shah et al., 2025, Wen et al., 2025, Semenov et al., 2025] have shown its consistent speedup over AdamW [Loshchilov and Hutter, 2019], and it has become a new industrial paradigm for pretraining massive-scale large language models (LLMs) [Team et al., 2025, 2026, Zeng et al., 2025, 2026, Cheng et al., 2026, DeepSeek-AI, 2026].

*Equal Contribution. † Corresponding Author.

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

The empirical success of sign-based optimizers has motivated a fruitful line of work that attempts to justify their advantages theoretically. From an optimization theory perspective, the convergence of sign-based methods under various assumptions has been extensively studied [Bernstein et al., 2018, 2019, Safaryan and Richtárik, 2021, Sun et al., 2023, Jiang et al., 2025b, Yu et al., 2026e]. However, a complete theoretical understanding of why sign-based optimizers outperform SGD remains elusive. The biggest obstacle preventing any theoretical advances is that *SGD is already worst-case optimal when the objective function is smooth and stochastic gradients are unbiased with bounded variance [Arjevani et al., 2023]*. Under these standard assumptions, any first-order algorithm requires at least $\Omega(\epsilon^{-4})$ queries to find an ϵ-stationary point measured by the ℓ_2-norm, a complexity tightly matched by SGD [Ghadimi and Lan, 2013]. Consequently, evaluating the performance of sign-based methods under the aforementioned problem class and ℓ_2-stationary measure inherently fails to deliver any provable gains we desire.

In this paper, we systematically investigate when and why sign-based methods can outperform SGD by delving into a new problem class and performance measure that depart from the canonical setting in Arjevani et al. [2023]. Specifically, instead of the traditional ℓ_2-geometry based on ℓ_2-smoothness, finite variance, and ℓ_2-stationarity, we consider an ℓ_∞-geometry with ℓ_∞-smoothness, coordinate-wise finite variance, and ℓ_1-stationarity. This geometry shift is motivated by the insight that sign-based updates are intrinsically aligned with ℓ_∞-geometry [Balles et al., 2020, Bernstein and Newhouse, 2024, Xie et al., 2025a,b]. Under this alternative geometry, we provide the first rigorous characterization of a problem class in which SignSGD provably outperforms SGD.

To clarify this geometric shift, it is essential to highlight the distinction between these conditions. Standard ℓ_2-smoothness assumes the loss landscape curves uniformly across all directions, while ℓ_∞-smoothness bounds the gradient change based on the maximum coordinate-wise distance, providing a hypercube-based geometric assumption that aligns with coordinate-wise, scaled updates of sign-based methods. Define $D_f(\mathbf{x}, \mathbf{y}) := f(\mathbf{y}) - (f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle)$ as the Bregman Divergence of f, the above two smooth models can be expressed as below.

ℓ_2-smoothness

$$|D_f(\mathbf{x}, \mathbf{y})| \leq \frac{L_2}{2} \|\mathbf{y} - \mathbf{x}\|_2^2$$

ℓ_∞-smoothness

$$|D_f(\mathbf{x}, \mathbf{y})| \leq \frac{L_\infty}{2} \|\mathbf{y} - \mathbf{x}\|_\infty^2$$

Furthermore, we adopt a separable noise model where variance is tracked independently for each coordinate via a vector $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \ldots, \sigma_d]$. This fine-grained characterization is more suitable for analyzing the highly imbalanced noise prevalent in modern deep learning [Sagun et al., 2017, Zhu et al., 2019, Zhang et al., 2020, Pan et al., 2022, Wu et al., 2022, Pan et al., 2024].

Standard noise model

$$\mathbb{E}\left[\|\mathbf{g}_t^b - \nabla f(\mathbf{x}_t)\|_2^2 \middle| \mathcal{F}_{t-1} \right] \leq \sigma^2$$

Separable noise model, $\forall i \in [d]$

$$\mathbb{E}\left[\left|\mathbf{g}_{t,i}^b - \nabla_i f(\mathbf{x}_t)\right|^2 \middle| \mathcal{F}_{t-1} \right] \leq \sigma_i^2$$

Under this refined geometry, we develop a clean, self-contained lower bound analysis for SignSGD. We prove an ℓ_1-norm lower bound that matches its upper bound, thereby giving a tight characterization of its convergence rate. We further derive an ℓ_1-norm lower bound for SGD under the same setting. Comparing the *upper bound of SignSGD* with the *lower bound of SGD* reveals a strict separation: when the coordinate-wise noise is sparse or highly heterogeneous, SignSGD achieves a better dimension dependence. Finally, we extend this framework to matrix optimization and establish the first lower bound for Muon measured by the nuclear norm with matching upper bounds.

The main contributions of this paper are summarized as follows.

- **Tight Bounds for SignSGD.** Let L_∞ and $\boldsymbol{\sigma}$ denote the ℓ_∞-smooth Lipschitz constant and the noise standard-deviation bound vector, respectively. We establish that the convergence rate of SignSGD with a constant step size is: $\mathbb{E}\left[\min_t \|\nabla f(\mathbf{x}_t)\|_1\right] = \Theta\left(\sqrt{L_\infty \Delta / N} + \left(\|\boldsymbol{\sigma}\|_1^2 L_\infty \Delta / N\right)^{\frac{1}{4}}\right)$.
- **Provable Dimensional Gains of SignSGD over SGD.** By comparing SignSGD's *upper bound* with SGD's *lower bound*, we prove that SignSGD achieves a strictly superior dimension dependence. We further show that when the noise $\boldsymbol{\sigma}$ is sparse, SignSGD's complexity can be d times better than that of SGD. We validate these theoretical findings by demonstrating accelerated convergence during the pretraining of nanoGPT and LLaMA.

- **Unified Extension to Muon.** We elevate our theoretical framework into the matrix domain, mapping separable operations to orthogonalized matrix steps. By applying *anisotropic matrix variance* (Assumption 2b) and *spectral norm smoothness* (Assumption 3b), we establish matching upper and lower bounds for the recently proposed Muon optimizer [Jordan et al., 2024], proving for the first time that Muon handles large-scale matrix parameters just as effectively as SignSGD handles flat vectors.

2 Related Work

Sign-based methods: vector optimizers The idea of using only gradient signs dates back at least to RProp [Riedmiller and Braun, 1993], while the modern stochastic formulation was introduced by Bernstein et al. [2018], who studied SignSGD and its momentum variant Signum for smooth non-convex optimization. Since SignSGD replaces the stochastic gradient by its coordinate-wise sign, its natural stationarity measure is ℓ_1-norm rather than the usual ℓ_2-norm. This geometry is also closely tied to communication efficiency, as one-bit updates and majority vote make SignSGD attractive in distributed learning [Bernstein et al., 2019, Jin et al., 2020, Safaryan and Richtárik, 2021].

A fundamental limitation of vanilla sign updates appears in non-smooth optimization. Even for simple convex non-smooth objectives, SignSGD can fail to converge because the discontinuity of the sign map may repeatedly select an unfavorable subgradient direction [Karimireddy et al., 2019, Xiao et al., 2023, Yu et al., 2026b]. Error feedback repairs this issue by accumulating and correcting compression errors [Karimireddy et al., 2019]. More recently, StoSignSGD [Yu et al., 2026c] injects unbiased structural stochasticity directly into the sign conversion, using coordinate-wise max-buffer to control the stochastic sign scale, and coupling its noise level with the gradient signal. It preserves the numerical simplicity of sign updates while resolving the non-smooth non-convergence pathology of deterministic SignSGD illustrated in Karimireddy et al. [2019], Yu et al. [2026b].

In smooth stochastic optimization, a separate line of work sharpens the convergence theory of sign-based methods through momentum-based analyses [Sun et al., 2023, Jiang et al., 2025b], stochastic sign mechanisms [Jin et al., 2020, Safaryan and Richtárik, 2021], variance reduction [Chzhen and Schechtman, 2023, Jiang et al., 2024], and heavy-tailed noise analyses [Kornilov et al., 2025, Yu et al., 2026e]. Sign-based vector optimizers are also closely related to adaptive methods. Several works interpret the optimization dynamics of Adam [Kingma and Ba, 2015] through signed or normalized update directions [Balles and Hennig, 2018, Crawshaw et al., 2022, Kunstner et al., 2023, Peng et al., 2025], while Lion [Chen et al., 2023] has emerged as a representative momentum-based sign optimizer with strong empirical performance and growing theoretical support [Chen et al., 2024, Dong et al., 2024, Jiang et al., 2026, Sfyraki and Wang, 2026, Yu et al., 2026e]. Yu et al. [2026e] demonstrate that SignSGD- and Lion-type methods can provably outperform AdamW, especially in LLM training regimes where the noise is heavy-tailed [Zhang et al., 2020, Yu et al., 2026d].

Sign-based methods: matrix optimizers Matrix sign optimizers extend the sign-based philosophy from coordinate-wise vector updates to structured matrix updates. The representative example is Muon [Jordan et al., 2024], which applies an orthogonalized matrix sign direction: for a matrix gradient $G = U\mathbf{\Sigma}V^\top$, the update direction is $\operatorname{msign}(G) = UV^\top$. In practice, this matrix sign operation is efficiently approximated by Newton–Schulz iterations [Kovarik, 1970, Björck and Bowie, 1971]. Subsequent variants, such as MuonLight, incorporate Nesterov momentum, learning-rate alignment, and implementation refinements, and have been used or benchmarked in massive-scale language model training [Liu et al., 2025, Zeng et al., 2025, Team et al., 2025, Wen et al., 2025, Semenov et al., 2025, DeepSeek-AI, 2026]. These works suggest that matrix-level normalization can provide optimization benefits beyond coordinate-wise sign updates.

Theoretical understanding of Muon and related matrix optimizers is still developing. Existing studies analyze Muon from several complementary perspectives, including spectral and nuclear-norm geometry, normalized matrix descent, preconditioning, connections to Adam-type methods, and empirical justifications for its superiority [Li and Hong, 2025, Shen et al., 2025, Chang et al., 2025, Si et al., 2025, Sfyraki and Wang, 2026, Chen et al., 2025, Huang et al., 2025, Li et al., 2025b, Qian et al., 2025, Tveit et al., 2025, Page et al., 2025, Mehta et al., 2025, Vasudeva et al., 2025a,b, Frans et al., 2025, Pan et al., 2025, Wang et al., 2025, Zhang and Gao, 2025, Zhang et al., 2025, Su, 2025, Crawshaw et al., 2025, Ma et al., 2026, Du and Su, 2026, Yang et al., 2026].

Lower complexity bounds Complexity lower bounds are the standard tool for determining whether algorithmic improvement is genuine or merely an artifact of analysis. The oracle-complexity view-

point goes back to the classical framework of Nemirovski and Yudin [1983], and was later developed extensively for stochastic convex optimization [Agarwal et al., 2012]. These results identify how noise, dimension, and geometry constrain the best possible rates in stochastic first-order optimization.

For smooth non-convex optimization, the classical target is an ϵ-stationary point measured by $\|\nabla f(x)\|_2$. In the deterministic setting, Carmon et al. [2020, 2021] established sharp lower bounds for finding stationary points of smooth high-dimensional functions, while Chewi et al. [2023] showed that even in one dimension the query complexity depends delicately on whether the algorithm is deterministic or randomized and whether it can access first-order or zeroth-plus-first-order information. In the stochastic setting, Ghadimi and Lan [2013] gave the standard $O(\epsilon^{-4})$ upper bound for SGD under unbiased bounded-variance gradients, and subsequent lower-bound work showed that this rate is essentially unavoidable. In particular, Drori and Shamir [2020] studied worst-case lower bounds for SGD itself, and Arjevani et al. [2023] proved that any stochastic first-order method requires $\Omega(\epsilon^{-4})$ stochastic-gradient queries under the standard bounded-variance model, with a corresponding $\Omega(\epsilon^{-3})$ lower bound under mean-squared smoothness. Thus, in the usual ℓ_2-geometry, SGD is minimax optimal, and one should not expect a worst-case improvement without changing the geometry, the oracle model, or the stationarity criterion.

A more recent line of work asks whether adaptive or sign-based methods can provably improve over SGD once the problem structure is refined. Jiang et al. [2025a] demonstrates the possibility for AdaGrad under coordinate-wise smoothness and coordinate-wise noise assumptions: by adopting an ℓ_1-stationarity measure, they prove both an AdaGrad upper bound and an SGD lower bound, yielding regimes where AdaGrad improves over SGD by a factor of the dimension. Crawshaw and Liu [2025] study lower bounds for adaptive gradient algorithms under (L_0, L_1)-smoothness, showing several AdaGrad variants necessarily incur higher-order dependence on relaxed-smoothness parameters.

3 SignSGD: Upper Bound and Lower Bound

In this section, we give the first tight characterization of the convergence of SignSGD.

3.1 Notations and Assumptions

We write $[T]$ for $\{1, 2, \ldots, T\}$, $|\mathbf{x}|$ for element-wise absolute value of $\mathbf{x} \in \mathbb{R}^d$. The $\mathbf{L}$-weighted vector norm for $\mathbf{L} \in \mathbb{R}^d_+$ is defined as $\|\mathbf{x}\|^2_{\mathbf{L}} := \mathbf{x}^\top \operatorname{diag}(\mathbf{L})\mathbf{x}$. For vector descent methods, we study the optimization problem $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$, where $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable. Given a point $\mathbf{x} \in \mathbb{R}^d$, we can only access the gradient $\nabla f(\mathbf{x}) = [\nabla_1 f(\mathbf{x}), \cdots, \nabla_d f(\mathbf{x})] \in \mathbb{R}^d$ in a noisy manner which we will later define in Assumption 2a. We use $\operatorname{sign}(\cdot)$ to denote the sign operator, and WLOG, use the convention $\operatorname{sign}(0) = 0$. Below, we list some necessary assumptions.

Assumption 1a (Lower bounded objective)**.** The function f is bounded from below. There exists $f^* > -\infty$ such that $f(\mathbf{x}) \geq f^*$, for all $\mathbf{x} \in \mathbb{R}^d$. We further denote $\Delta = f(\mathbf{x}_0) - \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$.

Assumption 2a (Separable noise model)**.** At step t we observe a mini-batch of gradients $g_t = \{\mathbf{g}_t^1, \cdots, \mathbf{g}_t^B\}$, mutually independent conditioned on $\mathcal{F}_{t-1}$ and satisfying $\mathbb{E}\left[\mathbf{g}_t^b | \mathcal{F}_{t-1}\right] = \nabla f(\mathbf{x}_t), \forall b \in [B]$ where $\mathcal{F}_t = \sigma(g_0, \ldots, g_t)$ denotes the natural filtration, with $\mathcal{F}_{-1}$ trivial. Moreover, the coordinate-wise conditional variance is bounded separably: there exist non-negative constants $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \cdots, \sigma_d]$ such that

$$\mathbb{E}\left[\left|\mathbf{g}_{t,i}^b - \nabla_i f(\mathbf{x}_t)\right|^2 \middle| \mathcal{F}_{t-1}\right] \leq \sigma_i^2, \quad \forall i \in [d], \forall b \in [B].$$

Assumption 1a is standard and necessary for stochastic non-convex optimization [Arjevani et al., 2023]. Assumption 2a is widely used in analysis of adaptive methods like AdaGrad [Duchi et al., 2011], and sign-based methods like SignSGD [Bernstein et al., 2018, Li et al., 2025a, Jiang et al., 2025a, Yu et al., 2026e].

Assumption 3a (ℓ_∞-smoothness)**.** The objective function f is ℓ_∞-smooth if there exists non-negative constant L_∞ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle)| \leq \frac{L_\infty}{2}\|\mathbf{y} - \mathbf{x}\|^2_\infty.$$

Besides Assumption 4a, separable smoothness (Assumption 4a) is also widely used in analyses of SignSGD methods [Bernstein et al., 2018, 2019, Safaryan and Richtárik, 2021, Yu et al., 2026e], which naturally aligns with sign descent methods due to their separable nature.

Algorithm 1 SignSGD [Bernstein et al., 2018]

Require iteration number T , initial point $\mathbf{x}_0$, batch size B, learning rate η
for step $t = 0$ **to** $T - 1$ **do**
  Compute stochastic gradient: $\mathbf{g}_t = {}^1\!/_B \sum_{b=1}^{B} \mathbf{g}_t^b$
  $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \operatorname{sign}(\mathbf{g}_t)$
end for

Algorithm 2 Muon [Jordan et al., 2024]

Require iteration number T , initial point $\mathbf{W}_0$, batch size B, learning rate η.
for step $t = 0$ **to** $T - 1$ **do**
  Compute stochastic gradient: $\mathbf{G}_t = {}^1\!/_B \sum_{b=1}^{B} \mathbf{G}_t^b$
  $\mathbf{W}_{t+1} = \mathbf{W}_t - \eta \operatorname{msign}(\mathbf{G}_t)$
end for

Assumption 4a (Separable smoothness)**.** The objective function f is $\mathbf{L}$-separable smooth if there exists a non-negative vector $\mathbf{L} = [L_1, L_2, \ldots, L_d]$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle)| \leq \frac{1}{2} \|\mathbf{y} - \mathbf{x}\|_{\mathbf{L}}^2.$$

Balles et al. [2020] further discussed the relationship between Assumptions 3a and 4a, pointing out that separable smoothness is a sufficient condition for ℓ_∞-smoothness. We present the following lemma to formalize this claim, whose proof can be found in Appendix A.

Lemma 1. If f is $\mathbf{L}$-separable smooth, then it is also ℓ_∞-smooth with $L_\infty = \|\mathbf{L}\|_1$.

For a unified view of this non-Euclidean smoothness, we refer the readers to Yu et al. [2026a].

3.2 Upper Bound Theory of SignSGD

For completeness, the whole algorithmic procedure of SignSGD is listed in Algorithm 1. Though the upper bounds of SignSGD and its variants under various settings have been widely studied [Bernstein et al., 2018, Karimireddy et al., 2019, Sun et al., 2023, Jiang et al., 2024, 2025b, 2026, Yu et al., 2026e], we present the upper bound analysis here for consistency of the paper.

Theorem 1. Run Algorithm 1 for T iterations under Assumptions 1a to 3a, with the hyperparameters specified in (10). Denote $N = BT$ as the total complexity, Algorithm 1 guarantees:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|_1\right] = \mathcal{O}\left(\sqrt{\frac{L_\infty \Delta}{N}} + \left(\frac{\|\boldsymbol{\sigma}\|_1^2 L_\infty \Delta}{N}\right)^{\frac{1}{4}}\right). \tag{1}$$

3.3 Lower Bound Theory of SignSGD

After establishing an upper bound for Algorithm 1, we present a corresponding lower bound, which immediately verifies the sharpness of the upper bound in Theorem 1.

Theorem 2. For any given $L_\infty, \Delta > 0$, $\boldsymbol{\sigma} \in \mathbb{R}_+^d$, and deterministic initial point $\mathbf{x}_0 \in \mathbb{R}^d$, fix integers $B, T \geq 1$ in the parameter range specified in Equation (17) and an arbitrary constant step size $\eta > 0$. There exist a function $f : \mathbb{R}^d \to \mathbb{R}$ and a stochastic gradient oracle such that:

1. f satisfies Assumptions 1a and 3a;
2. the per-sample stochastic gradients $\mathbf{g}_t^b$ satisfy Assumption 2a;
3. denote $N = BT$, the iterates generated by Algorithm 1 satisfy

$$\mathbb{E}\left[\min_{0\leq t<T}\|\nabla f(\mathbf{x}_t)\|_1\right] = \Omega\left(\sqrt{\frac{L_\infty \Delta}{N}} + \left(\frac{\|\boldsymbol{\sigma}\|_1^2 L_\infty \Delta}{N}\right)^{\frac{1}{4}}\right). \tag{2}$$

To the best of our knowledge, Theorem 2 is the first known tight lower bound for the SignSGD algorithm that does not depend on the dimension d explicitly. We briefly illustrate our proof techniques below, with the full analysis postponed to Appendix B.2.

Proof Sketch. To analyze the complexity of SignSGD, we divide the proof into four main steps.

Step 1: Dimensional Decomposition. We first reduce the d-dimensional optimization into d parallel one-dimensional problems by constructing a separable objective function $f(\mathbf{x}) = \sum_{i=1}^{d} p_i(\mathbf{x}_i)$,

where $\mathbf{x}_i$ denotes the i-th coordinate. Utilizing Lemma 1, we can safely distribute the global ℓ_∞-smoothness constant L_∞ and the suboptimality Δ across coordinates such that $\|\mathbf{L}\|_1 = L_\infty$ and $\sum_{i=1}^d \Delta_i = \Delta$.

Step 2: Constructing a 1D Resisting Oracle. We establish the worst-case lower bound for a single coordinate by constructing a "resisting oracle" [Nesterov et al., 2018]. Specifically, in the deterministic setting, we design a hard 1D function p_i that maintains a constant slope $p_i'(x) = -\epsilon$ at a finite sequence of equally spaced query points. Consequently, iterates that remain on this grid have gradient magnitude ϵ.

Step 3: Inducing Stall via Adversarial Bimodal Noise. To extend the 1D construction to the stochastic setting, we introduce an adversarial noise distribution. At these grid points, we define a specialized gradient oracle

$$\mathbb{P}\left(g_t^b = 0 \mid x_t\right) = \frac{\sigma_i^2}{\sigma_i^2 + \epsilon^2}, \quad \mathbb{P}\left(g_t^b = \frac{\sigma_i^2 + \epsilon^2}{\epsilon^2} p_i'(x_t) \mid x_t\right) = \frac{\epsilon^2}{\sigma_i^2 + \epsilon^2}. \tag{3}$$

At these grid points, the oracle is unbiased with variance σ_i^2; outside the grid, it returns the exact derivative. Crucially, this specific noise structure is designed to force the optimizer to stall, so that more total steps are needed to escape from the previously defined "resisting oracle" when the variance σ_i^2 is large. This yields the one-dimensional lower bound:

$$\mathbb{E}\left[\min_t |p'(x_t)|\right] \geq C \max\left\{\sqrt{\frac{L\Delta}{N}}, \left(\frac{L\Delta\sigma^2}{N}\right)^{\frac{1}{4}}\right\}.$$

Step 4: Dimensional Lifting and Adversarial Tuning. Finally, we aggregate the 1D lower bounds across all d dimensions. By adversarially tuning the coordinate-wise smoothness ($L_i \propto \sigma_i$) and the suboptimality (Δ_i) subject to their global constraints, we maximize the aggregated lower bound. This step successfully lifts the 1D result to the d-dimensional setting and recovers the exact dependency on $\|\boldsymbol{\sigma}\|_1$ as stated in the theorem. □

4 Why Sign Operator Works: Comparison Between SignSGD and SGD

In this section, we first give the lower bound of SGD, and then compare SignSGD with SGD under theoretical perspectives. This comparison highlights the role of sign descent beyond its classical interpretation as a gradient-compression mechanism [Bernstein et al., 2018]: under Assumption 3a, the sign operator can also lead to provably improved complexity bounds.

4.1 Theoretical Study: SignSGD Converges Provably Faster than SGD

We introduce the following ℓ_1-norm lower bound for SGD. This result characterizes the intrinsic hardness faced by SGD under our ℓ_∞-smooth geometry, and therefore serves as the key benchmark for comparison with the upper bound of SignSGD.

Theorem 3 (Mini-batch SGD lower bound)**.** For any given $d \geq 1$, $L_\infty, \Delta > 0$, $\boldsymbol{\sigma} \in \mathbb{R}_+^d$, and deterministic initial point $\mathbf{x}_0 \in \mathbb{R}^d$, fix an arbitrary constant step size $\eta > 0$ and consider vanilla mini-batch SGD with batch size B and T iterations in the parameter range specified in Equation (20). There exists a function $f : \mathbb{R}^d \to \mathbb{R}$ and a globally unbiased stochastic gradient oracle such that:

1. f satisfies Assumptions 1a and 3a, with $f(\mathbf{x}_0) - \inf f = \Delta$;
2. the per-sample stochastic gradients $\mathbf{g}_t^b$ satisfy Assumption 2a;
3. denote $N = BT$, the iterates satisfy

$$\mathbb{E}\left[\min_{0 \leq t < T} \|\nabla f(\mathbf{x}_t)\|_1\right] = \Omega\left(\sqrt{\frac{dL_\infty\Delta}{N}} + \left(\frac{d\|\boldsymbol{\sigma}\|_2^2 L_\infty \Delta}{N}\right)^{\frac{1}{4}}\right). \tag{4}$$

The proof is deferred to Appendix C.

Comparison based on concentration coefficient $\chi(\boldsymbol{\sigma})$. We proceed to compare the rate obtained in Theorem 1, which is the *upper bound* for SignSGD, and Theorem 3, which is the *lower bound* for SGD. We define the concentration coefficient $\chi : \mathbb{R}^d \setminus \{\mathbf{0}\} \to [1, d]$ as follows:

$$\chi(\boldsymbol{\sigma}) = \frac{d\|\boldsymbol{\sigma}\|_2^2}{\|\boldsymbol{\sigma}\|_1^2} \in [1, d]. \tag{5}$$

The above function χ is a measure of how "concentrated" a vector is. $\chi(\boldsymbol{\sigma}) \approx d$ corresponds to highly skewed noise acting on only a few coordinates (with $\chi(\boldsymbol{\sigma}) = d$ for a single coordinate), whereas $\chi(\boldsymbol{\sigma}) = 1$ represents perfectly uniform noise across all parameters. To find an ϵ-stationary point in the ℓ_1-norm, the sample-complexity bounds are

SignSGD

$$\Theta\left(\frac{L_\infty \Delta}{\epsilon^2} + \frac{\|\boldsymbol{\sigma}\|_1^2 L_\infty \Delta}{\epsilon^4}\right) \tag{6}$$

SGD

$$\Omega\left(\frac{dL_\infty \Delta}{\epsilon^2} + \frac{\chi(\boldsymbol{\sigma})\|\boldsymbol{\sigma}\|_1^2 L_\infty \Delta}{\epsilon^4}\right) \tag{7}$$

In view of the above two complexity bounds, we make the following observations.

1. **Deterministic term.** For the first noiseless term, SignSGD reduces the complexity of SGD by a factor of d, revealing its superior dependence on the dimensional factors.
2. **Stochastic term.** Since $1 \le \chi(\boldsymbol{\sigma}) \le d$, the displayed stochastic terms coincide at $\chi(\boldsymbol{\sigma}) = 1$, while their ratio reaches d at $\chi(\boldsymbol{\sigma}) = d$ (one-coordinate noise).

As shown above, our results provide the first problem setting in which SignSGD achieves a provably better complexity bound (in terms of the dimensional dependence) than SGD.

5 Matrix Optimizers

While vector-based sign methods effectively handle separable scaling, modern LLM architectures rely heavily on matrix multiplications. This motivates extending our framework to the matrix domain. Now that we have already illustrated the edge of SignSGD in Section 4, we will show in the sequel that it's natural to extend our previous framework from vectors to the matrix domain.

5.1 Notations and Assumptions

Throughout this section, we consider the matrix optimization problem $\min_{\mathbb{R}^{m\times n}} F(\mathbf{W})$ under the following assumptions. We suppose that matrices are oriented so that $m \le n$.

We denote the set of $m \times m$ positive semi-definite (PSD) matrices by $\mathbb{S}^m$. For a matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ of rank r, we write its compact singular value decomposition (SVD) as $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{m\times r}$ and $\mathbf{V} \in \mathbb{R}^{n\times r}$ have orthonormal columns and $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ retains only the positive singular values $\sigma_1 \ge \cdots \ge \sigma_r > 0$. The matrix sign operator is defined as $\operatorname{msign}(\mathbf{X}) := \mathbf{U}\mathbf{V}^\top$, with $\operatorname{msign}(\mathbf{0}) := \mathbf{0}$. In this section, we consider pure polar-update Muon, which uses no Newton–Schulz iterations or momentum. The inner product between matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}$ is denoted by $\langle \mathbf{Y}, \mathbf{X}\rangle := \operatorname{Tr}\left(\mathbf{Y}^\top \mathbf{X}\right)$. We let $\|\cdot\|_{\mathrm{op}}$, $\|\cdot\|_*$, and $\|\cdot\|_{\mathrm{F}}$ denote the matrix operator norm, nuclear norm, and Frobenius norm, respectively. For a vector $\mathbf{v} \in \mathbb{R}^m$, $\operatorname{diag}(\mathbf{v}) \in \mathbb{R}^{m\times m}$ denotes the diagonal matrix whose elements are given by $(\operatorname{diag}(\mathbf{v}))_{ij} = v_i$ if $i = j$ and 0 otherwise.

Assumption 1b (Lower bounded objective)**.** The function $F : \mathbb{R}^{m\times n} \to \mathbb{R}$ is bounded from below. There exists $F^* > -\infty$ such that $F(\mathbf{W}) \ge F^*$ holds for all $\mathbf{W} \in \mathbb{R}^{m\times n}$. We further denote $\Delta = F(\mathbf{W}_0) - \inf_{\mathbf{W}\in\mathbb{R}^{m\times n}} F(\mathbf{W})$.

Assumption 2b (Anisotropic matrix variance bound)**.** At step t we observe a mini-batch of gradients $G_t = \{\mathbf{G}_t^1, \cdots, \mathbf{G}_t^B\}$, mutually independent conditioned on $\mathcal{F}_{t-1}$ and satisfying $\mathbb{E}\left[\mathbf{G}_t^b | \mathcal{F}_{t-1}\right] = \nabla F(\mathbf{W}_t), \forall b \in [B]$ where $\mathcal{F}_t = \sigma(G_0, \ldots, G_t)$ denotes the natural filtration, with $\mathcal{F}_{-1}$ trivial. Denote by $\mathbf{N}_t^b = \mathbf{G}_t^b - \nabla F(\mathbf{W}_t)$, there exists $\mathbf{\Sigma} \in \mathbb{S}^m$ such that

$$\mathbb{E}\left[(\mathbf{N}_t^b)(\mathbf{N}_t^b)^\top | \mathcal{F}_{t-1}\right] \preceq \mathbf{\Sigma}^2, \quad \forall b \in [B].$$

Assumption 3b (Spectral norm smoothness)**.** We say $F : \mathbb{R}^{m\times n} \to \mathbb{R}$ is L_*-spectral norm smooth if for any $\mathbf{W}, \mathbf{W}' \in \mathbb{R}^{m\times n}$, it holds that

$$|F(\mathbf{W}') - (F(\mathbf{W}) + \langle \nabla F(\mathbf{W}), \mathbf{W}' - \mathbf{W}\rangle)| \le \frac{L_*}{2}\|\mathbf{W}' - \mathbf{W}\|_{\mathrm{op}}^2.$$

Assumption 2b can be viewed as an extension of Assumption 2a into matrix space, and has been widely used for analyzing Muon and other matrix optimizers [An et al., 2025, Pan et al., 2025]. Assumption 3b accounts for the distinct structure of matrix parameters, which can be viewed as an extension of Assumption 3a into matrix space. This extension also connects SignSGD with Muon, as Bernstein and Newhouse [2024] pointed out that SignSGD can be regarded as *steepest descent under infinity norm*, while Muon can be regarded as *steepest descent under spectral norm*.

5.2 Upper Bound Theory of Muon

The Muon algorithm is presented in Algorithm 2. We state its upper bound as below, with the proof deferred to Appendix D.1.

Theorem 4 (Muon upper bound)**.** Run Algorithm 2 for T iterations under Assumptions 1b to 3b, with the hyperparameters specified in (26). Denote $N = BT$ as the total complexity, we have:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\mathbf{W}_t)\|_*\right] = \mathcal{O}\left(\sqrt{\frac{L_*\Delta}{N}} + \left(\frac{\|\mathbf{\Sigma}\|_*^2 L_*\Delta}{N}\right)^{1/4}\right). \tag{8}$$

A key feature of Theorem 4 is that the bound does not depend explicitly on the matrix dimension $\min\{m, n\}$, in sharp contrast to existing analyses of Muon that incur such dimension dependence [Li and Hong, 2025, Shen et al., 2025, Chang et al., 2025, Huang et al., 2025].

5.3 Lower Bound Theory of Muon

After establishing the upper bound for Algorithm 2, we present a corresponding lower bound, which verifies the sharpness of Theorem 4. The proof is deferred to Appendix D.2.

Theorem 5 (Muon lower bound)**.** For any given $L_*, \Delta > 0$, $\mathbf{\Sigma} \succeq 0$, and deterministic initial point $\mathbf{W}_0 \in \mathbb{R}^{m\times n}$, fix integers $B, T \geq 1$ in the parameter range specified in Equation (33) and constant step size $\eta > 0$. There exist a function $F : \mathbb{R}^{m\times n} \to \mathbb{R}$ and a stochastic gradient oracle such that:

1. F satisfies Assumptions 1b and 3b;
2. the per-sample stochastic gradients $\mathbf{G}_t^b$ satisfy Assumption 2b;
3. denote $N = BT$, the iterates generated by Algorithm 2 satisfy

$$\mathbb{E}\left[\min_{0\leq t<T}\|\nabla F(\mathbf{W}_t)\|_*\right] = \Omega\left(\sqrt{\frac{L_*\Delta}{N}} + \left(\frac{\|\mathbf{\Sigma}\|_*^2 L_*\Delta}{N}\right)^{1/4}\right). \tag{9}$$

To our best knowledge, Theorem 5 is the first lower complexity bound for Muon. Existing stochastic lower bounds, such as Arjevani et al. [2023], are formulated for vector-valued methods under Euclidean geometry, and thus do not directly apply to spectral-norm smoothness and nuclear-norm stationarity. Our theorem fills this gap and shows that the dimension-free upper bound in Theorem 4 is unimprovable in our matrix geometry. We outline the proof below.

Proof Sketch. Our proof establishes a strict geometric and dynamic equivalence, mathematically reducing the Muon optimization in the matrix domain to the SignSGD optimization in the vector domain. We outline the proof in the following five steps.

Step 1: Constructing the Matrix Objective with Orthogonal Alignments. We lift the hard vector instance $f(\mathbf{x})$ into the matrix domain by defining $F(\mathbf{W}) = \sum_{i=1}^m f_i\big((\mathbf{Q}\mathbf{W}\mathbf{P}^\top)_{ii}\big)$. The reasons why we additionally introduce the orthogonal matrix $\mathbf{Q}$ and the projection matrix $\mathbf{P}$ here are:

- $\mathbf{Q}$ is explicitly used to align the separable vector noise with the target noise-scale matrix $\mathbf{\Sigma}$;
- $\mathbf{P}$ is used as the projection matrix, projecting the $m \times n$ matrix into an $m \times m$ subspace.

Step 2: Designing the Structured Matrix Oracle. To mirror the stochasticity, we construct a structured stochastic matrix gradient $\mathbf{G}_t = \mathbf{Q}^\top \operatorname{diag}(\mathbf{g}_t)\mathbf{P}$. Driven by the alignment in Step 1, this specific design perfectly satisfies the required noise-scale matrix $\mathbf{\Sigma} = \mathbf{Q}^\top \operatorname{diag}(\boldsymbol{\sigma})\mathbf{Q}$, while safely isolating the worst-case vector noise $\mathbf{g}_t$ strictly within the projected subspace.

Step 3: Establishing Trajectory Equivalence via SVD. We factor the constructed gradient over the active coordinates $S = \{i : g_{t,i} \neq 0\}$. Let $\mathbf{E}_S$ be the coordinate selector collecting the corresponding standard basis vectors, and define $\mathbf{U} = \mathbf{Q}^\top\mathbf{E}_S \operatorname{diag}(\operatorname{sign}(\mathbf{g}_{t,S}))$, $\mathbf{\Sigma} = \operatorname{diag}(|\mathbf{g}_{t,S}|)$, and $\mathbf{V} = \mathbf{P}^\top\mathbf{E}_S$, where $\mathbf{g}_{t,S}$ denotes the subvector of $\mathbf{g}_t$ indexed by S. The diagonal entries of $\mathbf{\Sigma}$ are strictly positive and $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns, so $\mathbf{G}_t = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is the thin SVD up to the ordering of the columns. The SVD orthogonalization therefore acts as a scalar $\operatorname{sign}(\cdot)$ operator on the active diagonal entries: $\operatorname{msign}(\mathbf{G}_t) = \mathbf{Q}^\top\mathbf{E}_S \operatorname{diag}(\operatorname{sign}(\mathbf{g}_{t,S}))\mathbf{E}_S^\top\mathbf{P} = \mathbf{Q}^\top \operatorname{diag}(\operatorname{sign}(\mathbf{g}_t))\mathbf{P}$, with $\operatorname{msign}(\mathbf{G}_t) = \mathbf{0}$ when S is empty. Consequently, Muon's matrix update $\mathbf{W}_{t+1} = \mathbf{W}_t - \eta \operatorname{msign}(\mathbf{G}_t)$ mathematically collapses into the exact SignSGD vector update: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \operatorname{sign}(\mathbf{g}_t)$.

Step 4: Aligning Geometric Constraints. We establish a rigorous one-to-one mapping between the governing metrics of both domains. We prove that the vector ℓ_∞-smoothness flawlessly translates to the matrix spectral norm smoothness ($L_* = L_\infty$), and the trace norm of the noise-scale matrix matches the ℓ_1-norm of the vector noise ($\|\mathbf{\Sigma}\|_* = \|\boldsymbol{\sigma}\|_1$).

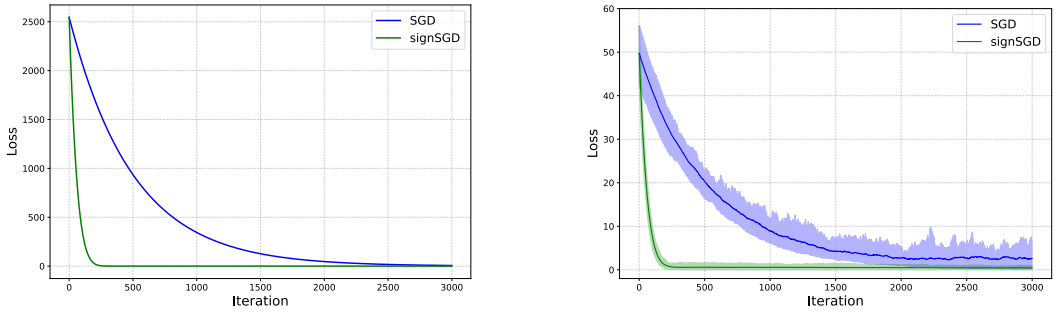
Step 5: Transferring the Complexity Bound. Finally, because orthogonal transformations preserve singular values, we show that the trace norm (nuclear norm) of the matrix gradient $\|\nabla F(\mathbf{W}_t)\|_*$ is precisely equivalent to the ℓ_1-norm of the vector gradient $\|\nabla f(\mathbf{x}_t)\|_1$, enabling us to directly invoke the vector bounds from Theorem 2 to establish the exact $\Omega(\cdot)$ lower bound for Muon. □

6 Empirical Study

We examine controlled numerical examples, optimizer comparisons on image classification and language modeling, and direct coordinate-noise measurements.

6.1 Experiments on Numerical Examples

To amplify the difference between the two optimizers, we manually crafted two numerical example functions that fit into the extreme cases stated in Section 4.1. The experimental results of these numerical examples are shown in Figure 1. Our findings are summarized as follows.



(a) Deterministic case with anisotropic curvature.

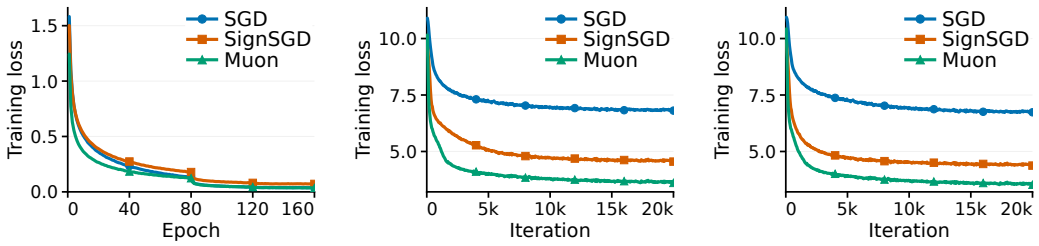
(b) Stochastic example with sparse noise.

Figure 1: Training loss of SignSGD and SGD on the two quadratic examples.

1. **Deterministic example.** SignSGD outperforms SGD when facing a ℓ_∞-geometry target and a sufficiently large d. Therefore, we set our target function as an imbalanced quadratic objective $f(\mathbf{x}) = 1/2 \sum_{i=1}^{d} L_i x_i^2$ for $\mathbf{x} \in \mathbb{R}^d$. To better illustrate the difference, we set $d = 5000$ and $L_1 = 1000, L_i = 1$ for $i \geq 2$. From Figure 1a, we can observe that SGD performs worse than SignSGD because its learning rate is constrained by the largest coordinate curvature L_1.
2. **Stochastic example.** SignSGD outperforms SGD when facing sparse noise structure. We switch to a simple quadratic objective function $f(x) = \|\mathbf{x}\|_2^2/2$ for $\mathbf{x} \in \mathbb{R}^{100}$. Then we add Gaussian noise $\mathcal{N}(0, 100^2)$ only to the first component of the gradient, which further maximizes the sparsity of $\boldsymbol{\sigma}$ as $\chi(\boldsymbol{\sigma}) = d$ in this setting. Results in Figure 1b align with our findings, as SignSGD consistently outperforms SGD facing sparse noise.

6.2 Experiments on CNNs and LLM Pretraining

We compare SGD, SignSGD [Bernstein et al., 2018], and practical Muon [Jordan et al., 2024] on ResNet-20 [He et al., 2016] trained on CIFAR-10 [Krizhevsky et al., 2009], and on nanoGPT [Karpathy, 2022] and LLaMA trained on C4 [Raffel et al., 2020]. Each configuration uses three training seeds and the same training budget. In Figure 2, we observe that SignSGD and Muon reach lower training losses than SGD, while the three optimizers attain similar results for the CNN model. Complete curves, final metrics, and implementation differences are reported in Appendix E.1.



(a) ResNet-20/CIFAR-10 (b) nanoGPT/C4 (c) LLaMA/C4

Figure 2: Training losses of SGD, SignSGD, and Muon under different settings.

6.3 Direct Coordinate-Noise Measurements

To understand the contrasting performance of SGD and SignSGD across CNN and LLM training, we compare the coordinate-wise gradient-noise variances along the corresponding SGD trajectories.

As illustrated in Figure 3, the coordinate-noise distributions differ between CNN and LLM training with SGD. ResNet-20's noise becomes more homogeneous during training, whereas nanoGPT and LLaMA exhibit stronger coordinate-wise variance heterogeneity and correspondingly a larger estimated concentration coefficient $\widehat{\chi}$ at all five measured stages. Measurement details are reported in Appendix E.2.

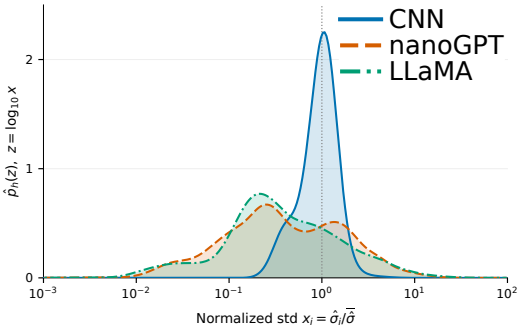


Figure 3: SGD coordinate-wise noise std distributions, estimated at 50% of training.

These findings provide a noise-geometry interpretation of the contrasting optimization behavior in Figure 2. SGD's update magnitudes depend on gradient magnitudes, so coordinates with larger noise scales can dominate its stochastic updates. SignSGD uses equal-magnitude coordinate-wise sign updates [Bernstein et al., 2018], avoiding this direct dependence on gradient magnitudes. The more homogeneous CNN noise leaves less scope for this noise-dependent advantage, and SGD attains lower training loss than SignSGD. In LLM pretraining, the stronger variance heterogeneity accompanies lower training losses for SignSGD than for SGD. Our theory in Section 4.1 identifies larger $\chi(\boldsymbol{\sigma})$ with a greater noise-dependent complexity advantage for SignSGD, connecting the noise distributions and update rules to the observed cross-model performance contrast.

7 Conclusion

In this work, we bridge the longstanding gap between the empirical superiority of sign-based optimizers and their theoretical guarantees in non-convex stochastic optimization. By moving from the standard ℓ_2 framework to an ℓ_∞-smooth, coordinate-wise noise setting with ℓ_1-stationarity, we give an optimal complexity-theoretic characterization of SignSGD and identify a problem geometry in which it provably outperforms SGD. Specifically, we prove matching upper and lower bounds for SignSGD, establish a strictly worse lower bound for SGD, and show that the resulting separation becomes most pronounced under sparse or highly heterogeneous noise. We further extend this framework to matrix optimization and obtain matching bounds for Muon under anisotropic matrix variance, spectral norm smoothness, and nuclear norm stationarity. Our empirical results corroborate these theoretical findings, with both SignSGD and Muon outperforming SGD in nanoGPT and LLaMA pretraining. Controlled sparse noise toy problems and direct coordinate-noise measurements during LLM pretraining reveal the skewed gradient noise structure leveraged by our theory, supporting the viewpoint that sign-based methods enjoy a genuine geometric advantage in real-world high-dimensional problems, such as LLM pretraining.

Acknowledgments and Disclosure of Funding

This work was partially supported by NSFC (U23A20382), the "111 Center" (No. B26023), and the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118). Dingzhi Yu is supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (NO.26CXJH0364). The authors declare no competing interests.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Kang An, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, and Tong Zhang. ASGO: Adaptive structured gradient optimization. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pages 126775–126814, 2025.

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.

Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 404–413, 2018.

Lukas Balles, Fabian Pedregosa, and Nicolas Le Roux. The geometry of sign gradient descent. *arXiv preprint arXiv:2002.08056*, 2020.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *OPT 2024: Optimization for Machine Learning*, 2024. URL https://openreview.net/forum?id=ux18f5nOpD.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. SignSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 560–569, 2018.

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. SignSGD with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations (ICLR)*, 2019.

Åke Björck and Clazett Bowie. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364, 1971.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 1877–1901, 2020.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120, 2020.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points II: First-order methods. *Mathematical Programming*, 185(1):315–355, 2021.

Da Chang, Yongxiang Liu, and Ganzhao Yuan. On the convergence of Muon and beyond. *arXiv preprint arXiv:2509.15816*, 2025.

Lizhang Chen, Bo Liu, Kaizhao Liang, and Qiang Liu. Lion secretly solves a constrained optimization: As lyapunov predicts. In *International Conference on Learning Representations (ICLR)*, pages 35404–35439, 2024.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 49205–49233, 2023.

Xin Cheng, Wangding Zeng, Damai Dai, Qinyu Chen, Bingxuan Wang, Zhenda Xie, Kezhao Huang, Xingkai Yu, Zhewen Hao, Yukun Li, et al. Conditional memory via scalable lookup: A new axis of sparsity for large language models. *arXiv preprint arXiv:2601.07372*, 2026.

Sinho Chewi, Sébastien Bubeck, and Adil Salim. On the complexity of finding stationary points of smooth functions in one dimension. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory (ALT)*, pages 358–374, 2023.

Evgenii Chzhen and Sholom Schechtman. SignSVRG: fixing SignSGD via variance reduction. *arXiv preprint arXiv:2305.13187*, 2023.

Michael Crawshaw and Mingrui Liu. Complexity lower bounds of adaptive gradient algorithms for non-convex stochastic optimization under relaxed smoothness. In *International Conference on Learning Representations (ICLR)*, 2025.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized SignSGD. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 9955–9968, 2022.

Michael Crawshaw, Chirag Modi, Mingrui Liu, and Robert M Gower. An exploration of non-Euclidean gradient descent: Muon and its many variants. *arXiv preprint arXiv:2510.09827*, 2025.

DeepSeek-AI. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence, 2026.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

Yiming Dong, Huan Li, and Zhouchen Lin. Convergence rate analysis of LION. *arXiv preprint arXiv:2411.07724*, 2024.

Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 2658–2667, 2020.

Zhehang Du and Weijie Su. The newton-muon optimizer. *arXiv preprint arXiv:2604.01472*, 2026.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research (JMLR)*, 12(7), 2011.

Kevin Frans, Pieter Abbeel, and Sergey Levine. What really matters in matrix-whitening optimizers? *arXiv preprint arXiv:2510.25000*, 2025.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 30016–30030, 2022.

Feihu Huang, Yuning Luo, and Songcan Chen. LiMuon: Light and fast Muon optimizer for large models. *arXiv preprint arXiv:2509.14562*, 2025.

Ruichen Jiang, Devyani Maladkar, and Aryan Mokhtari. Provable complexity improvement of AdaGrad over SGD: Upper and lower bounds in stochastic non-convex optimization. In *Proceedings of the 38th Conference on Learning Theory (COLT)*, pages 3124–3158, 2025a.

Wei Jiang, Sifan Yang, Wenhao Yang, and Lijun Zhang. Efficient sign-based optimization: Accelerating convergence via variance reduction. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pages 33891–33932, 2024.

Wei Jiang, Dingzhi Yu, Sifan Yang, Wenhao Yang, Zechao Li, and Lijun Zhang. Better convergence guarantees for sign-based momentum methods. *arXiv preprint arXiv:2507.12091*, 2025b.

Wei Jiang, Mao Xu, Wenhao Yang, Yibo Wang, Zechao Li, and Lijun Zhang. Convergence analysis of the Lion optimizer in centralized and distributed settings. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, pages 53308–53333, 2026.

Richeng Jin, Yufan Huang, Xiaofan He, Huaiyu Dai, and Tianfu Wu. Stochastic-Sign SGD for federated learning with theoretical guarantees. *arXiv preprint arXiv:2002.10940*, 2020.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 3252–3261, 2019.

Andrej Karpathy. NanoGPT, 2022. URL https://github.com/karpathy/nanoGPT.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Nikita Kornilov, Philip Zmushko, Andrei Semenov, Mark Ikonnikov, Alexander Gasnikov, and Alexander Beznosikov. Sign operator for coping with heavy-tailed noise in non-convex optimization: High probability bounds under (L_0, L_1)-smoothness. *arXiv preprint arXiv:2502.07923*, 2025.

Zdislav Kovarik. Some iterative methods for improving orthonormality. *SIAM Journal on Numerical Analysis*, 7(3):386–389, 1970.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between SGD and Adam on transformers, but sign descent might be. In *International Conference on Learning Representations (ICLR)*, 2023.

Huan Li, Yiming Dong, and Zhouchen Lin. On the $O(\sqrt{d}/T^{1/4})$ convergence rate of rmsprop and its momentum extension measured by ℓ_1 norm. *Journal of Machine Learning Research (JMLR)*, 26(131):1–25, 2025a.

Jiaxiang Li and Mingyi Hong. A note on the convergence of Muon and further. *arXiv preprint arXiv:2502.02900*, 2025.

Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. NorMuon: Making Muon more efficient and scalable. *arXiv preprint arXiv:2510.05491*, 2025b.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in Muon. *arXiv preprint arXiv:2601.13474*, 2026.

Sushant Mehta, Raj Dandekar, Rajat Dandekar, and Sreedath Panat. Muon: Training and trade-offs with latent attention and moe. *arXiv preprint arXiv:2509.24406*, 2025.

Arkadi Semen Nemirovski and David Berkovich Yudin. Problem complexity and method efficiency in optimization. *Wiley-Interscience*, 1983.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Saurabh Page, Advait Joshi, and SS Sonawane. Muonall: Muon variant for efficient finetuning of large language models. *arXiv preprint arXiv:2511.06086*, 2025.

Rui Pan, Haishan Ye, and Tong Zhang. Eigencurve: Optimal learning rate schedule for SGD on quadratic objectives with skewed Hessian spectrums. In *International Conference on Learning Representations (ICLR)*, 2022.

Rui Pan, Yuxing Liu, Xiaoyu Wang, and Tong Zhang. Accelerated convergence of stochastic heavy ball method under anisotropic gradient noise. In *International Conference on Learning Representations (ICLR)*, pages 55989–56028, 2024.

Rui Pan, Yang Luo, Yuxing Liu, Yang You, and Tong Zhang. Unbiased gradient low-rank projection. *arXiv preprint arXiv:2510.17802*, 2025.

Hanyang Peng, Shuang Qin, Yue Yu, Fangqing Jiang, Hui Wang, and Zhouchen Lin. Simple convergence proof of Adam from a sign-like descent perspective. *arXiv preprint arXiv:2507.05966*, 2025.

Xun Qian, Hussein Rammal, Dmitry Kovalev, and Peter Richtarik. Muon is provably faster with momentum variance reduction. *arXiv preprint arXiv:2512.16598*, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(140):1–67, 2020.

Martin Riedmiller and Heinrich Braun. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591, 1993.

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

Mher Safaryan and Peter Richtárik. Stochastic sign descent methods: New algorithms and better theory. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 9224–9234, 2021.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Andrei Semenov, Matteo Pagliardini, and Martin Jaggi. Benchmarking optimizers for large language model pretraining. *arXiv preprint arXiv:2509.01440*, 2025.

Maria-Eleni Sfyraki and Jun-Kun Wang. Lions and Muons: Optimization via stochastic frank-wolfe under heavy-tailed noise. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, pages 109388–109428, 2026.

Ishaan Shah, Anthony M Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, et al. Practical efficiency of Muon for pretraining. *arXiv preprint arXiv:2505.02222*, 2025.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon. *arXiv preprint arXiv:2505.23737*, 2025.

Chongjie Si, Debing Zhang, and Wei Shen. Adamuon: Adaptive Muon optimizer. *arXiv preprint arXiv:2507.11005*, 2025.

Weijie Su. Isotropic curvature model for understanding deep learning optimization: Is gradient orthogonalization optimal? *arXiv preprint arXiv:2511.00674*, 2025.

Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum ensures convergence of SIGNSGD under weaker assumptions. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 33077–33099, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi K2.5: Visual Agentic Intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Amund Tveit, Bjørn Remseth, and Arve Skogvold. Muon optimizer accelerates grokking. *arXiv preprint arXiv:2504.16041*, 2025.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How Muon's spectral design benefits generalization: A study on imbalanced data. *arXiv preprint arXiv:2510.22980*, 2025a.

Bhavya Vasudeva, Jung Whan Lee, Vatsal Sharan, and Mahdi Soltanolkotabi. The rich and the simple: On the implicit bias of Adam and SGD. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, page to appear, 2025b.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent YF Tan. Muon outperforms Adam in tail-end associative memory learning. *arXiv preprint arXiv:2509.26030*, 2025.

Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. *arXiv preprint arXiv:2509.02046*, 2025.

Lei Wu, Mingze Wang, and Weijie Su. The alignment property of sgd noise and how it helps select flat minima: A stability analysis. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 4680–4693, 2022.

Nachuan Xiao, Xiaoyin Hu, and Kim-Chuan Toh. Stochastic subgradient methods with guaranteed global stability in nonsmooth nonconvex optimization. *arXiv preprint arXiv:2307.10053*, 2023.

Shuo Xie, Mohamad Amin Mohamadi, and Zhiyuan Li. Adam exploits ℓ_∞-geometry of loss landscape via coordinate-wise adaptivity. In *International Conference on Learning Representations (ICLR)*, pages 7937–7965, 2025a.

Shuo Xie, Tianhao Wang, Beining Wu, and Zhiyuan Li. A tale of two geometries: Adaptive optimizers and non-Euclidean descent. *arXiv preprint arXiv:2511.20584*, 2025b.

Sifan Yang, Hongyi Tao, Dingzhi Yu, Haomin Bai, Wei Jiang, Silin Gao, Pi Bu, and Lijun Zhang. Memory-efficient subspace optimization with residual correction: From momentum sgd to muon. *arXiv preprint arXiv:2610.19844*, 2026.

Dingzhi Yu, Wei Jiang, Hongyi Tao, Yuanyu Wan, and Lijun Zhang. Mirror descent under generalized smoothness. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, pages 150427–150481, 2026a.

Dingzhi Yu, Yuxing Liu, Rui Pan, Lijun Zhang, Yang You, and Difan Zou. Finite-time nonsmooth convergence of Signum and Muon, 2026b. URL https://www.lamda.nju.edu.cn/yudz/pdf/Finite-Time_Nonsmooth_Convergence_of_Signum_and_Muon.pdf.

Dingzhi Yu, Rui Pan, Yuxing Liu, Difan Zou, and Tong Zhang. StoSignSGD: Unbiased Structural Stochasticity Fixes SignSGD for Training Large Language Models. *arXiv preprint arXiv:2604.15416*, 2026c.

Dingzhi Yu, Hongyi Tao, Yuanyu Wan, Luo Luo, and Lijun Zhang. Open problem: Is AdamW effective under heavy-tailed noise? *arXiv preprint arXiv:2606.23676*, 2026d.

Dingzhi Yu, Hongyi Tao, Yuanyu Wan, and Lijun Zhang. Sign-based optimizers are effective under heavy-tailed noise. In *Advances in Neural Information Processing Systems 39 (NeurIPS)*, page to appear, 2026e.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. GLM-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chengxing Xie, Cunxiang Wang, et al. GLM-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 15383–15393, 2020.

Minxin Zhang, Yuxuan Liu, and Hayden Schaeffer. Adagrad meets Muon: Adaptive stepsizes for orthogonal updates. *arXiv preprint arXiv:2509.02981*, 2025.

Xinwen Zhang and Hongchang Gao. On provable benefits of Muon in federated learning. *arXiv preprint arXiv:2510.03866*, 2025.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 7654–7663, 2019.

A Proof of Lemma 1

Suppose f is $\mathbf{L}$-separable smooth, we have

$$|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle)| \leq \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_{\mathbf{L}}^2 = \frac{1}{2}\sum_{i=1}^{d} L_i(\mathbf{y}_i - \mathbf{x}_i)^2$$
$$\leq \frac{1}{2}\sum_{i=1}^{d} L_i\|\mathbf{y} - \mathbf{x}\|_\infty^2 = \frac{\|\mathbf{L}\|_1}{2}\|\mathbf{y} - \mathbf{x}\|_\infty^2.$$

Therefore f also satisfies Assumption 3a with $L_\infty = \|\mathbf{L}\|_1$.

B Analysis for SignSGD

B.1 Upper Bound for SignSGD

Proof of Theorem 1. Set the hyperparameters to

$$\eta = \sqrt{\frac{2\Delta}{L_\infty T}}, \quad B = \left\lceil \max\left\{1, \frac{\|\boldsymbol{\sigma}\|_1^2}{\Delta L_\infty} T\right\}\right\rceil. \tag{10}$$

The batch size in (10) grows with T when $\|\boldsymbol{\sigma}\|_1 > 0$. At fixed B, (13) retains the term $2\|\boldsymbol{\sigma}\|_1/\sqrt{B}$, so this theorem is not a constant-batch guarantee for arbitrarily small stationarity.

Under Assumption 3a, we have:

$$\begin{aligned}
f(\mathbf{x}_{t+1}) \leq\ & f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L_\infty}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_\infty^2 \\
\leq\ & f(\mathbf{x}_t) - \langle \nabla f(\mathbf{x}_t), \eta \operatorname{sign}(\mathbf{g}_t) \rangle + \frac{\eta^2 L_\infty}{2} \\
=\ & f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \eta(\operatorname{sign}(\nabla f(\mathbf{x}_t)) - \operatorname{sign}(\mathbf{g}_t)) \rangle \\
& - \langle \nabla f(\mathbf{x}_t), \eta \operatorname{sign}(\nabla f(\mathbf{x}_t)) \rangle + \frac{\eta^2 L_\infty}{2} \\
=\ & f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \eta(\operatorname{sign}(\nabla f(\mathbf{x}_t)) - \operatorname{sign}(\mathbf{g}_t)) \rangle - \eta\|\nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 L_\infty}{2} \\
\leq\ & f(\mathbf{x}_t) + 2\eta\|\nabla f(\mathbf{x}_t) - \mathbf{g}_t\|_1 - \eta\|\nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 L_\infty}{2},
\end{aligned}$$

where the last inequality is due to

$$\begin{aligned}
& \langle \nabla f(\mathbf{x}_t), \eta(\operatorname{sign}(\nabla f(\mathbf{x}_t)) - \operatorname{sign}(\mathbf{g}_t)) \rangle \\
&= \sum_{i=1}^{d} \nabla_i f(\mathbf{x}_t) * \eta(\operatorname{sign}(\nabla_i f(\mathbf{x}_t)) - \operatorname{sign}(\mathbf{g}_{t,i})) \\
&\leq \sum_{i=1}^{d} 2\eta\,|\nabla_i f(\mathbf{x}_t)| * \mathbb{I}(\operatorname{sign}(\nabla_i f(\mathbf{x}_t)) \neq \operatorname{sign}(\mathbf{g}_{t,i})) \\
&\leq 2\eta \sum_{i=1}^{d} |\nabla_i f(\mathbf{x}_t) - \mathbf{g}_{t,i}| * \mathbb{I}(\operatorname{sign}(\nabla_i f(\mathbf{x}_t)) \neq \operatorname{sign}(\mathbf{g}_{t,i})) \\
&\leq 2\eta \sum_{i=1}^{d} |\nabla_i f(\mathbf{x}_t) - \mathbf{g}_{t,i}| = 2\eta\|\nabla f(\mathbf{x}_t) - \mathbf{g}_t\|_1.
\end{aligned}$$

Rearranging the obtained relation and summing up yields

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|_1\right] \leq \frac{\Delta}{\eta T} + 2\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t) - \mathbf{g}_t\|_1\right] + \frac{\eta L_\infty}{2}. \tag{11}$$

we denote $\xi_t = \nabla f(\mathbf{x}_t) - \mathbf{g}_t$ as the noise of the gradient at iteration t. Conditioned on the history $\mathcal{F}_{t-1}$, the batch gradients $\mathbf{g}_t^1, \ldots, \mathbf{g}_t^B$ are mutually independent and each is centered at $\nabla f(\mathbf{x}_t)$, so that:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t) - \mathbf{g}_t\|_1\right] &= \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\xi_t\|_1\right] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\xi_t\|_1\right] \\ &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{d}\mathbb{E}[|\xi_{t,i}|] \leq \frac{\|\boldsymbol{\sigma}\|_1}{\sqrt{B}},\end{aligned} \tag{12}$$

where the inequality follows from Jensen's inequality for the variance of the mini-batch mean. Putting (12) back to (11), and setting $\eta = \sqrt{2\Delta/L_\infty T}$, We get:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|_1\right] \leq \sqrt{\frac{2\Delta L_\infty}{T}} + 2\frac{\|\boldsymbol{\sigma}\|_1}{\sqrt{B}}. \tag{13}$$

If

$$\|\boldsymbol{\sigma}\|_1^2 T \leq \Delta L_\infty, \tag{14}$$

by setting $B = 1$, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|_1\right] \leq \sqrt{\frac{2\Delta L_\infty}{T}} + 2\frac{\|\boldsymbol{\sigma}\|_1}{\sqrt{B}} \leq (2+\sqrt{2})\sqrt{\frac{\Delta L_\infty}{N}}.$$

Otherwise, we set $B = \left\lceil \|\boldsymbol{\sigma}\|_1^2/\Delta L_\infty T \right\rceil \leq 2\|\boldsymbol{\sigma}\|_1^2/\Delta L_\infty T$, which implies

$$\sqrt{\frac{2\Delta L_\infty}{T}} \leq \left(\frac{8\|\boldsymbol{\sigma}\|_1^2 \Delta L_\infty}{N}\right)^{\frac{1}{4}}. \tag{15}$$

Plugging (15) into (13), we get

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|_1\right] \leq \left(\frac{8\|\boldsymbol{\sigma}\|_1^2 \Delta L_\infty}{N}\right)^{\frac{1}{4}} + 2\left(\frac{\|\boldsymbol{\sigma}\|_1^2 \Delta L_\infty}{N}\right)^{\frac{1}{4}}. \tag{16}$$

Combining the above two cases together, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|_1\right] = \mathcal{O}\left(\sqrt{\frac{\Delta L_\infty}{N}} + \frac{(\|\boldsymbol{\sigma}\|_1)^{\frac{1}{2}}(\Delta L_\infty)^{\frac{1}{4}}}{(N)^{\frac{1}{4}}}\right).$$

This completes the proof. □

B.2 Lower Bound for SignSGD

The parameter range in Theorem 2 is

$$B, T \in \mathbb{N}_{\geq 1}, \qquad N = BT, \qquad N \geq \max\left\{1, \frac{2\|\boldsymbol{\sigma}\|_1^2}{L_\infty \Delta}\right\}. \tag{17}$$

To exploit the separable nature of SignSGD before generalizing to d dimensions, we start by bounding the complexity of Algorithm 1 in finding stationary points in one dimension:

Lemma 2. For any positive integer n, suppose that ϵ satisfies

$$\epsilon \leq \frac{1}{2\sqrt{n}} \tag{18}$$

Let $x_1 \in \mathbb{R}$ and $x_t = x_1 + (t-1)\eta$ for any $2 \leq t \leq n+1$. Then there exists a function $p : \mathbb{R} \to \mathbb{R}$ such that: (i) p has a 1-Lipschitz gradient; (ii) $p(x_1) - \inf p \leq 1$ (iii) $p'(x_t) = -\epsilon$ for any $t \in [1, n]$.

Proof. We construct the function p as:

$$p(x) = \begin{cases} -\epsilon(x - x_1) & x \in (-\infty, x_1]; \\ \chi_{t,\epsilon}(x) + c_t & x \in (x_t, x_{t+1}]; \\ \dfrac{1}{2}(x - x_{n+1})^2 - \epsilon(x - x_{n+1}) + c_{n+1} & x \in (x_{n+1}, \infty), \end{cases}$$

where the values $\{c_t\}_{t=1}^{n+1}$ are chosen to ensure that the function p is continuous. The function $\chi_{t,\epsilon}(x)$ is defined as below:

$$\chi_{t,\epsilon}(x) = \begin{cases} \dfrac{1}{2}(x - x_t)^2 - \epsilon(x - x_t) & x \in \left(x_t, \dfrac{x_t + x_{t+1}}{2}\right]; \\ -\dfrac{1}{2}(x - x_{t+1})^2 - \epsilon(x - x_{t+1}) \\ + \dfrac{(x_{t+1} - x_t)^2}{4} - (x_{t+1} - x_t)\epsilon & x \in \left(\dfrac{x_t + x_{t+1}}{2}, x_{t+1}\right]. \end{cases}$$

Without loss of generality, we can assume $x_1 = \eta$, otherwise, we simply need to perform a translation on our function p. Plugging in $x_t = t\eta$, we have:

$$\chi_{t,\epsilon}(x) = \begin{cases} \dfrac{1}{2}(x - t\eta)^2 - \epsilon(x - t\eta) & x \in \left(x_t, \dfrac{x_t + x_{t+1}}{2}\right]; \\ -\dfrac{1}{2}(x - (t+1)\eta)^2 - \epsilon(x - (t+1)\eta) + \dfrac{\eta^2}{4} - \eta\epsilon & x \in \left(\dfrac{x_t + x_{t+1}}{2}, x_{t+1}\right]. \end{cases}$$

Further, we can give the value c_t as follows:

$$c_1 = 0, \quad c_{t+1} = t\left(\frac{1}{4}\eta^2 - \eta\epsilon\right), \quad t = 1, \ldots, n. \tag{19}$$

From the definition of p, we have

$$p(x) \geq \begin{cases} 0, & x \in (-\infty, x_1]; \\ \min\left\{c_t - \dfrac{1}{2}\epsilon^2, c_{t+1}\right\}, & x \in (x_t, x_{t+1}]; \\ c_{n+1} - \dfrac{1}{2}\epsilon^2, & x \in (x_{n+1}, +\infty), \end{cases}$$

which can be written as

$$\begin{aligned} \inf p &\geq \min_{1 \leq t \leq n+1} c_t - \frac{1}{2}\epsilon^2 \\ &= \min\left\{0, n\left(\frac{1}{4}\eta^2 - \eta\epsilon\right)\right\} - \frac{1}{2}\epsilon^2. \end{aligned}$$

This implies that

$$\begin{aligned} p(\mathbf{x}_1) - \inf p &\leq \max\left\{\frac{1}{2}\epsilon^2, n\left(\eta\epsilon - \frac{1}{4}\eta^2\right) + \frac{1}{2}\epsilon^2\right\} \\ &\leq \max\left\{\frac{1}{2}\epsilon^2, n\epsilon^2 + \frac{1}{2}\epsilon^2\right\} \\ &= n\epsilon^2 + \frac{1}{2}\epsilon^2. \end{aligned}$$

With (18), we have

$$n\epsilon^2 + \frac{1}{2}\epsilon^2 \leq \frac{1}{4} + \frac{1}{8n} < 1.$$

This completes the proof. □

Lemma 3. Fix $L, \Delta, \eta > 0$, $\sigma \geq 0$, an initial point $x_0 \in \mathbb{R}$, and integers $B, T \geq 1$, and write $N = BT$. For every ϵ satisfying

$$0 < \epsilon \leq \frac{\sqrt{L\Delta}}{4}, \qquad N \leq \frac{L\Delta(\sigma^2 + 4\epsilon^2)}{128\epsilon^4},$$

there exists a lower-bounded function $p : \mathbb{R} \to \mathbb{R}$ with an L-Lipschitz derivative and a globally unbiased stochastic gradient oracle such that $p(x_0) - \inf p = \Delta$. The per-sample gradients are conditionally independent, have conditional variance at most σ^2, and the SignSGD iterates satisfy

$$\mathbb{E} \min_{0 \le t < T} |p'(x_t)| \ge \epsilon.$$

In particular, if $N \ge \max\{1, 2\sigma^2/(L\Delta)\}$, there is such an instance satisfying, for a universal constant $C > 0$,

$$\mathbb{E} \min_{0 \le t < T} |p'(x_t)| \ge C \max\left\{ \sqrt{\frac{L\Delta}{N}}, \left(\frac{L\Delta\sigma^2}{N}\right)^{1/4} \right\}.$$

Proof. Set

$$a = 2\epsilon, \qquad K = \left\lfloor \frac{L\Delta}{16\epsilon^2} \right\rfloor \ge 1, \qquad z_j = x_0 + j\eta, \quad 0 \le j \le K.$$

Initially define $p(x_0) = 0$ and

$$p'(x) = \begin{cases} -a, & x \le z_0, \\ -a + L\min\{x - z_j, z_{j+1} - x\}, & z_j \le x \le z_{j+1}, \quad 0 \le j < K, \\ -a + L(x - z_K), & x \ge z_K. \end{cases}$$

The derivative is continuous and piecewise affine with slopes of magnitude at most L, and $p'(z_j) = -a$. On each grid interval, the first quadratic piece is bounded below by $p(z_j) - a^2/(2L)$, and the second piece is concave. Since $p(z_{j+1}) - p(z_j) = L\eta^2/4 - a\eta$, it follows that

$$\begin{aligned} p(x_0) - \inf p &\le K\left(a\eta - \frac{L\eta^2}{4}\right)_+ + \frac{a^2}{2L} \\ &\le \left(K + \frac{1}{2}\right)\frac{a^2}{L} \le \frac{3\Delta}{8}. \end{aligned}$$

Here $u_+ = \max\{u, 0\}$, and $a\eta - L\eta^2/4 \le a^2/L$ holds for every $\eta > 0$.

To make the gap equal to Δ, choose $R > \max\{x_0 + T\eta, z_K + a/L\}$ and write

$$h = p'(R) > 0, \qquad c = \sqrt{L(p(R) + \Delta) + h^2/2} > 0.$$

Beyond R, let p' decrease from h to $-c$ with slope $-L$, then increase from $-c$ to zero with slope L, and remain zero thereafter. The change in function value along this tail is

$$\frac{h^2 - c^2}{2L} - \frac{c^2}{2L} = -p(R) - \Delta.$$

The new tail first increases and then decreases to $-\Delta$, so $\inf p = -\Delta$ and $p(x_0) - \inf p = \Delta$. The derivative remains L-Lipschitz. Every SignSGD query lies in $[x_0 - T\eta, x_0 + T\eta]$, so this modification does not change any queried derivative.

Let $\mathcal{G} = \{z_0, \dots, z_K\}$ and

$$q = \frac{4\epsilon^2}{\sigma^2 + 4\epsilon^2}.$$

For all $0 \le t < T$ and $b \in [B]$, draw mutually independent $\xi_t^b \sim \mathrm{Bernoulli}(q)$ and define

$$g_t^b = \begin{cases} -2\epsilon\,\xi_t^b/q, & x_t \in \mathcal{G}, \\ p'(x_t), & x_t \notin \mathcal{G}. \end{cases}$$

At every query point, the samples are conditionally independent and unbiased. Their conditional variance equals $4\epsilon^2(q^{-1} - 1) = \sigma^2$ on $\mathcal{G}$ and zero elsewhere.

Define the latent batch-success variables and their total count by

$$Z_t = \mathbf{1}\left\{\sum_{b=1}^{B} \xi_t^b \ge 1\right\}, \qquad M = \sum_{t=0}^{T-1} Z_t.$$

These variables are defined even after a possible exit from $\mathcal{G}$, and

$$\mathbb{E}M = T[1-(1-q)^B] \le Nq \le \frac{L\Delta}{32\epsilon^2}.$$

Since $K+1 > L\Delta/(16\epsilon^2)$, Markov's inequality gives

$$\mathbb{P}(M > K) \le \frac{\mathbb{E}M}{K+1} < \frac{1}{2}.$$

On the event $M \le K$, induction yields $x_t = z_{\sum_{s=0}^{t-1} Z_s}$ for every $0 \le t \le T$. Indeed, at a grid point, the algorithm moves right by η exactly when $Z_t = 1$. Thus every queried derivative has magnitude 2ϵ on this event, and

$$\mathbb{E} \min_{0\le t<T} |p'(x_t)| \ge 2\epsilon\, \mathbb{P}(M \le K) \ge \epsilon.$$

For the final bound, take

$$\epsilon = \max\left\{\sqrt{\frac{L\Delta}{32N}}, \left(\frac{L\Delta\sigma^2}{128N}\right)^{1/4}\right\}.$$

This choice satisfies the required sample-budget inequality, and $N \ge \max\{1, 2\sigma^2/(L\Delta)\}$ ensures $\epsilon \le \sqrt{L\Delta}/4$. This proves the claim. □

Lemma 3 states the complexity lower bound for Algorithm 1 for a one-dimensional function. Now we can extend to a result in dimension d, which will yield the bound in Theorem 2.

Proof of Theorem 2. Write $S = \|\boldsymbol{\sigma}\|_1$. For $S > 0$, assign

$$L_i = \frac{\sigma_i}{S} L_\infty, \qquad \Delta_i = \frac{\sigma_i}{S}\Delta$$

to each coordinate with $\sigma_i > 0$. For these coordinates,

$$\frac{\sigma_i^2}{L_i\Delta_i} = \frac{S^2}{L_\infty\Delta},$$

so Equation (17) implies the sufficient-budget condition in Lemma 3 for every coordinate. Apply that lemma with the same η, B, T, initial point $x_{0,i}$, and parameters $(L_i, \Delta_i, \sigma_i)$. For coordinates with $\sigma_i = 0$, take $p_i \equiv 0$ and a zero oracle. Choose the oracle samples independently across coordinates and set $f(\mathbf{x}) = \sum_i p_i(x_i)$.

The resulting vector samples satisfy Assumption 2a. Moreover, $\sum_i L_i = L_\infty$ and $f(\mathbf{x}_0) - \inf f = \sum_i \Delta_i = \Delta$, so Lemma 1 gives Assumption 3a. For every trajectory,

$$\min_{0\le t<T} \|\nabla f(\mathbf{x}_t)\|_1 \ge \sum_i \min_{0\le t<T} |p_i'(x_{t,i})|.$$

Taking expectations and applying Lemma 3 therefore gives

$$\begin{aligned}\mathbb{E} \min_{0\le t<T} \|\nabla f(\mathbf{x}_t)\|_1 &\ge C \sum_{i:\sigma_i>0} \max\left\{\sqrt{\frac{L_i\Delta_i}{N}}, \left(\frac{L_i\Delta_i\sigma_i^2}{N}\right)^{1/4}\right\} \\ &\ge \frac{C}{2}\left[\sqrt{\frac{L_\infty\Delta}{N}} + \left(\frac{S^2 L_\infty\Delta}{N}\right)^{1/4}\right].\end{aligned}$$

If $S = 0$, apply Lemma 3 with $(L, \Delta, \sigma) = (L_\infty, \Delta, 0)$ to one coordinate and take all remaining coordinates constant. This proves (2). □

C Proof of Theorem 3

We first establish a one-dimensional lower bound, motivated by Jiang et al. [2025a, Theorem 4.1], and then extend it to d dimensions.

Let $C > 0$ be a sufficiently large universal constant. The parameter range in Theorem 3 is

$$B, T \in \mathbb{N}_{\geq 1}, \qquad N = BT, \qquad T \geq Cd^3, \qquad N \geq C\frac{d^5\|\boldsymbol{\sigma}\|_2^2}{L_\infty \Delta}. \tag{20}$$

The constant step size $\eta > 0$ is arbitrary. At each iteration, SGD uses B conditionally independent per-sample gradients and updates

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\eta}{B}\sum_{b=1}^{B} \mathbf{g}_t^b, \qquad 0 \leq t < T.$$

All constants in the proof are universal.

Lemma 4. For any $L, \delta > 0$, $\sigma \geq 0$, $\eta > 0$, and integers $B, T \geq 1$, there exist a function $p : \mathbb{R} \to \mathbb{R}$, an initial point x_0, and a stochastic gradient oracle such that:

(i) p has an L-Lipschitz derivative, is lower bounded, and satisfies $p(x_0) - \inf p \leq \delta$;

(ii) at every query point $x \in \mathbb{R}$, the oracle samples are conditionally independent and satisfy

$$\mathbb{E}[g_t^b \mid x_t = x] = p'(x), \qquad \mathbb{E}[(g_t^b - p'(x))^2 \mid x_t = x] \leq \sigma^2, \qquad b \in [B];$$

(iii) the mini-batch SGD iterates $x_{t+1} = x_t - \frac{\eta}{B}\sum_{b=1}^{B} g_t^b$ satisfy

$$\mathbb{E}\min_{0\leq t<T} |p'(x_t)| \geq \sqrt{2L\delta} \qquad \text{if } \eta \geq 2/L.$$

If $\eta < 2/L$, they satisfy

$$\mathbb{E}\min_{0\leq t<T} |p'(x_t)| \geq \max\left\{\frac{1}{2}\sqrt{\frac{\delta}{2\eta T + 1/(2L)}}, \min\left(\sigma\sqrt{\frac{L\eta}{2B}}, \sqrt{2L\delta}\right)\right\}. \tag{21}$$

Proof. For $a > 0$, we define

$$p(x) = \begin{cases} \dfrac{Lx^2}{2}, & |x| \leq a/L, \\ a|x| - \dfrac{a^2}{2L}, & |x| > a/L. \end{cases}$$

Then $p'(x) = \mathrm{clip}(Lx, -a, a)$, so p' is L-Lipschitz and $\inf p = 0$. For $\eta \geq 2/L$, we choose $a = \sqrt{2L\delta}$, set $x_0 = a/L$, and use the exact oracle $g_t^b = p'(x_t)$. The iterates satisfy

$$x_{2t} = \frac{a}{L}, \qquad x_{2t+1} = \frac{a}{L} - \eta a \leq -\frac{a}{L}.$$

Thus $|p'(x_t)| = a$ at every iteration, and $p(x_0) - \inf p = a^2/(2L) = \delta$.

Now suppose $\eta < 2/L$. For $\sigma > 0$, we first choose

$$a = \min\left\{\sigma\sqrt{\frac{L\eta}{2B}}, \sqrt{2L\delta}\right\}, \qquad \alpha = \frac{a^2}{\sigma^2 + a^2}, \qquad x_0 = \frac{a}{L}.$$

We use the same function p and define the oracle by

$$g_t^b = \frac{\xi_t^b}{\alpha} p'(x_t), \qquad \xi_t^b \sim \mathrm{Bernoulli}(\alpha),$$

where the variables ξ_t^b, $0 \leq t < T$, $b \in [B]$, are independent. By construction,

$$\mathbb{E}[g_t^b \mid x_t = x] = p'(x),$$
$$\mathbb{E}[(g_t^b - p'(x))^2 \mid x_t = x] = (\alpha^{-1} - 1)p'(x)^2 \leq (\alpha^{-1} - 1)a^2 = \sigma^2.$$

Let $K_t = \sum_{b=1}^{B} \xi_t^b$. Whenever $|p'(x_t)| = a$, the update is

$$x_{t+1} = x_t - hK_t \operatorname{sign}\left(p'(x_t)\right), \qquad h = \frac{\eta a}{B\alpha} \geq \frac{\eta\sigma^2}{Ba} \geq \frac{2a}{L}.$$

Starting from $x_0 = a/L$, induction therefore gives $x_t \in a/L + h\mathbb{Z}$ for every t. This lattice has no point in $(-a/L, a/L)$, so $|p'(x_t)| = a$. Consequently,

$$p(x_0) - \inf p = \frac{a^2}{2L} \leq \delta, \qquad \mathbb{E} \min_{0\leq t<T} |p'(x_t)| = a = \min\left\{\sigma\sqrt{\frac{L\eta}{2B}}, \sqrt{2L\delta}\right\}.$$

For the other term in (21), we set $x_0 = 0$ and choose

$$e = \frac{1}{2}\sqrt{\frac{\delta}{2\eta T + 1/(2L)}}, \qquad \alpha = \frac{4e^2}{\sigma^2 + 4e^2}, \qquad s = \frac{2\eta e}{B\alpha}, \qquad M = \lfloor 2BT\alpha \rfloor.$$

Let $q_j = js$ for $0 \leq j \leq M$. We now define $p(x) = \int_0^x p'(u)\, du$, where

$$p'(x) = \begin{cases} -2e + Lx, & x \leq 0, \\ -2e + L\min\{x - q_j, q_{j+1} - x\}, & q_j \leq x \leq q_{j+1}, \quad 0 \leq j < M, \\ -2e + L(x - q_M), & q_M \leq x \leq q_M + 2e/L, \\ 0, & x \geq q_M + 2e/L. \end{cases}$$

The pieces agree at their endpoints and have slopes of magnitude at most L. Thus p' is globally L-Lipschitz, and $p'(q_j) = -2e$ for every grid point. Integrating the derivative gives

$$p(q_j) - p(q_{j+1}) = 2es - \frac{Ls^2}{4} \leq 2es = \frac{\eta(\sigma^2 + 4e^2)}{B},$$

$$\inf p \geq \min_{0\leq j\leq M} p(q_j) - \frac{2e^2}{L}.$$

Here $p(x) \geq p(0)$ for $x \leq 0$. On each grid interval, the function falls at most $2e^2/L$ below the smaller endpoint value, and the final tail has minimum $p(q_M) - 2e^2/L$. It follows that p is lower bounded and

$$\begin{aligned} p(0) - \inf p &\leq M\frac{\eta(\sigma^2 + 4e^2)}{B} + \frac{2e^2}{L} \\ &\leq 8e^2\eta T + \frac{2e^2}{L} = \delta. \end{aligned}$$

Let ξ_t^b, $0 \leq t < T$, $b \in [B]$, be independent $\mathrm{Bernoulli}(\alpha)$ variables and define the oracle by

$$g_t^b = \begin{cases} -2e\xi_t^b/\alpha, & x_t \in \{q_0, \ldots, q_M\}, \\ p'(x_t), & \text{otherwise.} \end{cases}$$

For every $x \in \mathbb{R}$, this gives

$$\mathbb{E}[g_t^b \mid x_t = x] = p'(x),$$

$$\mathbb{E}[(g_t^b - p'(x))^2 \mid x_t = x] = \begin{cases} 4e^2(\alpha^{-1} - 1) = \sigma^2, & x \in \{q_0, \ldots, q_M\}, \\ 0, & \text{otherwise.} \end{cases}$$

Set $K_t = \sum_{b=1}^{B} \xi_t^b$ and $S = \sum_{t=0}^{T-1} K_t$. Since S is the sum of all BT independent Bernoulli variables, we have

$$\mathbb{E}S = BT\alpha, \qquad \mathbb{P}(S > M) = \mathbb{P}(S \geq M + 1) \leq \frac{BT\alpha}{M+1} < \frac{1}{2}.$$

On the event $S \leq M$, induction gives

$$x_t = q_{\sum_{r=0}^{t-1} K_r}, \qquad x_{t+1} = q_{\sum_{r=0}^{t} K_r}, \qquad 0 \leq t < T.$$

All the indices are at most M on this event, so every queried gradient has magnitude $2e$. Therefore,

$$\mathbb{E} \min_{0\leq t<T} |p'(x_t)| \geq 2e\mathbb{P}(S \leq M) \geq e = \frac{1}{2}\sqrt{\frac{\delta}{2\eta T + 1/(2L)}}.$$

Choosing the function that gives the larger of the two lower bounds proves (21). □

Proof of Theorem 3. We first fix $\mathbf{L} = (L_1, \dots, L_d)$ with $L_i > 0$ and $\sum_i L_i = L_\infty$, and set $\delta = \Delta/d$. For each coordinate, we apply Lemma 4 with parameters (L_i, δ, σ_i) and the same η, B, T. We translate the functions so that their initial points agree with the prescribed coordinates of $\mathbf{x}_0$, and choose the samples independently across coordinates. Then $f(\mathbf{x}) = \sum_i p_i(x_i)$ satisfies Assumption 4a, and hence Assumption 3a by Lemma 1. Its initial gap is at most Δ, and the vector oracle is globally unbiased with the required coordinatewise variance. For every trajectory, we have

$$\min_{0\le t<T} \|\nabla f(\mathbf{x}_t)\|_1 = \min_{0\le t<T} \sum_i |p_i'(x_{t,i})| \ge \sum_i \min_{0\le t<T} |p_i'(x_{t,i})|.$$

Write $L_{\max} = \max_i L_i$, $L_{\min} = \min_i L_i$, and $C_\sigma = \sum_i \sigma_i \sqrt{L_i}$. If $\eta \ge 2/L_{\max}$, the coordinate with largest curvature gives a lower bound of $\sqrt{2L_{\max}\Delta/d}$. If $\eta < 2/L_{\max}$, all coordinates satisfy (21). Summing these bounds and using $\max\{a,b\} \ge (a+b)/2$, we get, for a universal $c > 0$,

$$\mathbb{E} \min_{0\le t<T} \|\nabla f(\mathbf{x}_t)\|_1 \ge c \left[\sqrt{\frac{d\Delta}{\eta T + 1/L_{\min}}} + \sum_i \min\left\{ \sigma_i \sqrt{\frac{L_i \eta}{B}}, \sqrt{\frac{L_i \Delta}{d}} \right\} \right]. \tag{22}$$

When $\eta T \le 1/L_{\min}$, the first term is at least $c\sqrt{d\Delta L_{\min}}$. When a stochastic term is capped, it gives at least $c\sqrt{L_{\min}\Delta/d}$. Otherwise, the right side of (22) is at least

$$c \left[\sqrt{\frac{d\Delta}{\eta T}} + C_\sigma \sqrt{\frac{\eta}{B}} \right].$$

Since $\eta < 2/L_{\max}$, the first summand is at least a constant multiple of $\sqrt{dL_{\max}\Delta/T}$. By the arithmetic-geometric mean inequality, the sum is also at least a constant multiple of $(d\Delta C_\sigma^2/(BT))^{1/4}$. In each of the other cases, namely $\eta \ge 2/L_{\max}$, $\eta T \le 1/L_{\min}$, or a capped stochastic term, the lower bound is at least $c\sqrt{L_{\min}\Delta/d}$. These lower bounds imply the two required rates provided that

$$T \ge C \frac{d^2 L_{\max}}{L_{\min}}, \qquad BT \ge C \frac{d^3 C_\sigma^2}{L_{\min}^2 \Delta}. \tag{23}$$

Indeed, these conditions give

$$\sqrt{\frac{dL_{\max}\Delta}{T}} \le C^{-1/2} \sqrt{\frac{L_{\min}\Delta}{d}},$$
$$\left(\frac{d\Delta C_\sigma^2}{BT}\right)^{1/4} \le C^{-1/4} \sqrt{\frac{L_{\min}\Delta}{d}}.$$

Combining these cases, under (23) we obtain

$$\mathbb{E} \min_{0\le t<T} \|\nabla f(\mathbf{x}_t)\|_1 = \Omega\left(\sqrt{\frac{dL_{\max}\Delta}{T}} + \left(\frac{d\Delta C_\sigma^2}{BT}\right)^{1/4} \right). \tag{24}$$

For the deterministic term, we choose $L_1 = (d+1)L_\infty/(2d)$ and $L_i = L_\infty/(2d)$ for $i > 1$. Then $L_{\max} = \Omega(L_\infty)$, $L_{\min} \ge L_\infty/(2d)$, and $C_\sigma^2 \le 2L_\infty \|\boldsymbol{\sigma}\|_2^2$. For $\boldsymbol{\sigma} \ne 0$, we construct a second instance with

$$L_i = \frac{L_\infty}{2}\left(\frac{\sigma_i^2}{\|\boldsymbol{\sigma}\|_2^2} + \frac{1}{d} \right) > 0.$$

This choice gives $L_{\min} \ge L_\infty/(2d)$ and $\sqrt{L_\infty/2}\|\boldsymbol{\sigma}\|_2 \le C_\sigma \le \sqrt{2L_\infty}\|\boldsymbol{\sigma}\|_2$. For both choices, (20) implies (23) after enlarging the same universal constant C. Applying (24) to these two instances and using $\max\{a,b\} \ge (a+b)/2$, we can choose one of them to obtain

$$\mathbb{E}\left[\min_{0\le t<T} \|\nabla f(\mathbf{x}_t)\|_1 \right] = \Omega\left(\sqrt{\frac{dBL_\infty\Delta}{N}} + \left(\frac{d\|\boldsymbol{\sigma}\|_2^2 L_\infty \Delta}{N}\right)^{\frac{1}{4}} \right). \tag{25}$$

Since $B \ge 1$, this implies (4). If $\boldsymbol{\sigma} = 0$, we use only the first instance.

It remains to make the initial gap equal to Δ. Each one-dimensional construction gives a gap at most its assigned budget. Since B and T are finite and the oracle has only finitely many Bernoulli outcomes, the set of all possible queried points is finite. Outside this set, we extend a tail by a sufficiently wide smooth well so that the gap equals the assigned budget. This preserves the L_i-smoothness and all queried gradients. For the stochastic clipped-quadratic construction, we also preserve $|p_i'| \le a$, so the global variance bound is unchanged. This completes the proof. □

D Analysis for Muon

D.1 Upper Bound for Muon

Proof of Theorem 4. Set the hyperparameters to

$$\eta = \sqrt{\frac{2\Delta}{L_* T}}, \quad B = \left\lceil \max\left\{1, \frac{\|\mathbf{\Sigma}\|_*^2}{\Delta L_*} T\right\}\right\rceil. \tag{26}$$

The batch size in (26) grows with T when $\|\mathbf{\Sigma}\|_* > 0$. At fixed B, (31) retains the term $2\|\mathbf{\Sigma}\|_*/\sqrt{B}$, so this theorem is not a constant-batch guarantee for arbitrarily small stationarity.

Under Assumption 3b, we have:

$$\begin{aligned}
F(\mathbf{W}_{t+1}) &\le F(\mathbf{W}_t) + \langle \nabla F(\mathbf{W}_t), \mathbf{W}_{t+1} - \mathbf{W}_t \rangle + \frac{L_*}{2}\|\mathbf{W}_{t+1} - \mathbf{W}_t\|_{\text{op}}^2 \\
&= F(\mathbf{W}_t) - \langle \nabla F(\mathbf{W}_t), \eta\, \text{msign}\,(\mathbf{G}_t) \rangle + \frac{L_*}{2}\|\eta\, \text{msign}\,(\mathbf{G}_t)\|_{\text{op}}^2 \\
&\le F(\mathbf{W}_t) - \eta \langle \mathbf{G}_t, \text{msign}\,(\mathbf{G}_t) \rangle - \eta \langle \nabla F(\mathbf{W}_t) - \mathbf{G}_t, \text{msign}\,(\mathbf{G}_t) \rangle + \frac{L_* \eta^2}{2} \\
&\le F(\mathbf{W}_t) - \eta \|\mathbf{G}_t\|_* + \eta \|\nabla F(\mathbf{W}_t) - \mathbf{G}_t\|_* \|\text{msign}\,(\mathbf{G}_t)\|_{\text{op}} + \frac{L_* \eta^2}{2} \\
&\le F(\mathbf{W}_t) - \eta \|\nabla F(\mathbf{W}_t)\|_* + 2\eta \|\nabla F(\mathbf{W}_t) - \mathbf{G}_t\|_* + \frac{L_* \eta^2}{2},
\end{aligned}$$

where we use $\|\text{msign}\,(\mathbf{G}_t)\|_{\text{op}} \le 1$ and $|\langle A, B \rangle| \le \|A\|_* \|B\|_{\text{op}}$. Summing over $t = 0, \ldots, T-1$, we get

$$F(\mathbf{W}_T) - F(\mathbf{W}_0) \le -\eta \sum_{t=0}^{T-1} \|\nabla F(\mathbf{W}_t)\|_* + 2\eta \sum_{t=0}^{T-1} \|\nabla F(\mathbf{W}_t) - \mathbf{G}_t\|_* + \frac{L_* \eta^2 T}{2}. \tag{27}$$

Under Assumption 1b, $F(\mathbf{W}_0) - F(\mathbf{W}_T) \le F(\mathbf{W}_0) - \inf F \le \Delta$, take expectation on both sides, and we get

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla F(\mathbf{W}_t)\|_*\right] \le \frac{\Delta}{\eta T} + \frac{2}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla F(\mathbf{W}_t) - \mathbf{G}_t\|_*\right] + \frac{L_* \eta}{2}. \tag{28}$$

Define $\mathbf{N}_t = \nabla F(\mathbf{W}_t) - \mathbf{G}_t$. Consider Batch size B, we denote the gradient on each sample is $\mathbf{G}_t^b$, $1 \le b \le B$, and the noise on each sample as $\mathbf{N}_t^b = \nabla F(\mathbf{W}_t) - \mathbf{G}_t^b$. We have the following lemma:

Lemma 5. (An et al. [2025, Lemma 9]) Under Assumption 2b, with batch size B, we have

$$\mathbb{E}\left[\mathbf{N}_t \mathbf{N}_t^\top\right] \preceq \frac{\mathbf{\Sigma}^2}{B}. \tag{29}$$

Under the fact that the map $X \mapsto \text{Tr}\left(X^{\frac{1}{2}}\right)$ is concave on the positive semi-definite cone, we can apply Jensen's inequality to get

$$\mathbb{E}\left[\|\mathbf{N}_t\|_*\right] = \mathbb{E}\left[\text{Tr}\left(\sqrt{\mathbf{N}_t \mathbf{N}_t^\top}\right)\right] \le \text{Tr}\left(\sqrt{\mathbb{E}\left[\mathbf{N}_t \mathbf{N}_t^\top\right]}\right) \le \frac{\|\mathbf{\Sigma}\|_*}{\sqrt{B}}. \tag{30}$$

Plugging (30) back to (28) and set $\eta = \sqrt{2\Delta / L_* T}$, we get

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla F(\mathbf{W}_t)\|_*\right] \le \sqrt{\frac{2\Delta L_*}{T}} + 2\frac{\|\mathbf{\Sigma}\|_*}{\sqrt{B}}. \tag{31}$$

If $\|\mathbf{\Sigma}\|_*^2 T \le \Delta L_*$, by setting $B = 1$, we have

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(\mathbf{W}_t)\|_*\right] \le \sqrt{\frac{2\Delta L_*}{T}} + 2\|\mathbf{\Sigma}\|_* \le (2 + \sqrt{2})\sqrt{\frac{\Delta L_*}{N}}$$

Otherwise, we set $B = \left\lceil \|\mathbf{\Sigma}\|_*^2/\Delta L_* T \right\rceil \leq 2\|\mathbf{\Sigma}\|_*^2/\Delta L_* T$, which implies

$$\sqrt{\frac{2\Delta L_*}{T}} \leq \left(\frac{8\|\mathbf{\Sigma}\|_*^2 \Delta L_*}{BT}\right)^{\frac{1}{4}}. \tag{32}$$

Bringing (32) back to (31)

$$\begin{aligned}\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\mathbf{W}_t)\|_*\right] &\leq 2\left(\frac{\|\mathbf{\Sigma}\|_*^2 \Delta L_*}{BT}\right)^{\frac{1}{4}} + \left(\frac{8\|\mathbf{\Sigma}\|_*^2 \Delta L_*}{BT}\right)^{\frac{1}{4}} \\ &= \mathcal{O}\left(\left(\frac{\|\mathbf{\Sigma}\|_*^2 \Delta L_*}{BT}\right)^{\frac{1}{4}}\right).\end{aligned}$$

Putting the above two cases together, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\mathbf{W}_t)\|_*\right] = \mathcal{O}\left(\sqrt{\frac{\Delta L_*}{N}} + \left(\frac{\|\mathbf{\Sigma}\|_*^2 \Delta L_*}{N}\right)^{\frac{1}{4}}\right).$$

□

D.2 Lower Bound for Muon

The parameter range in Theorem 5 is

$$B, T \in \mathbb{N}_{\geq 1}, \qquad N = BT, \qquad N \geq \max\left\{1, \frac{2\|\mathbf{\Sigma}\|_*^2}{L_* \Delta}\right\}. \tag{33}$$

Assume the target weight matrix is $\mathbf{W} \in \mathbb{R}^{m\times n}$. Without loss of generality, we assume $m \leq n$. We introduce a projection matrix $\mathbf{P} = [\mathbf{I}_m, \mathbf{0}_{m\times(n-m)}] \in \mathbb{R}^{m\times n}$ which pads an $m \times m$ matrix with zeros to match the $m \times n$ dimension. Let $\mathbf{\Sigma} \in \mathbb{R}^{m\times m}$ be the target noise-scale matrix, which can be diagonalized via an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{m\times m}$ as $\mathbf{\Sigma} = \mathbf{Q}^\top \operatorname{diag}(\boldsymbol{\sigma})\mathbf{Q}$.

We construct our worst-case matrix objective function by extracting the diagonal elements of the transformed matrix $\mathbf{QWP}^\top \in \mathbb{R}^{m\times m}$:

$$F(\mathbf{W}) = \sum_{i=1}^{m} f_i\big((\mathbf{QWP}^\top)_{ii}\big), \tag{34}$$

where $f(\mathbf{x}) = \sum_{i=1}^m f_i(\mathbf{x}_i)$ is the hard 1D vector instance constructed in Theorem 2. Apply that theorem with $L_\infty = L_*$, the eigenvalue vector $\boldsymbol{\sigma}$ of $\mathbf{\Sigma}$, and $x_{0,i} = (\mathbf{QW}_0\mathbf{P}^\top)_{ii}$. Since $\|\boldsymbol{\sigma}\|_1 = \|\mathbf{\Sigma}\|_*$, (33) is exactly the vector budget condition. The diagonal map is onto, with right inverse $\mathbf{x} \mapsto \mathbf{Q}^\top \operatorname{diag}(\mathbf{x})\mathbf{P}$, so $\inf F = \inf f$ and $F(\mathbf{W}_0) - \inf F = \Delta$.

First, we provide two lemmas to establish the strict geometric equivalence between the vector domain and the matrix domain.

Lemma 6. Assume that the separable vector function $f(\mathbf{x}) : \mathbb{R}^m \to \mathbb{R}$ satisfies Assumption 2a with bounded variance $\boldsymbol{\sigma}^2$. Then the constructed matrix function $F(\mathbf{W}) : \mathbb{R}^{m\times n} \to \mathbb{R}$ satisfies Assumption 2b with matrix covariance bound $\mathbf{\Sigma}^2 = \mathbf{Q}^\top \operatorname{diag}(\boldsymbol{\sigma}^2)\mathbf{Q}$.

Proof. By the chain rule, the exact gradient of $F(\mathbf{W})$ is given by:

$$\nabla F(\mathbf{W}) = \mathbf{Q}^\top \operatorname{diag}\big(\nabla f(\mathbf{x})\big)\mathbf{P}, \tag{35}$$

where $\mathbf{x} \in \mathbb{R}^m$ with $\mathbf{x}_i = (\mathbf{QWP}^\top)_{ii}$. At every matrix query $\mathbf{W}_t$, draw fresh vector samples $\mathbf{g}_t^b$, $b \in [B]$, conditionally independently given $\mathcal{F}_{t-1}$, from the global vector oracle at $x_{t,i} = (\mathbf{QW}_t\mathbf{P}^\top)_{ii}$. Lift each sample linearly as $\mathbf{G}_t^b = \mathbf{Q}^\top \operatorname{diag}(\mathbf{g}_t^b)\mathbf{P}$; write $\mathbf{g} = \mathbf{g}_t^b$ and $\mathbf{G} = \mathbf{G}_t^b$ for one such sample:

$$\mathbf{G} = \mathbf{Q}^\top \operatorname{diag}(\mathbf{g})\mathbf{P}. \tag{36}$$

The estimation error matrix is $\mathbf{G} - \nabla F(\mathbf{W}) = \mathbf{Q}^\top \operatorname{diag}(\mathbf{g} - \nabla f(\mathbf{x}))\mathbf{P}$. Computing the uncentered covariance, we have:

$$\begin{aligned}
&\mathbb{E}\left[(\mathbf{G} - \nabla F(\mathbf{W}))(\mathbf{G} - \nabla F(\mathbf{W}))^\top \mid \mathcal{F}_{t-1}\right] \\
&\quad = \mathbb{E}\left[\mathbf{Q}^\top \operatorname{diag}(\mathbf{g} - \nabla f(\mathbf{x}))\mathbf{P}\mathbf{P}^\top \operatorname{diag}(\mathbf{g} - \nabla f(\mathbf{x}))\mathbf{Q} \mid \mathcal{F}_{t-1}\right] \\
&\quad = \mathbf{Q}^\top \mathbb{E}\left[\operatorname{diag}\left((g_i - \nabla f_i(\mathbf{x}_i))^2\right) \mid \mathcal{F}_{t-1}\right] \mathbf{Q}.
\end{aligned}$$

Here we used $\mathbf{P}\mathbf{P}^\top = \mathbf{I}_m$. Applying the coordinate-wise conditional variance bound in Assumption 2a gives:

$$\mathbb{E}\left[(\mathbf{G} - \nabla F(\mathbf{W}))(\mathbf{G} - \nabla F(\mathbf{W}))^\top \mid \mathcal{F}_{t-1}\right] \preceq \mathbf{Q}^\top \operatorname{diag}(\boldsymbol{\sigma}^2)\mathbf{Q} = \boldsymbol{\Sigma}^2.$$

Conditional unbiasedness, $\mathbb{E}[\mathbf{G}_t^b \mid \mathcal{F}_{t-1}] = \nabla F(\mathbf{W}_t)$, and conditional independence across batch samples are inherited from the vector oracle under this deterministic linear lifting. This completes the proof. □

Lemma 7. Assume that $f(\mathbf{x}) : \mathbb{R}^m \to \mathbb{R}$ satisfies Assumption 3a with L_∞-smoothness parameter L. Then the constructed matrix function $F(\mathbf{W}) : \mathbb{R}^{m\times n} \to \mathbb{R}$ satisfies Assumption 3b with spectral norm smoothness $L_* = L$.

Proof. Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}$ be two arbitrary matrices, and define $\Delta = \mathbf{Y} - \mathbf{X}$. We map these matrices to vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ by defining $\mathbf{x}_i = (\mathbf{Q}\mathbf{X}\mathbf{P}^\top)_{ii}$ and $\mathbf{y}_i = (\mathbf{Q}\mathbf{Y}\mathbf{P}^\top)_{ii}$. Let their difference be $\boldsymbol{\delta} = \mathbf{y} - \mathbf{x}$, where $\delta_i = (\mathbf{Q}\Delta\mathbf{P}^\top)_{ii}$. By definition, $F(\mathbf{X}) = f(\mathbf{x})$ and $F(\mathbf{Y}) = f(\mathbf{y})$.

First, we bridge the Frobenius inner product of the matrices with the Euclidean inner product of the vectors. Let $\mathbf{S} = \operatorname{diag}(\nabla f(\mathbf{x}))$. Using (35) and the cyclic property of the trace operator, we have:

$$\begin{aligned}
\langle \nabla F(\mathbf{X}), \Delta \rangle_F &= \operatorname{tr}\big((\mathbf{Q}^\top \mathbf{S}\mathbf{P})^\top \Delta\big) \\
&= \operatorname{tr}\big(\mathbf{P}^\top \mathbf{S}^\top \mathbf{Q}\Delta\big) \\
&= \operatorname{tr}\big(\mathbf{S}^\top (\mathbf{Q}\Delta\mathbf{P}^\top)\big).
\end{aligned}$$

Let $\mathbf{M} = \mathbf{Q}\Delta\mathbf{P}^\top \in \mathbb{R}^{m\times m}$. Since $\mathbf{S}$ is diagonal, the trace $\operatorname{tr}(\mathbf{S}^\top\mathbf{M})$ perfectly reduces to the element-wise sum of their main diagonals:

$$\operatorname{tr}\big(\mathbf{S}^\top \mathbf{M}\big) = \sum_{i=1}^m \mathbf{S}_{ii}\mathbf{M}_{ii} = \sum_{i=1}^m \nabla f_i(\mathbf{x}_i)\delta_i = \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle .$$

Therefore, the Taylor expansion residuals in both domains are strictly identical:

$$F(\mathbf{Y}) - (F(\mathbf{X}) + \langle \nabla F(\mathbf{X}), \mathbf{Y} - \mathbf{X} \rangle_F) = f(\mathbf{y}) - (f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle) . \tag{37}$$

Taking the absolute value and applying the L_∞-smoothness of $f(\mathbf{x})$, we obtain:

$$|F(\mathbf{Y}) - F(\mathbf{X}) - \langle \nabla F(\mathbf{X}), \mathbf{Y} - \mathbf{X} \rangle_F| \le \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_\infty^2. \tag{38}$$

To bound the vector ℓ_∞-norm using the matrix spectral norm, we examine the maximum component of $\boldsymbol{\delta}$. For any $i \in \{1, \dots, m\}$:

$$|\mathbf{y}_i - \mathbf{x}_i| = \left|(\mathbf{Q}\Delta\mathbf{P}^\top)_{ii}\right| = \left|\mathbf{e}_i^\top \mathbf{Q}\Delta\mathbf{P}^\top \mathbf{e}_i\right| = \left|\mathbf{q}_i^\top \Delta \tilde{\mathbf{e}}_i\right| ,$$

where $\mathbf{e}_i \in \mathbb{R}^m$ is the standard basis vector, $\mathbf{q}_i = \mathbf{Q}^\top \mathbf{e}_i \in \mathbb{R}^m$, and $\tilde{\mathbf{e}}_i = \mathbf{P}^\top \mathbf{e}_i \in \mathbb{R}^n$. Since $\mathbf{Q}$ is orthogonal and $\mathbf{P}$ is a standard projection, both vectors preserve their unit Euclidean lengths: $\|\mathbf{q}_i\|_2 = 1$ and $\|\tilde{\mathbf{e}}_i\|_2 = 1$. By the definition of the induced matrix 2-norm (spectral norm), we have:

$$\left|\mathbf{q}_i^\top \Delta \tilde{\mathbf{e}}_i\right| \le \|\mathbf{q}_i\|_2 \|\Delta\|_{\text{op}} \|\tilde{\mathbf{e}}_i\|_2 = \|\mathbf{Y} - \mathbf{X}\|_{\text{op}}.$$

Taking the maximum over all i yields $\|\mathbf{y} - \mathbf{x}\|_\infty \le \|\mathbf{Y} - \mathbf{X}\|_{\text{op}}$. Substituting this into (38) establishes that $F(\mathbf{W})$ is L_*-spectral norm smooth with $L_* = L$. □

Proof of Theorem 5. We now demonstrate that under this construction, the optimization trajectory of Muon in the matrix space exactly replicates that of SignSGD in the vector space.

Recall the constructed stochastic matrix gradient is $\mathbf{G}_t = \mathbf{Q}^\top \operatorname{diag}(\mathbf{g}_t)\mathbf{P}$. The Muon optimizer updates the weight matrix $\mathbf{W}$ by moving in the direction of the orthogonalized gradient, computed via the thin SVD: $\mathbf{G}_t = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where the update direction is $\operatorname{msign}(\mathbf{G}_t) = \mathbf{U}\mathbf{V}^\top$.

We can explicitly write the SVD of our constructed $\mathbf{G}_t$ by restricting to the active coordinates. Let $S = \{i : g_{t,i} \neq 0\}$ and let $\mathbf{E}_S \in \mathbb{R}^{m\times|S|}$ be the coordinate selector whose columns are the standard basis vectors indexed by S. Define

$$\mathbf{U} = \mathbf{Q}^\top \mathbf{E}_S \operatorname{diag}\big(\operatorname{sign}(\mathbf{g}_{t,S})\big), \quad \mathbf{\Sigma} = \operatorname{diag}\big(|\mathbf{g}_{t,S}|\big), \quad \mathbf{V} = \mathbf{P}^\top \mathbf{E}_S,$$

where $\mathbf{g}_{t,S}$ denotes the subvector of $\mathbf{g}_t$ indexed by S. Since the diagonal entries of $\mathbf{\Sigma}$ are strictly positive, and since $\mathbf{Q}^\top$ is orthogonal while $\mathbf{P}\mathbf{P}^\top = \mathbf{I}_m$, both $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns. Ordering the columns by $|\mathbf{g}_{t,S}|$ yields the thin SVD of $\mathbf{G}_t$, and this ordering does not affect $\mathbf{U}\mathbf{V}^\top$. We have

$$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{Q}^\top \mathbf{E}_S \operatorname{diag}(\operatorname{sign}(\mathbf{g}_{t,S}))\operatorname{diag}(|\mathbf{g}_{t,S}|)\mathbf{E}_S^\top \mathbf{P} = \mathbf{Q}^\top \operatorname{diag}(\mathbf{g}_t)\mathbf{P} = \mathbf{G}_t,$$

so the orthogonal factor is

$$\operatorname{msign}(\mathbf{G}_t) = \mathbf{U}\mathbf{V}^\top = \mathbf{Q}^\top \mathbf{E}_S \operatorname{diag}\big(\operatorname{sign}(\mathbf{g}_{t,S})\big)\mathbf{E}_S^\top \mathbf{P} = \mathbf{Q}^\top \operatorname{diag}(\operatorname{sign}(\mathbf{g}_t))\mathbf{P}.$$

When S is empty, $\mathbf{G}_t = \mathbf{0}$ and $\operatorname{msign}(\mathbf{G}_t) = \mathbf{0}$. The Muon update rule $\mathbf{W}_{t+1} = \mathbf{W}_t - \eta \operatorname{msign}(\mathbf{G}_t)$ then translates to:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta \mathbf{Q}^\top \operatorname{diag}(\operatorname{sign}(\mathbf{g}_t))\mathbf{P}.$$

Left-multiplying by $\mathbf{Q}$ and right-multiplying by $\mathbf{P}^\top$, we obtain:

$$\mathbf{Q}\mathbf{W}_{t+1}\mathbf{P}^\top = \mathbf{Q}\mathbf{W}_t\mathbf{P}^\top - \eta \operatorname{diag}(\operatorname{sign}(\mathbf{g}_t)).$$

By isolating the diagonal entries $\mathbf{x}_{t,i} = (\mathbf{Q}\mathbf{W}_t\mathbf{P}^\top)_{ii}$, we recover the exact SignSGD coordinate-wise update:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \operatorname{sign}(\mathbf{g}_t).$$

This establishes a strict dynamic isomorphism: the sequence of the extracted diagonal entries perfectly replicates the trajectory of SignSGD running on the vector function $f(\mathbf{x})$.

By Lemma 6 and Lemma 7, $F(\mathbf{W})$ satisfies Assumption 2b with noise-scale matrix $\mathbf{\Sigma} = \mathbf{Q}^\top \operatorname{diag}(\boldsymbol{\sigma})\mathbf{Q}$ and Assumption 3b with $L_* = L_\infty$. Furthermore, the trace norm (nuclear norm) of the noise-scale matrix strictly aligns: $\|\mathbf{\Sigma}\|_* = \operatorname{tr}(\mathbf{\Sigma}) = \|\boldsymbol{\sigma}\|_1$.

Finally, observe the true gradient $\nabla F(\mathbf{W}_t) = \mathbf{Q}^\top \operatorname{diag}(\nabla f(\mathbf{x}_t))\mathbf{P}$. Because orthogonal transformations preserve singular values, the singular values of $\nabla F(\mathbf{W}_t)$ are exactly the absolute values of the components of $\nabla f(\mathbf{x}_t)$. Thus, the nuclear norm of the matrix gradient equals the ℓ_1-norm of the vector gradient:

$$\|\nabla F(\mathbf{W}_t)\|_* = \|\nabla f(\mathbf{x}_t)\|_1.$$

Because the optimization trajectories are identical and all bounding metrics perfectly correspond, we can directly invoke the vector lower bound from Theorem 2 under (33) to establish the matrix lower bound for Muon:

$$\mathbb{E}\left[\min_{0\le t<T} \|\nabla F(\mathbf{W}_t)\|_*\right] = \Omega\left(\sqrt{\frac{L_*\Delta}{N}} + \left(\frac{\|\mathbf{\Sigma}\|_*^2 L_*\Delta}{N}\right)^{\frac{1}{4}}\right).$$

This completes the proof. □

E Experimental Details

This section gives the numerical-example figures, real-data performance configurations, and coordinate-noise measurement protocol.

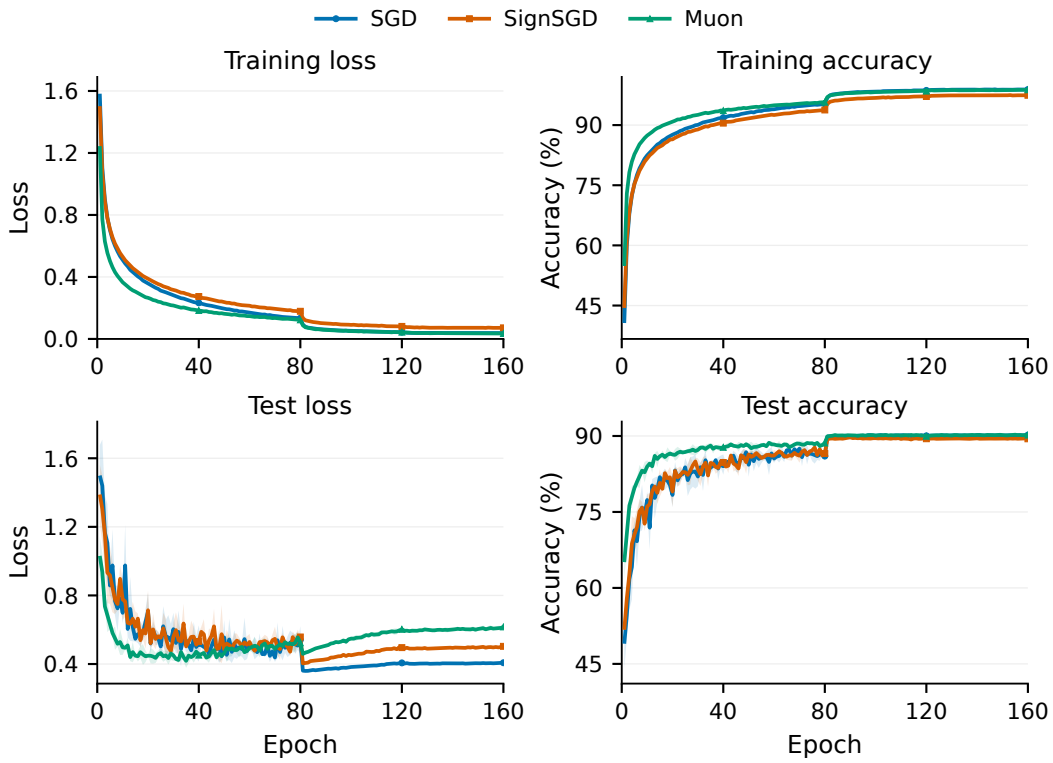


Figure 4: SGD, SignSGD, and Muon on ResNet-20/CIFAR-10. The top row shows training loss and training accuracy, and the bottom row shows test loss and test accuracy. Curves show means over three seeds, and shaded bands indicate ± 1 sample standard deviation.

E.1 CNN and LLM Performance Comparisons

All experiments are conducted in PyTorch on a single NVIDIA Pro 6000 GPU with 96 GB of memory.

For ResNet-20 on CIFAR-10, we follow the experimental setup of Bernstein et al. [2018]. We present the omitted result details in Figure 4.

For nanoGPT and the 130M-parameter LLaMA on C4, we use the codebase of Semenov et al. [2025] at `https://github.com/epfml/llm-optimizer-benchmark/tree/main/scripts`. All three methods use 20000 updates, global batch size 256, and sequence length 512, which gives approximately 2.6 billion training tokens, roughly $1\times$ Chinchilla-optimal token budget for these models [Hoffmann et al., 2022]. The best learning rates of SignSGD and SGD are found via grid search in $\{10^{-2}, 10^{-3}, 10^{-4}\}$. We use a cosine learning rate schedule, with peak learning rates of 0.01 for SGD, 10^{-4} for SignSGD, and 10^{-3} for Muon. Each configuration is trained under three multiple seeds, and we report their mean and sample standard deviation. Detailed curves for training loss, validation loss, and accuracy are presented in Figure 5 and Table 1. We set momentum to zero for each optimizer.

E.2 Direct Coordinate-Noise Measurement

We measure coordinate noise at frozen checkpoints of one run per model–optimizer pair, using the training configurations in Appendix E.1. Along the training trajectory, we store the checkpoints at different stages, and then we use a large number of batches to estimate the true gradient and noise distribution.

At a frozen checkpoint, let $g_i^{(k)}$ denote coordinate i of the kth raw training mini-batch gradient and let $\overline{g}_i = K^{-1}\sum_{k=1}^{K} g_i^{(k)}$. The estimator is

$$\widehat{\sigma}_i^2 = \frac{1}{K-1}\sum_{k=1}^{K}(g_i^{(k)} - \overline{g}_i)^2, \qquad \widehat{\sigma}_i = \sqrt{\widehat{\sigma}_i^2}. \tag{39}$$

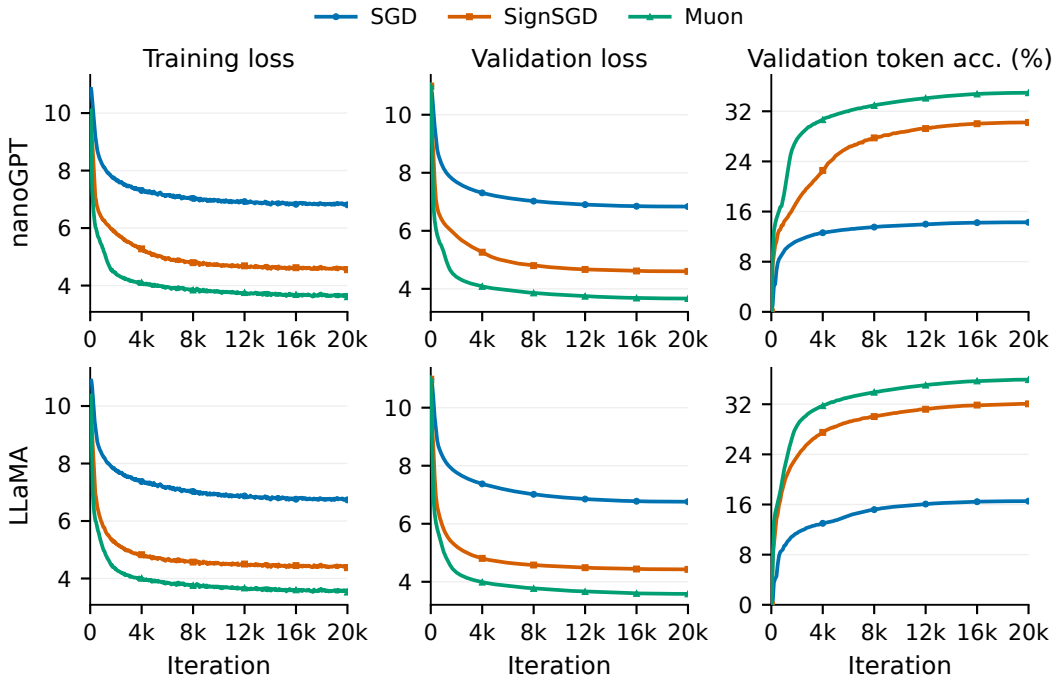


Figure 5: SGD, SignSGD, and Muon on C4. The top row is nanoGPT, and the bottom row is LLaMA. The columns show training loss, validation loss, and validation token accuracy. Curves show means over three runs.

Table 1: Means over three seeds, with sample standard deviations in small gray subscripts. Held-out loss and accuracy use the test set for ResNet-20 and the validation set for the language models.

Model	Optimizer	Training loss	Test/validation loss	Accuracy (%)
ResNet-20	SGD	$0.0381_{0.0018}$	$0.407_{0.006}$	$90.19_{0.24}$
	SignSGD	$0.0716_{0.0006}$	$0.503_{0.008}$	$89.44_{0.09}$
	Muon	$0.0350_{0.0021}$	$0.616_{0.015}$	$90.18_{0.14}$
nanoGPT	SGD	$6.809_{0.036}$	$6.835_{0.016}$	$14.29_{0.09}$
	SignSGD	$4.551_{0.013}$	$4.603_{0.014}$	$30.22_{0.05}$
	Muon	$3.611_{0.028}$	$3.666_{0.005}$	$34.96_{0.07}$
LLaMA	SGD	$6.735_{0.046}$	$6.759_{0.012}$	$16.54_{0.11}$
	SignSGD	$4.375_{0.014}$	$4.428_{0.006}$	$32.06_{0.10}$
	Muon	$3.528_{0.028}$	$3.583_{0.005}$	$35.92_{0.07}$

Each checkpoint estimate uses $K = 10000$ raw training mini-batch gradients.

We use the *empirical* counterpart of the vector concentration coefficient defined in (5),

$$\widehat{\chi} = \frac{d \sum_{i=1}^{d} \widehat{\sigma}_i^2}{(\sum_{i=1}^{d} \widehat{\sigma}_i)^2}. \tag{40}$$

All trainable coordinates are included, and $\widehat{\chi}$ is computed directly from their estimated standard deviations. We present estimated $\widehat{\chi}$ across different optimizers and models in Figure 6. We observe that under SGD, nanoGPT and LLaMA exhibit larger $\widehat{\chi}$ than ResNet-20 at all five measured stages, with LLaMA consistently the highest. Also, within each language model, SGD yields larger $\widehat{\chi}$ than SignSGD and Muon at every measured checkpoint after initialization.

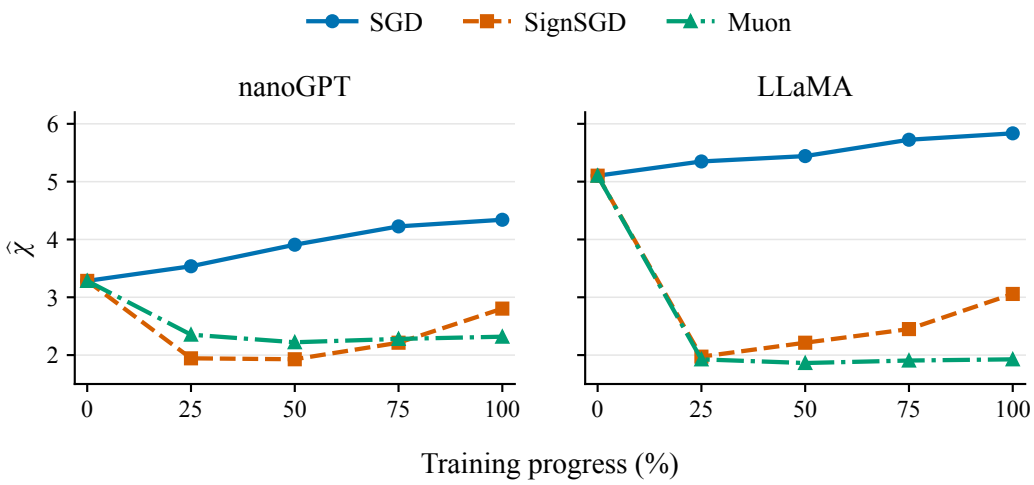


Figure 6: Coordinate-noise concentration $\widehat{\chi}$ during SGD, SignSGD, and Muon training on nanoGPT and LLaMA. Markers show measurements at five frozen checkpoints.

For the distribution plot presented in Figure 3, define $x_i = \widehat{\sigma}_i/\overline{\widehat{\sigma}}$, where $\overline{\widehat{\sigma}} = d^{-1}\sum_i \widehat{\sigma}_i$, and $z_i = \log_{10} x_i$. The curve is a smoothed density $\widehat{p}(z)$ whose area satisfies $\int_a^b \widehat{p}(z)\,dz \approx d^{-1}|\{i : a \le z_i \le b\}|$, the fraction of coordinates in the log interval $[a, b]$.

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: We explicitly state our theoretical contributions and empirical findings regarding optimization under sparse noise in the abstract and introduction. These claims are strictly supported by the rigorous mathematical proofs and experimental evaluations presented in the subsequent sections.

   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: The theoretical results rely on the assumptions explicitly stated in Sections 3.1 and 5.1. Section 5.1 specifies that the Muon analysis considers pure polar updates without Newton–Schulz iterations or momentum.

   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: The paper provides assumptions and proofs for each theoretical result.

   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 paper discloses the information necessary to reproduce the main experimental results in Appendix E.

   Guidelines:

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

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: We provide the code and instructions in the supplementary material. Please follow the instructions in `README.md` to reproduce our results. Our code is also open-sourced at https://github.com/Dingzhen230/SignSGD_Outperforms_SGD.

   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: The setup of our experiments is clearly introduced in Appendix E.

   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: Our real-data experiments report means and sample standard deviations over three runs in Appendix E.1.

   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: [No]

   Justification: We report the hardware and training configurations in Appendix E.1 and release the experimental code. Execution-time records are unavailable, and per-experiment runtimes and total compute are not fully reported.

   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 research conducted in the paper conforms with the NeurIPS Code of Ethics.

   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: This is primarily a theoretical paper with no potential negative social impact.

    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: This is primarily a theoretical paper with 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: [Yes]

Justification: We cite the creators of the dataset used in our experiments, and we explicitly respect the license and terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [N/A]

Justification: This work does not release new assets.

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.