Sign-Based Optimizers Under Heavy-Tailed Noise: On the Effectiveness of Lion and Muon

Dingzhi Yu[1,2] **Hongyi Tao**[1,2] **Yuanyu Wan**[3] **Lijun Zhang**[1,2,✉]
[1]State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210023, China
[2]School of Artificial Intelligence, Nanjing University, Nanjing 210023, China
[3]School of Software Technology, Zhejiang University, Ningbo 315100, China
{yudz,taohy,zhanglj}@lamda.nju.edu.cn, wanyy@zju.edu.cn
✉ Corresponding Author

Abstract

While adaptive gradient methods are the workhorse of modern machine learning, sign-based optimization algorithms such as Lion and Muon have recently demonstrated superior empirical performance over AdamW in training large language models (LLMs). However, a theoretical understanding of why sign-based updates outperform variance-adapted methods remains elusive. In this paper, we aim to bridge the gap between theory and practice through the lens of heavy-tailed gradient noise, a phenomenon frequently observed in language modeling tasks. Theoretically, we introduce a novel generalized heavy-tailed noise condition that captures the behavior of LLMs more accurately than standard finite variance assumptions. Under this noise model, we establish sharp convergence rates of SignSGD and Lion for generalized smooth function classes, matching or surpassing previous best-known bounds. Furthermore, we extend our analysis to Muon and Muonlight, providing what is, to our knowledge, the first rigorous analysis of matrix optimization under heavy-tailed stochasticity. These results offer a strong theoretical justification for the empirical superiority of sign-based optimizers, showcasing that they are naturally suited to handle the noisy gradients associated with heavy tails. Empirically, LLM pretraining experiments validate our theoretical insights and confirm that our proposed noise models are well-aligned with practice.

1 Introduction

The success of large foundation models [1–7] requires costly pretraining procedures that are supported by stochastic optimization algorithms [8]. Among them, adaptive gradient methods [9, 10] prevail, in which AdamW [11] is the *de facto* optimizer for training large language models (LLM). Recently, two representative sign-based optimizers, namely Lion [12] and Muon [13] have received significant attention due to their consistent speedup over AdamW [14–17], especially in scaling up both LLM pretraining [18–22] and post-training [23]. Despite their huge empirical success, we still lack a complete theoretical understanding of why Lion and Muon outperform AdamW for training LLMs.

In this work, we systematically investigate this phenomenon with theoretical analysis and empirical study. The significant empirical evidence that initiates our work comes from the widely observed heavy-tailed distribution of the stochastic gradients (see also our measurements in Figure 1) [24–27], especially in language modeling tasks [28, 29]. In sharp contrast to the finite variance condition commonly employed in theoretical studies [30, 31], such a regime only assumes the stochastic noise has a finite pth moment where $p \in (1, 2]$. Under this more realistic noise model, the challenges underlying both theoretical analysis and algorithmic design arise. For instance, SGD is known to diverge in the expected squared gradient norm $\mathbb{E}\|\nabla f(\mathbf{x}_t)\|_2^2$ on heavy-tailed examples with $p < 2$ [25, 32], so the standard second-moment convergence analysis no longer applies. As for the more popular adaptive methods like Adam [10] and AdaGrad [9], *no rigorous convergence theory has yet been*

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

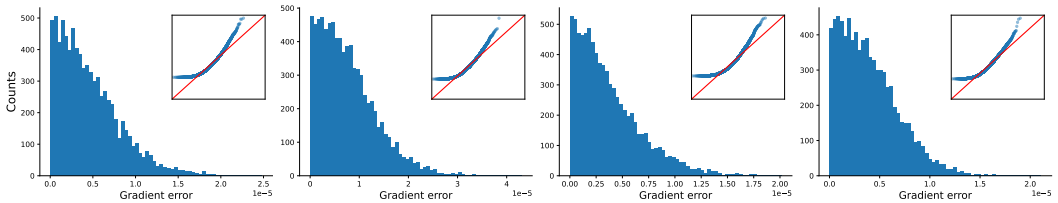


Figure 1: Heavy-tailed gradient noise observed at initialization in GPT-2 pretraining on C4. The histograms show gradient-error distributions at randomly sampled coordinates, while the inset Q-Q plots compare them against a Gaussian reference. The pronounced tail deviations provide empirical evidence of heavy-tailed stochastic gradients.

established under heavy-tailed noise. Indeed, whether vanilla AdamW admits a sharp convergence guarantee under finite-p heavy-tailed noise has recently been formalized as an open problem [33]. To make things worse, ***Chezhegov et al. [34] show that Adam and AdaGrad could suffer from worse convergence when the noise is heavy-tailed.*** Intriguingly, the empirical advantage of sign-based methods like Lion and Muon is most pronounced in LLM training regimes [18, 16], where heavy-tailed noise is prevalent. In contrast, this advantage diminishes when it comes to computer vision tasks [35, 36], which typically exhibit more concentrated gradient noise [25, 37, 26, 29]. Regarding this disparity, it is natural to ask the following question:

> Can we obtain theoretical guarantees of sign-based optimizers to explain their practical effectiveness and justify their advantage over AdamW under heavy-tailed gradient noise?

We answer the above question affirmatively by establishing new convergence theories for a series of representative vector- and matrix-based sign descent methods, including vector sign optimizer SignSGD [38] and Lion [12], matrix sign optimizer Muon [13] and Muonlight [18]. For a heavy-tailed index p and an iteration number T, we show that these optimizers converge at $O(T^{-\frac{p-1}{3p-2}})$, matching or surpassing the previous state-of-the-art. Our convergence analysis is derived under a newly proposed generalized heavy-tailed noise condition, allowing the noise level to grow w.r.t. the gradient norm, which is strictly weaker than existing assumptions and is further supported by empirical evidence from LLM pretraining. Furthermore, our theory holds for generalized smooth function classes [39], which have become a standard deployment in theory [40–44] and proven to align well with practice [45]. Our results suggest that the sign operation, whether applied element-wise or via orthogonalization in the matrix case, acts as a natural robustifier against heavy-tailed stochasticity, thus making sign-based optimizers like Lion and Muon a natural fit for training LLMs.

Our main contributions are summarized as follows.

- We propose a generalized heavy-tailed noise framework for both vector and matrix variables and derive sharp convergence rates of $O(T^{-\frac{p-1}{3p-2}})$ for SignSGD, Lion, Muon, and Muonlight.
- We characterize the mechanism by which the sign operator inherently handles heavy-tailed noise, providing a theoretical justification for their superiority over SGD or AdamW in LLMs.
- We conduct GPT-2 pretraining experiments to validate the proposed noise model and assess robustness and stability. Curvature and noise measurements over the full model, together with matrix effective ranks, reveal geometry that favors our bounds over prior Euclidean and Frobenius analyses.

Due to space limitations, the comprehensive literature review can be found in Appendix A.

2 Vector Sign Optimizer: SignSGD and Lion

In this section, we first introduce some notations and assumptions, and then present the convergence guarantee. Proofs of the main theorems are deferred to Appendix D.

2.1 Notations and Assumptions

We write $[T]$ for $\{1, 2, \ldots, T\}$, $|\mathbf{x}|$ for element-wise absolute value of $\mathbf{x} \in \mathbb{R}^d$, $\odot$ for element-wise product between two vectors. The $\mathbf{l}$-weighted vector norm for $\mathbf{l} \in \mathbb{R}^d_+$ is defined as $\|\mathbf{x}\|_{\mathbf{l}}^2 :=$

$\mathbf{x}^\top \operatorname{diag}(\mathbf{l})\,\mathbf{x}$. For vector descent methods, we study the optimization problem $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$, where $f:\mathbb{R}^d\mapsto\mathbb{R}$ 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, as later specified in Assumption 3a. Below, we list some necessary assumptions.

Assumption 1a (Lower bounded objective)**.** *The objective function f is possibly non-convex, almost everywhere twice-differentiable, and bounded from below: $f^*:=\inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})>-\infty$.*

Assumption 2a ($(\mathbf{l}_0,\mathbf{l}_1)$-vector smoothness)**.** *There exists non-negative vectors $\mathbf{l}_0=[\mathbf{l}_{0,1},\cdots,\mathbf{l}_{0,d}]\in\mathbb{R}^d_+$ and $\mathbf{l}_1=[\mathbf{l}_{1,1},\cdots,\mathbf{l}_{1,d}]\in\mathbb{R}^d_+$ such that for all $\|\mathbf{x}'-\mathbf{x}\|_\infty\le 1/\|\mathbf{l}_1\|_\infty$, it holds that*

$$|f(\mathbf{x}')-(f(\mathbf{x})+\langle\nabla f(\mathbf{x}),\mathbf{x}'-\mathbf{x}\rangle)|\le\frac{1}{2}\|\mathbf{x}'-\mathbf{x}\|^2_{\mathbf{l}_0+\mathbf{l}_1\odot|\nabla f(\mathbf{x})|}. \tag{1}$$

Assumption 1a is standard and necessary for stochastic non-convex optimization [46]. We further introduce twice-differentiability as a regularity condition coupled with our smoothness model (see Lemma D.4 for details). Assumption 2a is new, which characterizes a class of generalized smooth functions originally proposed in Zhang et al. [39]. Note that when $\mathbf{l}_1=\mathbf{0}$, Assumption 2a degenerates to the *coordinate-wise smoothness* condition widely employed in the pioneering works of SignSGD [38, 47, 48], which naturally aligns with sign descent methods due to their coordinate-wise nature [49, 50]. When $\mathbf{l}_1\succ\mathbf{0}$, Assumption 2a can be viewed as an extension of the *coordinate-wise smoothness* assumption, which aligns closely with practice [51–53]. In general, Assumption 2a is not only weaker than previous assumptions, but also a better fit for sign gradient descent. Detailed discussions are postponed to Appendix C.1.

Assumption 3a (Unbiasedness)**.** *At step t we observe a mini-batch of mutually independent gradients $G_t=\{\mathbf{g}_t^1,\cdots,\mathbf{g}_t^B\}$ satisfying $\mathbb{E}\left[\mathbf{g}_t^b|\mathcal{F}_{t-1}\right]=\nabla f(\mathbf{x}_t),\forall b\in[B]$ where $\mathcal{F}_t=\sigma(G_1,\ldots,G_t)$ denotes the natural filtration.*

Assumption 4a ($(\boldsymbol{\sigma}_0,\boldsymbol{\sigma}_1)$-vector heavy-tailed noise)**.** *There exists $p\in(1,2]$, $\boldsymbol{\sigma}_0=[\boldsymbol{\sigma}_{0,1},\cdots,\boldsymbol{\sigma}_{0,d}]\in\mathbb{R}^d_+$ and $\boldsymbol{\sigma}_1=[\boldsymbol{\sigma}_{1,1},\cdots,\boldsymbol{\sigma}_{1,d}]\in\mathbb{R}^d_+$ such that*

$$\mathbb{E}\left[\left|\mathbf{g}_{t,i}^b-\nabla_i f(\mathbf{x}_t)\right|^p\middle|\mathcal{F}_{t-1}\right]\le\boldsymbol{\sigma}_{0,i}^p+\boldsymbol{\sigma}_{1,i}^p\left|\nabla_i f(\mathbf{x}_t)\right|^p,\quad\forall i\in[d],\forall b\in[B].$$

Assumption 4a is new, which generalizes the classic *coordinate-wise* heavy-tailed noise assumption [54, Assumption 3] where $\boldsymbol{\sigma}_1=\mathbf{0}$. Assumption 4a is meaningful in two aspects. Theoretically, we demonstrate that the classical heavy-tailed noise model ($\boldsymbol{\sigma}_1=\mathbf{0}$) might fail even on linear regression tasks (see Appendix C.3). Empirically, we draw significant evidence from LLM pretraining to illustrate that Assumption 4a aligns well with practice (see Section 4).

2.2 Convergence Theory for SignSGD

We formally introduce the SignSGD optimizer [38] in Algorithm 1.[1] At step t, Algorithm 1 maintains the momentum buffer $\mathbf{m}_t$ using the batched stochastic gradient $\mathbf{g}_t$. Then, the sign of the momentum $\mathbf{m}_t$ is utilized to update the current model $\mathbf{x}_t$. Below, we present the convergence guarantee of Algorithm 1, showcasing that SignSGD remains effective under heavy-tailed gradient noise.

Theorem 1. *Under Assumptions 1a to 4a, define $\Delta_f:=f(\mathbf{x}_1)-f^*$. By setting*

$$\begin{aligned}&B=\max\left\{1,\left\lceil\left(64\sqrt{2}\,\|\boldsymbol{\sigma}_1\|_\infty\right)^{\frac{p}{p-1}}\right\rceil\right\},\quad\beta=\max\left\{0,1-B^{\frac{2p-2}{3p-2}}\left(\frac{\Delta_f\,\|\mathbf{l}_0\|_1}{\|\boldsymbol{\sigma}_0\|_1^2\,T}\right)^{\frac{p}{3p-2}}\right\},\\&\eta=\min\left\{\sqrt{\frac{2\Delta_f(1-\beta)}{9\,\|\mathbf{l}_0\|_1\,T}},\frac{1-\beta}{32\,\|\mathbf{l}_1\|_\infty}\right\},\end{aligned}\tag{2}$$

Algorithm 1 ensures

$$\frac{1}{T}\sum_{t=1}^T\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|_1\right]\le O\left(\left(\Delta_f\,\|\mathbf{l}_0\|_1\right)^{\frac{p-1}{3p-2}}\|\boldsymbol{\sigma}_0\|_1^{\frac{p}{3p-2}}(BT)^{-\frac{p-1}{3p-2}}\right).$$

[1] Algorithm 1 is referred to as Signum, signSGD-SIM, or SMM in previous work [38, 55, 56]. For simplicity, we keep the name SignSGD hereafter.

Algorithm 1 SignSGD [38]

Input: $T \in \mathbb{N}$, $\mathbf{x}_1 \in \mathbb{R}^d$, $\eta \in \mathbb{R}_+$
for $t = 1$ **to** T **do**
  $\mathbf{g}_t = \frac{1}{B}\sum_{b=1}^{B} \mathbf{g}_t^b$
  $\mathbf{m}_t = \beta \mathbf{m}_{t-1} + (1-\beta)\mathbf{g}_t \quad \{\mathbf{m}_0 := \mathbf{g}_1\}$
  $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \operatorname{sign}(\mathbf{m}_t)$
end for

Algorithm 2 Lion [12]

Input: $T \in \mathbb{N}$, $\mathbf{x}_1 \in \mathbb{R}^d$, $\eta \in \mathbb{R}_+$, $\lambda \in \mathbb{R}_+$
for $t = 1$ **to** T **do**
  $\mathbf{g}_t = \frac{1}{B}\sum_{b=1}^{B} \mathbf{g}_t^b$
  $\mathbf{v}_t = \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t \ \{\mathbf{m}_0 := \mathbf{g}_1\}$
  $\mathbf{m}_t = \beta_2 \mathbf{m}_{t-1} + (1-\beta_2)\mathbf{g}_t$
  $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \operatorname{sign}(\mathbf{v}_t) - \eta\lambda \mathbf{x}_t$
end for

The convergence rate in Theorem 1 implies a complexity of $O(\epsilon^{-\frac{3p-2}{p-1}})$ to reach an ϵ-stationary point, i.e., $\mathbb{E}\left[\|\nabla f(\mathbf{x})\|_1\right] \le \epsilon$. Since $\|\nabla f(\mathbf{x})\|_1 \ge \|\nabla f(\mathbf{x})\|_2$, this complexity matches the ℓ_2-norm lower bound in Liu and Zhou [57, Theorem 3.3] and also recovers the well-known $O(\epsilon^{-4})$ lower bound [46] when $p = 2$. Unlike previous works [56, 54], our result does not depend on dimension d explicitly and holds under any $\mathbf{l}_1 \ge \mathbf{0}, \boldsymbol{\sigma}_1 \ge \mathbf{0}$ and $p \in (1, 2]$, suggesting that our theory may capture a wider range of objectives. Compared to the well-known gradient normalization or gradient clipping techniques which effectively cope with heavy-tailed noise [57, 58], our results reveal that the sign operator not only remains robust to heavy-tailed stochasticity but also has an advantage over gradient normalization. Due to space limitations, the comprehensive illustrations can be found in Appendix B.

Finite-variance lower bound. For $p = 2$ and $\mathbf{l}_1 = \boldsymbol{\sigma}_1 = \mathbf{0}$, Tao et al. [59, Theorem 2] establish an ℓ_1-stationarity lower bound for SignSGD (without momentum). Its stochastic term is

$$\Omega\Big((\Delta_f \|\mathbf{l}_0\|_1)^{1/4} \|\boldsymbol{\sigma}_0\|_1^{1/2} N^{-1/4}\Big), \qquad N = BT,$$

which matches the dependence on curvature, noise, and oracle complexity in the $p = 2$ specialization of Theorem 1. This gives a tightness benchmark in the finite-variance, no-momentum subclass; the lower bound does not cover momentum or the generalized multiplicative terms. The assumption mapping and full lower bounds are given in Appendix B.1.

Remark 1. *When $\boldsymbol{\sigma}_1 = \mathbf{0}$, Kornilov et al. [54] analyze Algorithm 1 and derive a complexity of $O(\Delta_f l_0 d^{\frac{3p-2}{2(p-1)}} \|\boldsymbol{\sigma}_0\|_p^{\frac{p}{p-1}} \epsilon^{-\frac{3p-2}{p-1}})$ under the classic l_0-smoothness. Compare their result to ours:*

$$\textit{Kornilov et al. [54]} : O\left(\frac{\Delta_f \|\mathbf{l}_0\|_1 \|\boldsymbol{\sigma}_0\|_1^{\frac{p}{p-1}} d^{\frac{3p-2}{2(p-1)}}}{\epsilon^{\frac{3p-2}{p-1}} \phi_l \phi_\sigma^{\frac{p}{p-1}}}\right) \textit{ vs Ours} : O\left(\frac{\Delta_f \|\mathbf{l}_0\|_1 \|\boldsymbol{\sigma}_0\|_1^{\frac{p}{p-1}}}{\epsilon^{\frac{3p-2}{p-1}}}\right),$$

where $\phi_l := \|\mathbf{l}_0\|_1 / l_0 \in [1, d]$ and $\phi_\sigma := \|\boldsymbol{\sigma}_0\|_1 / \|\boldsymbol{\sigma}_0\|_p \in [1, d^{1-\frac{1}{p}}]$ are two quantities reflecting the problem geometry. When the curvature vector $\mathbf{l}_0$ is sparse and exhibits axis-alignment phenomenon [50], ϕ_l becomes close to 1. If $\boldsymbol{\sigma}_0$ is also sparse [60, 61], then it holds $\phi_\sigma = O(1)$. In this case, Theorem 1 achieve a remarkable complexity improvement of $d^{\frac{3p-2}{2(p-1)}}$. On the other hand, taking the maximum over ϕ_l, ϕ_σ gives the best case *complexity of Kornilov et al. [54], which still suffers from an extra $d^{\frac{2-p}{2(p-1)}}$ factor compared to ours.*

2.3 Convergence Theory for Lion

The Lion optimizer is discovered by Chen et al. [12] through program search. On top of SignSGD, it incorporates (i) different β values to track the exponential moving average of past gradients $\{\mathbf{g}_t\}_{t\in[T]}$ thus decoupling the state between stored momentum $\mathbf{m}_t$ and the update momentum $\mathbf{v}_t$, and (ii) decoupled weight decay into the final update. The whole procedure is summarized in Algorithm 2. The following theorem prescribes the convergence rate of Lion.

Theorem 2. *Under Assumptions 1a to 4a, define $\Delta_f := f(\mathbf{x}_1) - f^*$. Set*

$$B = \max\left\{1, \left\lceil \left(144\sqrt{2}\,\|\boldsymbol{\sigma}_1\|_\infty\right)^{\frac{p}{p-1}} \right\rceil\right\}, \quad \beta_2 = \max\left\{0, 1 - B^{\frac{2p-2}{3p-2}} \left(\frac{\Delta_f \|\mathbf{l}_0\|_1}{\|\boldsymbol{\sigma}_0\|_1^2 T}\right)^{\frac{p}{3p-2}}\right\},$$
$$\eta = \min\left\{\sqrt{\frac{8\Delta_f(1-\beta_2)}{33\,\|\mathbf{l}_0\|_1 T}}, \frac{1-\beta_2}{120\,\|\mathbf{l}_1\|_\infty}\right\}, \beta_1 \in \left[1 - (1-\beta_2)^{\frac{p-1}{p}}, 1\right], \lambda \in \left[0, \frac{1-2^{-\frac{1}{T}}}{\eta}\right]. \tag{3}$$

If $\|\mathbf{x}_1\|_\infty \le 1/(3\lambda)$*, then Algorithm 2 ensures* $\|\mathbf{x}_t\|_\infty \le 2/(3\lambda), \forall t \in [T]$*, and*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|_1\right] \le O\left(\left(\Delta_f \|\mathbf{l}_0\|_1\right)^{\frac{p-1}{3p-2}} \|\boldsymbol{\sigma}_0\|_1^{\frac{p}{3p-2}} (BT)^{-\frac{p-1}{3p-2}}\right).$$

The above convergence rate matches that of Theorem 1, since Lion reduces to SignSGD when $\beta_1 = \beta_2$. As far as we know, Theorem 2 is the first rigorous theoretical guarantee for Lion under heavy-tailed noise. The implied $O(\epsilon^{-\frac{3p-2}{p-1}})$ complexity, as aforementioned SignSGD, is also optimal.

Comparison with Jiang et al. [62]. A direct comparison with Jiang et al. [62] requires aligning the noise budget. Consider $p = 2$, $\mathbf{l}_0 = L\mathbf{1}$, $\boldsymbol{\sigma}_0 = \sigma\mathbf{1}$, and $\mathbf{l}_1 = \boldsymbol{\sigma}_1 = \mathbf{0}$. Our theorem gives

$$O\Big((\Delta_f dL)^{1/4}(d\sigma)^{1/2}T^{-1/4}\Big) = O\Big(d^{3/4}(\Delta_f L)^{1/4}\sigma^{1/2}T^{-1/4}\Big).$$

Their Euclidean variance parameter satisfies $\mathbb{E}\|\nabla f(\mathbf{x};\xi) - \nabla f(\mathbf{x})\|_2^2 \le \sigma_{[62]}^2$. Under the same coordinate-wise noise model, the matching budget is $\sigma_{[62]}^2 = d\sigma^2$, so their leading bound becomes

$$O\Big(d^{1/2}(\Delta_f L)^{1/4}\sigma_{[62]}^{1/2}T^{-1/4}\Big) = O\Big(d^{3/4}(\Delta_f L)^{1/4}\sigma^{1/2}T^{-1/4}\Big).$$

Thus the isotropic worst-case scaling coincides. More generally, the Euclidean conversion pays $\sqrt{d}\|\boldsymbol{\sigma}_0\|_2$ and $d\|\mathbf{l}_0\|_\infty$, whereas our bound depends directly on

$$\|\boldsymbol{\sigma}_0\|_1 \le \sqrt{d}\,\|\boldsymbol{\sigma}_0\|_2, \qquad \|\mathbf{l}_0\|_1 \le d\,\|\mathbf{l}_0\|_\infty.$$

Hence our geometry-adaptive bound is never worse in its smoothness/noise dependence under a common parametrization, and is strictly sharper for sparse noise or axis-aligned curvature. Section 4.2 measures the favorable geometry during GPT-2 pretraining.

Our analysis also allows more flexible momentum and weight-decay parameters. In particular, we relax their $\beta_1 \in [1-\sqrt{1-\beta_2}, 1-(1-\beta_2)^2]$ requirement as shown in Equation (3). By a finer analysis of the weight decay dynamics in Lemma D.8, we relax their $\lambda \le 1/(2\eta T)$ requirement to the milder $\lambda \le (1-2^{-1/T})/\eta$. Our initialization condition $\|\mathbf{x}_1\|_\infty \le 1/(3\lambda)$ is more general than their $\|\mathbf{x}_1\|_\infty \le \eta$, as a large λ ($\sim 10^{-1}$) and a small η ($\sim 10^{-5}$) is typically chosen for Lion [16]. Furthermore, we establish a uniform constant bound on the iterate trajectory $\|\mathbf{x}_t\|_\infty \le 2/(3\lambda)$, which is significantly tighter than the $\|\mathbf{x}_t\|_\infty \le \eta t$ bound in Jiang et al. [62] that grows over time. Our result also aligns with the bounded constraints identified for Lion-$\mathcal{K}$ dynamics [63].

3 Matrix Sign Optimizer: Muon and Muonlight

The organization of this section follows from Section 2. Proofs can be found at Appendix E.

3.1 Notations and Assumptions

We denote the set of $m \times m$ positive semi-definite (PSD) matrices by $\mathbb{S}^m$. For any $\mathbf{X} \in \mathbb{R}^{m\times n}$, its modulus (or absolute value) is defined as $|\mathbf{X}|_{\mathrm{m}} := \left(\mathbf{X}\mathbf{X}^\top\right)^{1/2} \in \mathbb{S}^m$. The matrix sign operator is defined as $\operatorname{msign}(\mathbf{X}) := \mathbf{U}\mathbf{V}^\top$, where $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ is the singular value decomposition (SVD) of $\mathbf{X} \in \mathbb{R}^{m\times n}$. 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. Finally, for any $\mathbf{L} \in \mathbb{S}^m$, the $\mathbf{L}$-weighted matrix norm is defined as $\|\mathbf{X}\|_{\mathbf{L}}^2 := \operatorname{tr}\left(\mathbf{X}^\top\mathbf{L}\mathbf{X}\right)$. For Muon and Muonlight, we consider the matrix optimization problem $\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} f(\mathbf{X})$ and list some necessary assumptions in the sequel.

Assumption 1b (Lower bounded objective)**.** *The objective function f is possibly non-convex, and bounded from below: $f^* := \inf_{\mathbf{X}\in\mathbb{R}^{m\times n}} f(\mathbf{X}) > -\infty$.*

Assumption 2b ($(\mathbf{L}_0, \mathbf{L}_1)$-matrix smoothness)**.** *There exists matrices $\mathbf{L}_0, \mathbf{L}_1 \in \mathbb{R}^{m\times n}$ where $\mathbf{L}_0$ has full row rank, such that for all $\|\mathbf{X}' - \mathbf{X}\|_{\mathrm{op}} \le 1/\|\mathbf{L}_1\|_{\mathrm{op}}$, it holds that*

$$\|\nabla f(\mathbf{X}') - \nabla f(\mathbf{X})\|_{(\mathbf{L}(\mathbf{X}))^{-1}} \le \|\mathbf{X}' - \mathbf{X}\|_{\mathbf{L}(\mathbf{X})}, \text{ where } \mathbf{L}(\mathbf{X}) = |\mathbf{L}_0|_{\mathrm{m}} + \left|\nabla f(\mathbf{X})\mathbf{L}_1^\top\right|_{\mathrm{m}}.$$

Algorithm 3 Muon [13]

Input: $T \in \mathbb{N}$, $\mathbf{X}_1 \in \mathbb{R}^{m\times n}$, $\eta \in \mathbb{R}_+$
for $t = 1$ **to** T **do**
  $\mathbf{G}_t = \frac{1}{B}\sum_{b=1}^{B} \mathbf{G}_t^b$
  $\mathbf{B}_t = \beta \mathbf{B}_{t-1} + \mathbf{G}_t \qquad \{\mathbf{B}_0 := \mathbf{0}\}$
  $\mathbf{O}_t = \texttt{NewtonSchulz}(\mathbf{B}_t)$
  $\mathbf{X}_{t+1} = \mathbf{X}_t - \eta \mathbf{O}_t$
end for

Algorithm 4 Muonlight [18]

Input: $T \in \mathbb{N}$, $\mathbf{X}_1 \in \mathbb{R}^{m\times n}$, $\eta, \lambda \in \mathbb{R}_+$
for $t = 1$ **to** T **do**
  $\mathbf{G}_t = \frac{1}{B}\sum_{b=1}^{B} \mathbf{G}_t^b$
  $\mathbf{B}_t = \beta_2 \mathbf{B}_{t-1} + \mathbf{G}_t \qquad \{\mathbf{B}_0 := \mathbf{0}\}$
  $\widetilde{\mathbf{B}}_t = \beta_1 \mathbf{B}_t + \mathbf{G}_t$
  $\mathbf{O}_t = \texttt{NewtonSchulz}(\widetilde{\mathbf{B}}_t)$
  $\mathbf{X}_{t+1} = \mathbf{X}_t - \eta \mathbf{O}_t - \eta\lambda \mathbf{X}_t$
end for

Assumption 2b serves as the matrix-based counterpart to Assumption 2a and represents, to the best of our knowledge, the *first* formulation of a **matrix generalized smoothness model**. Furthermore, when $\mathbf{L}_1 = \mathbf{0}$, this framework encompasses classical matrix L-smoothness, as defined in An et al. [64, Assumption 2] and Kovalev and Borodich [65, Assumption 1]. As pointed out by An et al. [64], this formulation effectively models the block-diagonal Hessian structure of transformer-based modern neural networks [66, 67].

Assumption 3b (Unbiasedness)**.** *At step t we observe a mini-batch of mutually independent gradients $G_t = \{\mathbf{G}_t^1, \cdots, \mathbf{G}_t^B\}$ satisfying $\mathbb{E}\left[\mathbf{G}_t^b | \mathcal{F}_{t-1}\right] = \nabla f(\mathbf{X}_t), \forall b \in [B]$ where $\mathcal{F}_t = \sigma(G_1, \ldots, G_t)$ denotes the natural filtration.*

Assumption 4b ($(\mathbf{V}_0, \mathbf{V}_1)$-matrix heavy-tailed noise)**.** *There exists $p \in (1, 2]$, $\mathbf{V}_0, \mathbf{V}_1 \in \mathbb{R}^{m\times n}$ where $\mathbf{V}_0$ has full row rank, such that*

$$\mathbb{E}\left[\|\mathbf{V}_0\|_*^{p/2} \cdot \left\|\mathbf{G}_t^b - \nabla f(\mathbf{X}_t)\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}^{p} \middle| \mathcal{F}_{t-1}\right] \le \|\mathbf{V}_0\|_*^p + |\langle \mathbf{V}_1, \nabla f(\mathbf{X}_t)\rangle|^p, \quad \forall b \in [B].$$

Assumption 4b introduces our newly proposed matrix-based generalized heavy-tailed noise model, which generalizes canonical assumptions in several key dimensions (see Appendix C.4 for details). First, when $\mathbf{V}_1 = \mathbf{0}$ and $p = 2$, our condition recovers the adaptive variance assumption $\mathbb{E}\left[\left\|\mathbf{G}_t^b - \nabla f(\mathbf{X}_t)\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}^{2} \middle| \mathcal{F}_{t-1}\right] \le \|\mathbf{V}_0\|_*$ in Kovalev [68, Assumption 3], Kovalev and Borodich [65, Assumption 2], Xie et al. [69, Definition 4.1] and Tao et al. [59, Assumption 2b]. Similar to coordinate-wise bounds in vector settings [38, 52], this formulation and its variants in An et al. [64, Assumption 3] and Pan et al. [70, Assumption 3] better capture the structural properties of the noise matrix. Second, by extending the tail-index to $p \in (1, 2]$, we provide a natural characterization of the heavy-tailed regime. Finally, Assumption 4b allows the pth moment of the noise to grow with the gradient, extending recent vector-based models [57, Assumption 2.4] to matrix optimization—a property we empirically validate in Section 4.

3.2 Convergence Theory for Muon

Muon [13] (Algorithm 3) has recently emerged as a powerful optimizer for the hidden layers of large-scale neural networks, utilizing the principle of orthogonalized momentum. Algorithm 3 first tracks a momentum estimate $\mathbf{B}_t$ and subsequently employs the Newton–Schulz iteration [71, 72] to compute the matrix sign, $\mathrm{msign}(\mathbf{B}_t)$, which is further utilized to update the parameter matrix $\mathbf{X}_t$. Following common practice [73–76, 70], we assume zero numerical error in Newton–Schulz algorithm, i.e., $\texttt{NewtonSchulz}(\mathbf{X}) = \mathrm{msign}(\mathbf{X}) = \mathbf{U}\mathbf{V}^\top$ with $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ as the SVD of $\mathbf{X} \in \mathbb{R}^{m\times n}$. *We remark that it is introduced for simplicity rather than necessity, with details deferred to Appendix C.5.* In the following theorem, we establish the first convergence guarantee for Muon under our generalized heavy-tailed noise framework and matrix-based generalized smoothness conditions.

Theorem 3. *Under Assumptions 1b to 4b, define $\Delta_f := f(\mathbf{X}_1) - f^*$. By setting*

$$B = \max\left\{1, \left\lceil \left(64\sqrt{2}\,\|\mathbf{V}_1\|_{\mathrm{op}}\right)^{\frac{p}{p-1}} \right\rceil\right\}, \quad \beta = \max\left\{0, 1 - B^{\frac{2p-2}{3p-2}} \left(\frac{\Delta_f \|\mathbf{L}_0\|_*}{\|\mathbf{V}_0\|_*^2 T}\right)^{\frac{p}{3p-2}}\right\},$$
$$\eta = \min\left\{\sqrt{\frac{2\Delta_f(1-\beta)}{5\|\mathbf{L}_0\|_* T}}, \frac{1-\beta}{16\|\mathbf{L}_1\|_{\mathrm{op}}}\right\}, \tag{4}$$

Algorithm 3 ensures

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(\mathbf{X}_t)\|_*\right] \le O\left(\left(\Delta_f \|\mathbf{L}_0\|_*\right)^{\frac{p-1}{3p-2}} \|\mathbf{V}_0\|_*^{\frac{p}{3p-2}} (BT)^{-\frac{p-1}{3p-2}}\right).$$

Comparison with Shen et al. [74]. To compare with Frobenius-based analyses such as Shen et al. [74], let $r = \min\{m, n\}$ and orient the matrix so that $m \le n$, consistently with the full-row-rank assumptions. Set $p = 2$, $\mathbf{L}_0 = L\mathbf{I}_{m\times n}$, $\mathbf{V}_0 = \sigma\mathbf{I}_{m\times n}$, and $\mathbf{L}_1 = \mathbf{V}_1 = \mathbf{0}$. Since $\|\mathbf{L}_0\|_* = rL$ and $\|\mathbf{V}_0\|_* = r\sigma$, Theorem 3 gives

$$O\Big((\Delta_f rL)^{1/4}(r\sigma)^{1/2}T^{-1/4}\Big) = O\Big(r^{3/4}(\Delta_f L)^{1/4}\sigma^{1/2}T^{-1/4}\Big).$$

Under the same parametrization, Assumption 4b implies $\mathbb{E}\|\mathbf{G}_t^b - \nabla f(\mathbf{X}_t)\|_{\mathrm{F}}^2 \le r\sigma^2$. The matching Frobenius variance budget is therefore $\sigma_{[74]}^2 = r\sigma^2$, and the prior leading bound becomes

$$O\Big(r^{1/2}(\Delta_f L)^{1/4}\sigma_{[74]}^{1/2}T^{-1/4}\Big) = O\Big(r^{3/4}(\Delta_f L)^{1/4}\sigma^{1/2}T^{-1/4}\Big).$$

The isotropic worst-case scaling again coincides after aligning the noise budgets. For general curvature and noise, the spectral structure enters through

$$\|\mathbf{V}_0\|_* \le \sqrt{r}\,\|\mathbf{V}_0\|_{\mathrm{F}}, \qquad \|\mathbf{L}_0\|_* \le r\,\|\mathbf{L}_0\|_{\mathrm{op}}.$$

Under a common parametrization, our bound has no worse dependence on smoothness and noise, and is sharper when curvature or noise has low effective rank. It also covers $p \in (1, 2)$ and gradient-dependent noise. Section 4.2 examines this geometry during pretraining.

Finite-variance lower bound. For $p = 2$ and $\mathbf{L}_1 = \mathbf{V}_1 = \mathbf{0}$, the diagonal matrix construction of Tao et al. [59, Theorem 5] gives a nuclear-norm lower bound for Muon without momentum whose stochastic term is

$$\Omega\Big((\Delta_f\|\mathbf{L}_0\|_*)^{1/4}\|\mathbf{V}_0\|_*^{1/2}N^{-1/4}\Big), \qquad N = BT.$$

This matches the curvature, noise, and sample dependence of Theorem 3 at $p = 2$ in the covered subclass. The covariance assumption of that construction implies our weighted noise condition, as shown in Appendix B.1. These lower bounds do not establish tightness for $p < 2$, momentum, or gradient-dependent noise.

Remark 2. *Sfyraki and Wang [77] recently established a suboptimal $O(T^{-\frac{p-1}{3p-2}}\log T)$ convergence rate for* clipped *versions of Lion and Muon under heavy-tailed noise. Their analysis exhibits several limitations compared to our work: (i) their assumptions and convergence criteria rely on Euclidean norms ($\|\cdot\|_2, \|\cdot\|_{\mathrm{F}}$), which fail to capture the intrinsic geometry of sign-based methods, particularly from their Frank-Wolfe perspective; (ii) they require an additional bounded gradient assumption ($\|\nabla f(\mathbf{X})\|_{\mathrm{F}} \le G$) and still incur extraneous logarithmic factors; and (iii) their theory is restricted to settings with a strictly positive weight decay parameter ($\lambda > 0$) and provides no guarantees for the standard algorithms* without clipping.

3.3 Convergence Theory for Muonlight

Proposed by Liu et al. [18], Muonlight has achieved remarkable success in large-scale LLM pretraining, as evidenced by the Moonlight-16B-A3B, Kimi-K2, GLM-4.5 models. It introduces two primary modifications to the original Muon algorithm: (i) the integration of Nesterov momentum and (ii) the addition of decoupled weight decay. The former is a standard technique in modern optimization [78, 79]—already a key feature in the popular PyTorch implementation of Adam [80]—whose effectiveness is often attributed to implicit variance reduction [16, 35, 36]. The latter, decoupled weight decay, has proven essential for stabilizing the training trajectories of large foundation models [19]. The complete procedure is detailed in Algorithm 4. Notably, while original implementations often restrict Nesterov momentum to $\beta_1 = \beta_2$, our framework supports a general (β_1, β_2) configuration. We present the theoretical analysis of Muonlight below.

Theorem 4. *Under Assumptions 1b to 4b, define $\Delta_f := f(\mathbf{X}_1) - f^*$. Set*

$$B = \max\left\{1, \left\lceil \left(5958 \left\|\mathbf{V}_1\right\|_{\text{op}}\right)^{\frac{p}{p-1}} \right\rceil\right\}, \quad \beta_2 = \max\left\{0, 1 - B^{\frac{2p-2}{3p-2}} \left(\frac{\Delta_f \left\|\mathbf{L}_0\right\|_*}{\left\|\mathbf{V}_0\right\|_*^2 T}\right)^{\frac{p}{3p-2}}\right\},$$

$$\eta = \min\left\{\sqrt{\frac{4\Delta_f(1-\beta_2)}{15\left\|\mathbf{L}_0\right\|_* T}}, \frac{3(1-\beta_2)}{625\left\|\mathbf{L}_1\right\|_{\text{op}}}\right\}, \ \beta_1 \in \left[\max\left\{0, \beta_2 - \frac{3}{20}\right\}, \min\left\{1, \beta_2 + \frac{3}{20}\right\}\right], \tag{5}$$

and $\lambda \in \left[0, \left(1 - 2^{-\frac{1}{T}}\right)/\eta\right]$. If $\left\|\mathbf{X}_1\right\|_{\text{op}} \leq 1/(3\lambda)$, then Algorithm 4 ensures $\left\|\mathbf{X}_t\right\|_{\text{op}} \leq 2/(3\lambda), \forall t \in [T]$, and

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left\|\nabla f(\mathbf{X}_t)\right\|_*\right] \leq O\left(\left(\Delta_f \left\|\mathbf{L}_0\right\|_*\right)^{\frac{p-1}{3p-2}} \left\|\mathbf{V}_0\right\|_*^{\frac{p}{3p-2}} (BT)^{-\frac{p-1}{3p-2}}\right).$$

Several remarks are in order: (i) Muonlight achieves the same convergence bounds as Muon under generalized heavy-tailed noise. (ii) Our result provides the first analysis for $p \in (1, 2)$ and improves upon previous *dimension-dependent* rates for the $p = 2, \mathbf{L}_1 = \mathbf{V}_1 = \mathbf{0}$ case [75, 76]. (iii) We establish a uniform trajectory bound $\left\|\mathbf{X}_t\right\|_{\text{op}} \leq 2/(3\lambda)$, mirroring the stability of Lion in Theorem 2. (iv) The theory supports a broad range for the Nesterov momentum β_1, reinforcing the empirical flexibility and practicality of the algorithm.

4 Empirical Study

We study GPT-2 (124M parameters) pretraining on C4 to test the proposed noise models, measure the geometry in our complexity comparisons, and evaluate training performance. All runs use 10,000 steps, or approximately 2.62B tokens. Experimental protocols are in Appendix F. Code is available at `https://github.com/Dingzhen230/Heavy-tailed-Noise-in-LLMs`.

4.1 Heavy-Tailed Noise and the Generalized Noise Model

The stochastic gradient errors in Figure 1 show marked tail deviations from the Gaussian reference. We next examine whether the noise level varies with the gradient, as allowed by the growth terms in Assumptions 4a and 4b. At each saved checkpoint, we freeze the parameters and use fresh mini-batches to estimate the mean gradient and centered noise moments. Figures 2 and 3 compare these moments with the corresponding gradient quantities across sampled states. The approximately linear relation in both plots supports modeling the noise magnitude as dependent on the gradient.

We test this relation along Adam, Lion, and Muon trajectories, using the same model, dataset, and random seed. We also add Gaussian or Student-t perturbations, which make the optimizers visit points they would otherwise never visit. Table 1 reports the linear-fit scores for all nine settings. The relation persists across optimizers and under both perturbations, suggesting that it reflects the local loss geometry beyond the unperturbed trajectories.

Table 1: Linear-fit scores for the matrix noise model under trajectory perturbations.

Optimizer	Natural	Gaussian	Student-t
Adam	0.5807	0.9339	0.9465
Lion	0.7986	0.9571	0.9592
Muon	0.8514	0.8678	0.8713

4.2 Geometry Predicted by the Theory

The comparison with Euclidean normalization in Appendix B depends on $R = R_1 R_2^{p/(2(p-1))}$, where R_1 and R_2 measure curvature and noise density relative to gradient density. A ratio $R < 1$ favors our bound under this comparison. We estimate these ratios along a Lion trajectory over all 75 trainable tensors (124M parameters), using 256 stochastic gradients per checkpoint and Hessian–vector products (HVPs).

In Figure 4, $R < 1$ at every measured checkpoint and ranges from approximately 1.7×10^{-4} to 1.7×10^{-3} after initialization. Although R_2 is greater than one, the small curvature ratio R_1 keeps

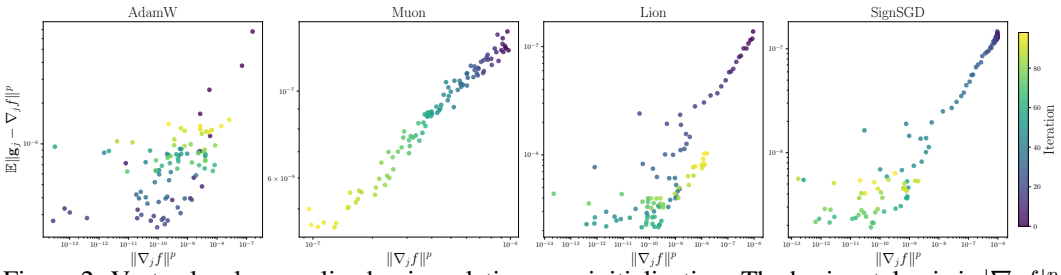


Figure 2: Vector-level generalized noise relation near initialization. The horizontal axis is $|\nabla_j f|^p$ and the vertical axis is $\mathbb{E}[|\mathbf{g}_j - \nabla_j f|^p]$, as in Assumption 4a.

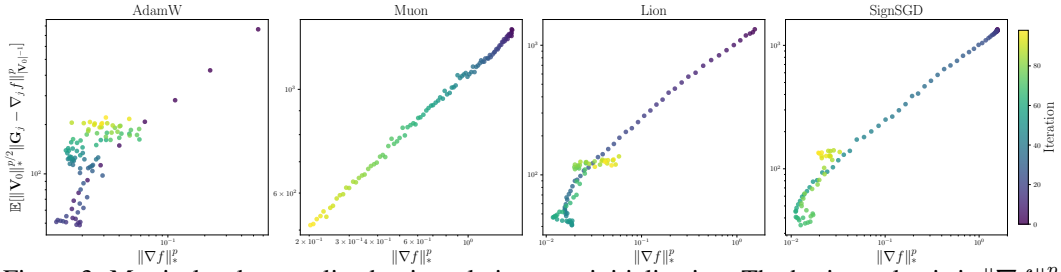


Figure 3: Matrix-level generalized noise relation near initialization. The horizontal axis is $\|\nabla f\|_*^p$ and the vertical axis is $\mathbb{E}[\|\mathbf{V}_0\|_*^{p/2}\|\mathbf{G} - \nabla f\|_{|\mathbf{V}_0|_m^{-1}}^p]$.

the combined ratio below one. Thus the measured advantage comes mainly from curvature that is concentrated relative to the gradient. These checkpoint estimates support the geometric regime in our comparison, while the theorem uses parameters valid along the trajectory. The full values and estimation protocol are in Appendix F.6.

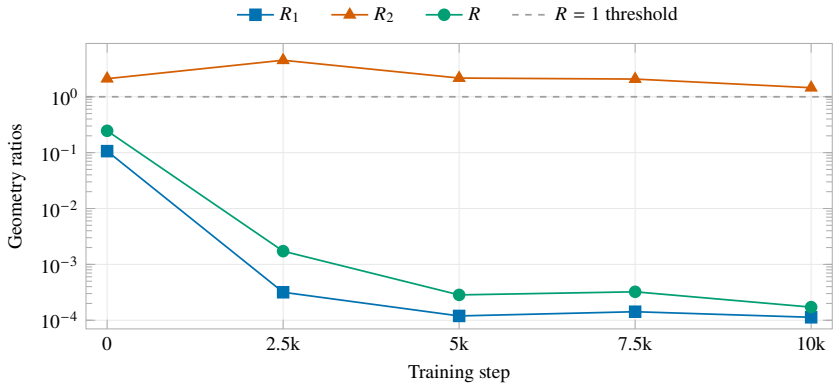


Figure 4: Geometry of the full model along the Lion trajectory. Curvature ratio R_1, noise ratio R_2, and combined ratio R are plotted on a logarithmic scale. The dashed line marks $R = 1$.

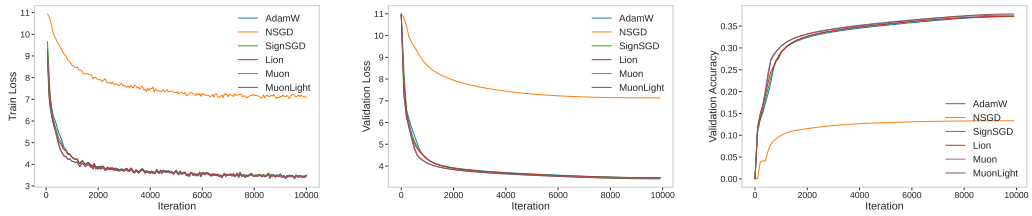
Table 2 reports the geometry of a 2304×768 attention weight in a Muon run. The Hessian ranks summarize eigenvalue concentration, while $\psi_2^2(\widehat{\mathbf{V}}_0)$ measures the noise effective dimension. At 10k steps, the Hessian ranks are 77.3 (nuclear) and 14.3 (stable), compared with the block Hessian dimension 1,769,472. The noise effective dimension falls from 76.4 to 37.2. These small effective dimensions matter because the bound in Section 3 depends on $\|\mathbf{L}_0\|_*$ and $\|\mathbf{V}_0\|_*$, whereas the isotropic conversion uses $r\|\mathbf{L}_0\|_{\mathrm{op}}$ and $\sqrt{r}\|\mathbf{V}_0\|_{\mathrm{F}}$, with $r = 768$. The measured noise factor gives $\|\widehat{\mathbf{V}}_0\|_*/(\sqrt{r}\|\widehat{\mathbf{V}}_0\|_{\mathrm{F}}) = \sqrt{37.2/768} \approx 0.22$ at 10k steps, illustrating how spectral concentration reduces the noise term. Its density relative to the gradient is 1.43 for the matrix normalization comparison (Appendix B). The Hessian ranks likewise indicate concentrated local curvature but do not identify the global smoothness matrix $\mathbf{L}_0$. Definitions and estimators are in Appendix F.6.

Table 2: Matrix geometry under Muon. Quantities are defined in Equation (38).

Step	Hessian $r_{\mathrm{nuc}}(H)$	Hessian $r_{\mathrm{st}}(H)$	Noise $\psi_2^2(\widehat{\mathbf{V}}_0)$
0	406.6	58.4	76.4
5k	130.3	15.5	51.9
10k	77.3	14.3	37.2

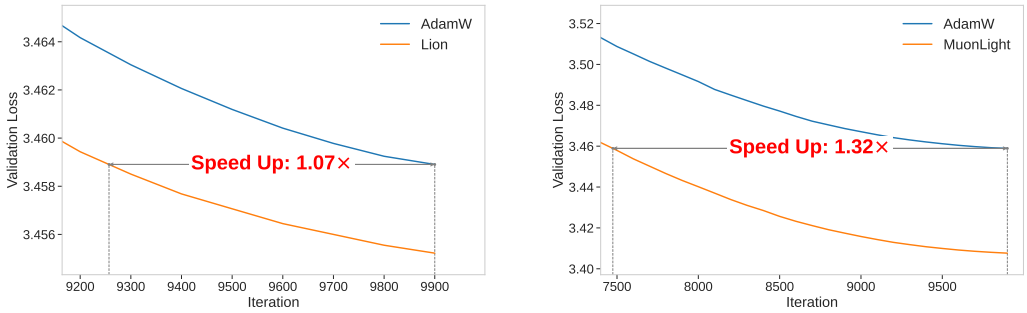
4.3 Pretraining Performance and Stability

Figure 5 compares the training loss, validation loss, and validation accuracy of six optimizers over the same 10,000-step budget. Sign-based methods outperform NSGD, and the matrix sign methods achieve the best validation performance. To quantify progress per update, we compare the iterations needed to reach a common validation loss, following Wen et al. [16]. The horizontal guides in Figure 6 mark the matched losses used for this comparison. Lion and Muon achieve iteration speedups of approximately $1.07\times$ and $1.32\times$ over AdamW, respectively. The full trajectories show how these differences develop over training.



(a) Train loss (b) Validation loss (c) Validation accuracy

Figure 5: Training loss, validation loss, and validation accuracy for GPT-2 pretraining on C4.



(a) Lion: $1.07\times$ iteration speedup (b) Muon: $1.32\times$ iteration speedup

Figure 6: Validation-loss comparisons with AdamW for Lion (left) and Muonlight (right).

Repeated runs give closely clustered final validation losses (Table 6): 3.467 ± 0.001 for AdamW, 3.464 ± 0.001 for Lion, and 3.429 ± 0.0005 for Muon. Lion thus has variability comparable to AdamW, while Muon reaches a lower mean loss with smaller dispersion. The average step takes 447 ms for AdamW, 443 ms for Lion, and 467 ms for Muon due to five Newton–Schulz iterations. Since time to a matched loss is the step count times the time per step, Lion attains a time speedup of approximately $1.08\times$ over AdamW. For Muon, fewer steps outweigh the per-step overhead, yielding approximately $1.27\times$. Perturbing representative final checkpoints along 100 random directions increases the mean validation loss by only 2.8×10^{-5} to 1.08×10^{-4}. These results describe efficiency and local sensitivity within the training budget, with details in Appendices F.8 and F.9.

5 Conclusion

This work aims to provide theoretical justifications for the empirical success of sign-based optimization algorithms such as Lion and Muon, particularly in the context of training LLMs. We introduce generalized heavy-tailed gradient noise assumptions that allow the noise magnitude to scale linearly with the gradient norm, a model we empirically validate as accurately reflecting the stochastic dynamics of real-world LLM training. Under this regime, we establish sharp convergence guarantees for SignSGD, Lion, Muon, and Muonlight, achieving rates that match or surpass the best-known bounds. Crucially, our theory indicates provable advantages of sign-based methods over NSGD and AdamW in the presence of heavy-tailed stochasticity. These findings are corroborated by LLM pretraining experiments, which confirm the practical superiority of sign-based optimizers.

Acknowledgments

This work was partially supported by NSFC (U23A20382), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118), and the Collaborative Innovation Center of Novel Software Technology and Industrialization. Dingzhi Yu is supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (NO.26CXJH0364).

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[6] 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.

[7] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

[9] 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.

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

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

[12] 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.

[13] 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/`.

[14] Rosie Zhao, Depen Morwani, David Brandfonbrener, Nikhil Vyas, and Sham Kakade. Deconstructing what makes a good optimizer for autoregressive language models. In *International Conference on Learning Representations (ICLR)*, pages 2830–2850, 2025.

[15] 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.

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

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 5827–5837, 2019.

[25] 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.

[26] Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 3964–3975, 2021.

[27] Barak Battash, Lior Wolf, and Ofir Lindenbaum. Revisiting the noise model of stochastic gradient descent. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4780–4788, 2024.

[28] 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.

[29] Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *International Conference on Learning Representations (ICLR)*, pages 16193–16205, 2024.

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

[31] Guanghui Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

[32] Ilyas Fatkhullin, Florian Hübler, and Guanghui Lan. Can SGD handle heavy-tailed noise? In *OPT 2025: Optimization for Machine Learning*, 2025.

[33] 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*, 2026.

[34] Savelii Chezhegov, Klyukin Yaroslav, Andrei Semenov, Aleksandr Beznosikov, Alexander Gasnikov, Samuel Horváth, Martin Takáč, and Eduard Gorbunov. Clipping improves Adam-norm and AdaGrad-norm when the noise is heavy-tailed. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pages 10269–10333, 2025.

[35] Huizhuo Yuan, Yifeng Liu, Shuang Wu, Zhou Xun, and Quanquan Gu. MARS: Unleashing the power of variance reduction for training large models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pages 73553–73587, 2025.

[36] Yifeng Liu, Angela Yuan, and Quanquan Gu. MARS-M: When variance reduction meets matrices. *arXiv preprint arXiv:2510.21800*, 2025.

[37] Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, and Weinan E. Towards theoretically understanding why SGD generalizes better than Adam in deep learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 21285–21296, 2020.

[38] 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.

[39] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations (ICLR)*, 2020.

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

[41] Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 5396–5427, 2023.

[42] Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive SGD. In *Proceedings of the 36th Conference on Learning Theory (COLT)*, pages 89–160, 2023.

[43] Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of AdaGrad for non-convex objectives: Simple proofs and relaxed assumptions. In *Proceedings of the 36th Conference on Learning Theory (COLT)*, pages 161–190, 2023.

[44] Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of Adam under relaxed assumptions. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 52166–52196, 2023.

[45] Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making Muon & Scion great again!(bridging theory and practice of LMO-based optimizers for LLMs). *arXiv preprint arXiv:2505.13416*, 2025.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

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

[51] 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.

[52] Yuxing Liu, Rui Pan, and Tong Zhang. AdaGrad under anisotropic smoothness. In *International Conference on Learning Representations (ICLR)*, pages 19574–19608, 2025.

[53] 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, 2026.

[54] 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.

[55] 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.

[56] 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*, 2025.

[57] Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations (ICLR)*, pages 92529–92554, 2025.

[58] Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed SGD. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2413–2421, 2025.

[59] Hongyi Tao, Dingzhi Yu, and Lijun Zhang. When and Why SignSGD Outperforms SGD: A Theoretical Study Based on ℓ_1-norm Lower Bounds. In *Advances in Neural Information Processing Systems 39 (NeurIPS)*, page to appear, 2026.

[60] Ganzhao Yuan and Bernard Ghanem. ℓ_0TV: A sparse optimization method for impulse noise image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 352–364, 2017.

[61] Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as SGD. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 10367–10376, 2020.

[62] 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.

[63] 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.

[64] 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.

[65] Dmitry Kovalev and Ekaterina Borodich. Non-Euclidean SGD for structured optimization: Unified analysis and improved rates. *arXiv preprint arXiv:2511.11466*, 2025.

[66] Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why Transformers need Adam: A Hessian perspective. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pages 131786–131823, 2024.

[67] Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik (Durk) Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. In *International Conference on Learning Representations (ICLR)*, pages 28033–28063, 2025.

[68] Dmitry Kovalev. SGD with adaptive preconditioning: Unified analysis and momentum acceleration. *arXiv preprint arXiv:2506.23803*, 2025.

[69] Shuo Xie, Tianhao Wang, Beining Wu, and Zhiyuan Li. A tale of two geometries: Adaptive optimizers and non-Euclidean descent. In *International Conference on Learning Representations (ICLR)*, 2026.

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

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

[72] 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.

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

[74] Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon. *Transactions on Machine Learning Research (TMLR)*, 2026.

[75] Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Convergence bound and critical batch size of Muon optimizer. *arXiv preprint arXiv:2507.01598*, 2025.

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

[77] 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.

[78] Timothy Dozat. Incorporating Nesterov momentum into Adam. In *International Conference on Learning Representations (ICLR) Workshop*, 2016.

[79] Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. Adan: Adaptive Nesterov momentum algorithm for faster optimizing deep models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9508–9520, 2024.

[80] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 8026–8037, 2019.

[81] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 1310–1318, 2013.

[82] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 5998–6008, 2017.

[83] Frederik Kunstner, Alan Milligan, Robin Yadav, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why Adam outperforms gradient descent on language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pages 30106–30148, 2024.

[84] Frederik Kunstner and Francis Bach. Scaling laws for gradient descent and sign descent for linear bigram models under Zipf's law. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pages 144435–144467, 2025.

[85] Robin Yadav, Shuo Xie, Tianhao Wang, and Zhiyuan Li. Provable benefit of sign descent: A minimal model under heavy-tail class imbalance. In *OPT 2025: Optimization for Machine Learning*, 2025.

[86] Daniel Hsu and Sivan Sabato. Heavy-tailed regression with a generalized median-of-means. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 37–45, 2014.

[87] Lijun Zhang and Zhi-Hua Zhou. ℓ_1-regression with heavy-tailed distributions. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 1076–1086, 2018.

[88] Jiujia Zhang and Ashok Cutkosky. Parameter-free regret in high probability with heavy tails. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 8000–8012, 2022.

[89] Zijian Liu. Online convex optimization with heavy tails: Old algorithms, new regrets, and applications. In *Proceedings of the 37th International Conference on Algorithmic Learning Theory (ALT)*, pages 1–47, 2026.

[90] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[91] Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Optimal algorithms for Lipschitz bandits with heavy-tailed rewards. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 4154–4163, 2019.

[92] Bo Xue, Guanghui Wang, Yimu Wang, and Lijun Zhang. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2936–2942, 2020.

[93] Yutian Gou, Jinfeng Yi, and Lijun Zhang. Stochastic graphical bandits with heavy-tailed rewards. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 734–744, 2023.

[94] Bo Xue, Yimu Wang, Yuanyu Wan, Jinfeng Yi, and Lijun Zhang. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 70880–70891, 2023.

[95] Chenlu Ye, Yujia Jin, Alekh Agarwal, and Tong Zhang. Catoni contextual bandits are robust to heavy-tailed rewards. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pages 71998–72040, 2025.

[96] Liam Hodgkinson and Michael Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 4262–4274, 2021.

[97] Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 4883–4895, 2021.

[98] Nuri Mert Vural, Lu Yu, Krishna Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In *Proceedings of 35th Conference on Learning Theory (COLT)*, pages 65–102, 2022.

[99] Zijian Liu, Jiawei Zhang, and Zhengyuan Zhou. Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In *Proceedings of 36th Conference on Learning Theory (COLT)*, pages 2266–2290, 2023.

[100] Zijian Liu and Zhengyuan Zhou. Stochastic nonsmooth convex optimization with heavy-tailed noises: High-probability bound, in-expectation rate and initial distance adaptation. *arXiv preprint arXiv:2303.12277*, 2023.

[101] Ta Duy Nguyen, Thien H Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 24191–24222. Curran Associates, Inc., 2023.

[102] Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 29563–29648, 2023.

[103] Nikita Kornilov, Ohad Shamir, Aleksandr Lobanov, Darina Dvinskikh, Alexander Gasnikov, Innokentiy Shibaev, Eduard Gorbunov, and Samuel Horváth. Accelerated zeroth-order method for non-smooth stochastic convex optimization problem with infinite variance. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 64083–64102, 2023.

[104] Nikita Puchkin, Eduard Gorbunov, Nickolay Kutuzov, and Alexander Gasnikov. Breaking the heavy-tailed noise barrier in stochastic optimization problems. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 856–864, 2024.

[105] Eduard Gorbunov, Abdurakhmon Sadiev, Marina Danilova, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability convergence for composite and distributed stochastic minimization and variational inequalities with heavy-tailed noise. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 15951–16070, 2024.

[106] Langqi Liu, Yibo Wang, and Lijun Zhang. High-probability bound for non-smooth non-convex stochastic optimization with heavy tails. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 32122–32138, 2024.

[107] Chuan He, Zhaosong Lu, Defeng Sun, and Zhanwang Deng. Complexity of normalized stochastic first-order methods with momentum under heavy-tailed noise. *arXiv preprint arXiv:2506.11214*, 2025.

[108] Zhuanghua Liu and Luo Luo. Stochastic bilevel optimization with heavy-tailed noise. *arXiv preprint arXiv:2509.14952*, 2025.

[109] Menglian Wang, Zhuanghua Liu, and Luo Luo. Near-optimal decentralized stochastic nonconvex optimization with heavy-tailed noise. *arXiv preprint arXiv:2601.11435*, 2026.

[110] Hongxu Chen, Ke Wei, Xiaoming Yuan, and Luo Luo. Stability and generalization of nonconvex optimization with heavy-tailed noise. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, pages 17344–17370, 2026.

[111] Yidong Wu and Luo Luo. Optimal asynchronous stochastic nonconvex optimization under heavy-tailed noise. *arXiv preprint arXiv:2601.19379*, 2026.

[112] Zijian Liu. Clipped gradient methods for nonsmooth convex optimization under heavy-tailed noise: A refined analysis. In *International Conference on Learning Representations (ICLR)*, pages 116560–116608, 2026.

[113] 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.

[114] 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, 2025.

[115] Wei Jiang, Li Zechao, and Lijun Zhang. Revisiting distributed sign-based variance reduction. *arXiv preprint arXiv:2609.18656*, 2026.

[116] 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.

[117] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Proceedings of Interspeech 2014*, pages 1058–1062, 2014.

[118] 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*, 2026.

[119] Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. SignSGD via zeroth-order oracle. In *International Conference on Learning Representations (ICLR)*, 2019.

[120] Egor Petrov, Grigoriy Evseev, Aleksey Antonov, Andrey Veprikov, Nikolay Bushkov, Stanislav Moiseev, and Aleksandr Beznosikov. Leveraging coordinate momentum in SignSGD and Muon: Memory-optimized zero-order. *arXiv preprint arXiv:2506.04430*, 2025.

[121] 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.

[122] Bingrui Li, Wei Huang, Andi Han, Zhanpeng Zhou, Taiji Suzuki, Jun Zhu, and Jianfei Chen. On the optimization and generalization of two-layer transformers with sign gradient descent. In *International Conference on Learning Representations (ICLR)*, pages 64406–64484, 2025.

[123] Enea Monzio Compagnoni, Tianlin Liu, Rustem Islamov, Frank Proske, Antonio Orvieto, and Aurelien Lucchi. Adaptive methods through the lens of sdes: Theoretical insights on the role of noise. In *International Conference on Learning Representations (ICLR)*, pages 95500–95557, 2025.

[124] Simon Elistratov, Andrey Podivilov, Timofei Iuzhakov, and Dmitry Vetrov. Lion's sign noise can make training more stable. In *OPT 2024: Optimization for Machine Learning Workshop of NeurIPS 2024*, 2024.

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

[126] Kwangjun Ahn, Byron Xu, Natalie Abreu, Ying Fan, Gagik Magakyan, Pratyusha Sharma, Zheng Zhan, and John Langford. Dion: Distributed orthonormalized updates. *arXiv preprint arXiv:2504.05295*, 2025.

[127] Kwangjun Ahn, Noah Amsel, and John Langford. Dion2: A simple method to shrink matrix in Muon. *arXiv preprint arXiv:2512.16928*, 2025.

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

[129] Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research (TMLR)*, 2026.

[130] Feihu Huang, Yuning Luo, and Songcan Chen. LiMuon: Light and fast Muon optimizer for large models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, pages 47263–47305, 2026.

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

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

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

[134] 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.

[135] 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*, 2025.

[136] 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)*, pages 161725–161761, 2025.

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

[138] Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Y. F. Tan. Muon outperforms Adam in tail-end associative memory learning. In *International Conference on Learning Representations (ICLR)*, 2026.

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

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

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

[142] Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. NorMuon: Making Muon more efficient and scalable. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, pages 68514–68533, 2026.

[143] Michael Crawshaw, Chirag Modi, Mingrui Liu, and Robert M. Gower. An exploration of non-euclidean gradient descent: Muon and its many variants. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, pages 21681–21716, 2026.

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

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

[146] 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.

[147] Yubo Zhang and Junhong Lin. On convergence of Muon for nonconvex stochastic optimization under generalized smoothness, February 2026. URL `https://doi.org/10.36227/techrxiv.177220100.05629800/v1`.

[148] Hideaki Iiduka. Muon converges under heavy-tailed noise: Nonconvex hölder-smooth empirical risk minimization. *arXiv preprint arXiv:2603.15059*, 2026.

[149] Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pages 49069–49104, 2025.

[150] Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-Euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

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

[152] Tao Sun, Xinwang Liu, and Kun Yuan. Revisiting gradient normalization and clipping for nonconvex SGD under heavy-tailed noise: Necessity, sufficiency, and acceleration. *Journal of Machine Learning Research (JMLR)*, 26(237):1–42, 2025.

[153] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, page 928–935, 2003.

[154] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[155] Yurii Nesterov. Minimization methods for nonsmooth convex and quasiconvex functions. *Matekon*, 29(3):519–531, 1984.

[156] Elad Hazan, Kfir Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 1594–1602, 2015.

[157] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[158] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training BERT in 76 minutes. In *International Conference on Learning Representations (ICLR)*, 2020.

[159] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 2260–2268, 2020.

[160] 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.

[161] 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, 2025.

[162] Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the Hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

[163] 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.

[164] Yikai Wu, Xingyu Zhu, Chenwei Wu, Annie Wang, and Rong Ge. Dissecting Hessian: Understanding common structure of Hessian in neural networks. *arXiv preprint arXiv:2010.04261*, 2020.

[165] Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

[166] Jiawei Zhao, Florian Tobias Schaefer, and Anima Anandkumar. Zero initialization: Initializing neural networks with only zeros and ones. *Transactions on Machine Learning Research (TMLR)*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=1AxQpKmiTc.

[167] Romain Cosson, Ali Jadbabaie, Anuran Makur, Amirhossein Reisizadeh, and Devavrat Shah. Low-rank gradient descent. *IEEE Open Journal of Control Systems*, 2:380–395, 2023.

[168] Greg Yang, James B Simon, and Jeremy Bernstein. A spectral condition for feature learning. *arXiv preprint arXiv:2310.17813*, 2023.

[169] Minhak Song, Kwangjun Ahn, and Chulhee Yun. Does SGD really happen in tiny subspaces? In *International Conference on Learning Representations (ICLR)*, pages 8086–8120, 2025.

[170] Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pages 3920–3930, 2025.

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

[172] Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton–Schulz. In *International Conference on Learning Representations (ICLR)*, pages 45672–45718, 2026. URL https://openreview.net/forum?id=lJSfxtLpLm.

[173] Noah Amsel, David Persson, Christopher Musco, and Robert M Gower. The polar express: Optimal matrix sign methods and their application to the Muon algorithm. *arXiv preprint arXiv:2505.16932*, 2025.

[174] Matthew J. Streeter and H. Brendan McMahan. Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*, 2010.

[175] Alexander Rakhlin and Karthik Sridharan. On equivalence of martingale tail bounds and deterministic regret inequalities. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, pages 1704–1722, 2017.

[176] Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Conference on Learning Theory (COLT)*, pages 244–256, 2010.

[177] Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213v8*, 2019.

[178] Adam Osękowski. Two inequalities for the first moments of a martingale, its square function and its maximal function. *Bulletin of the Polish Academy of Sciences. Mathematics*, 53(4): 441–449, 2005. doi: 10.4064/ba53-4-9.

[179] Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer Science & Business Media, 1996.

[180] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge University Press, 2012.

[181] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 1842–1850, 2018.

[182] Gilles Pisier and Quanhua Xu. Non-commutative martingale inequalities. *Communications in Mathematical Physics*, 189(3):667–698, 1997. doi: 10.1007/s002200050224.

[183] Shuo Xie, Tianhao Wang, Sashank J. Reddi, Sanjiv Kumar, and Zhiyuan Li. Structured preconditioners in adaptive optimization: A unified analysis. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pages 68746–68776, 2025.

[184] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

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

[186] 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.

[187] 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.

A Related Work

Heavy-tailed Noise The seminal work of Zhang et al. [25] theoretically explains why adaptive gradient methods like Clipped SGD [81] have a benefit for Attention models [82]. Inspired by Simsekli et al. [24], they identify the heavy-tailed noise distributions in language modeling tasks, and establish an $O(T^{-\frac{p-1}{3p-2}})$ convergence for clipped SGD with a matching lower bound under the standard heavy-tailed gradient noise condition, where the noise is only assumed to have a finite pth moment where $1 \leq p \leq 2$. Subsequent work studies heavy-tailed gradient noise and its implications for optimization and model architectures [26, 37, 29]. Complementary explanations of sign descent based on heavy-tailed class imbalance and Zipfian token frequencies concern data distributions [83–85]. As for the theory community, heavy-tailedness has been widely adopted and investigated in multiple realms, such as learning theory [86, 87], online learning [88, 89], bandits [90–95] and optimization [96–106, 57, 58, 32, 107–112].

Sign-based methods: vector optimizers The simplest signed gradient descent algorithm, namely SignSGD, is proposed by Bernstein et al. [38], where the authors establish convergence rates based on ℓ_1-norm and discuss its possible complexity improvement over SGD. Bernstein et al. [47] later shows that SignSGD with majority vote is communication efficient and robust to noise error. Further attempts have been made to improve the convergence of SignSGD for smooth functions [48, 55, 113, 56, 54, 114, 115]. While for non-smooth functions, Karimireddy et al. [116] provide a simple convex counterexample where SignSGD fails to converge, for which they utilize the error feedback techniques [117] to alleviate the divergence. Yu et al. [118] propose a new optimizer named StoSignSGD, provably fixing this non-convergence issue by injecting structural noise into the sign operator. Liu et al. [119], Petrov et al. [120] combine signed gradient with zeroth-order optimization to reduce memory cost in LLM training. SignSGD has also received great attention due to its close relationship with Adam [49]. Numerous efforts have been made to theoretically justify the success of Adam through the lens of SignSGD [51, 28, 121], and of course, study the effectiveness of SignSGD itself [50, 122, 123, 84]. Another important breakthrough is the Lion optimizer, proposed by Chen et al. [12] via symbolic search. By incorporating two momentum buffers, it achieves significant empirical success [14]. Likewise, there exists a rich body of literature that tries to explain its effectiveness [63, 124, 125, 62, 77]. Tao et al. [59] derive geometry-dependent lower bounds for SignSGD and Muon in the finite-variance setting, and establish a separation between SignSGD and vanilla SGD under sparse noise. Their lower bounds provide the finite-variance benchmarks discussed in Sections 2 and 3 and Appendix B.1.

Sign-based methods: matrix optimizers The Muon algorithm is introduced by Jordan et al. [13], which can be regarded as matrix sign descent. The matrix sign operation $\mathrm{msign}\,(\cdot)$ is implemented via Newton–Schulz iteration [71, 72]. Built upon Jordan et al. [13], Liu et al. [18] incorporate Nesterov momentum and learning rate alignment techniques into the original version of Muon. The modified algorithm Muonlight, has been applied to train massive-scale LLMs [18, 20, 19, 23] and exhibits clear performance gains over AdamW [16, 17, 15, 126, 127]. Many works have tried to theoretically analyze Muon (or its modified versions), or empirically explain its practical efficiency [73, 74, 76, 128–137, 70, 138–142, 77, 143–146].

Zhang and Lin [147] establish Muon convergence under generalized smoothness and finite-p heavy-tailed noise, while Iiduka [148] studies nonconvex Hölder-smooth empirical risk under heavy-tailed noise. Pethick et al. [149], Kovalev [150] interpret matrix orthogonalization through norm-constrained linear minimization and non-Euclidean trust-region optimization. Our analysis uses matrix-valued generalized smoothness and noise conditions to obtain bounds that depend on problem geometry, and covers SignSGD, Lion, Muon, and Muonlight. Yu et al. [151] study finite-time convergence of Signum and Muon in the nonsmooth setting.

How to handle heavy-tailed stochasticity? It is well-known that incorporating clipping or normalization into the vanilla SGD provably helps to tackle heavy-tailed noise [25, 97, 57, 58, 152, 107]. Chezhegov et al. [34] recently prove that Adam and AdaGrad might converge poorly in high probability with the presence of heavy-tailed gradient noise, while their clipped versions effectively fix this issue. Another important line of work focuses on vanilla versions of classic algorithms. Liu [89] studies SGD [153], Dual Averaging [154], and AdaGrad [9] from the perspective of online convex optimization. They show that these algorithms can converge optimally under heavy-tailedness. Fatkhullin

Algorithm 5 Normalized Stochastic Gradient Descent (NSGD)

Input: $T \in \mathbb{N}$, $\mathbf{x}_1 \in \mathbb{R}^d$, $\eta \in \mathbb{R}_+$
for $t = 1$ **to** T **do**
  $\mathbf{g}_t = \frac{1}{B}\sum_{b=1}^{B} \mathbf{g}_t^b$
  $\mathbf{m}_t = \beta \mathbf{m}_{t-1} + (1-\beta)\mathbf{g}_t \ \{\mathbf{m}_0 := \mathbf{g}_1\}$
  $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \frac{\mathbf{m}_t}{\|\mathbf{m}_t\|_2}$
end for

Algorithm 6 Matrix Normalized Stochastic Gradient Descent (MNSGD)

Input: $T \in \mathbb{N}$, $\mathbf{X}_1 \in \mathbb{R}^{m\times n}$, $\eta \in \mathbb{R}_+$
for $t = 1$ **to** T **do**
  $\mathbf{G}_t = \frac{1}{B}\sum_{b=1}^{B} \mathbf{G}_t^b$
  $\mathbf{M}_t = \beta_1 \mathbf{M}_{t-1} + (1-\beta_1)\mathbf{G}_t \ \{\mathbf{M}_0 := \mathbf{G}_1\}$
  $\mathbf{X}_{t+1} = \mathbf{X}_t - \eta \frac{\mathbf{M}_t}{\|\mathbf{M}_t\|_{\mathrm{F}}}$
end for

et al. [32] also analyze SGD under bounded pth moments of stochastic gradients. Their Theorem 5.10 further establishes convergence of mini-batch SGD under bounded central pth moments of gradient noise, using a weaker gradient-norm moment criterion than the expected squared gradient norm. Thus divergence in the latter measure does not preclude convergence in first-moment stationarity. Our focus is on geometry-adaptive guarantees for sign-based methods under generalized smoothness and gradient-dependent heavy-tailed noise.

B Provable Complexity Improvement of Sign Gradient Descent over Normalized Gradient Descent

This section provides a theoretical comparison between the complexities of sign-based methods and Normalized Stochastic Gradient Descent (NSGD) [155–159], highlighting scenarios where the former offers provable advantages. The comparison is motivated by two key observations: (i) NSGD is a robust baseline for heavy-tailed noise [57, 58], and (ii) both methods can be interpreted as steepest descent algorithms—or Linear Minimization Oracles (LMOs)—operating under distinct geometric constraints [160, 149, 77]. Consistent with the main text, both methods incorporate momentum, which is essential for tight convergence rates [159].

To quantify these geometric differences, we utilize the notion of *vector density function* following Bernstein et al. [38], Jiang et al. [161]:

$$\phi_q(\mathbf{v}) := \frac{\|\mathbf{v}\|_1}{\|\mathbf{v}\|_q} \in [1, d^{1-\frac{1}{q}}], \forall \mathbf{v} \in \mathbb{R}^d, q \in [1, \infty],$$

where a higher $\phi_q(\mathbf{v})$ indicates a denser vector. To compare different stationary measures, we define the trajectory-wide density as $\phi_2(\nabla_T) := \min_{t\in[T]} \phi_2(\nabla f(\mathbf{x}_t))$. This concept extends naturally to the matrix setting via the *matrix density function*:

$$\psi_q(\mathbf{Y}) := \frac{\|\mathbf{Y}\|_{S_1}}{\|\mathbf{Y}\|_{S_q}} \in [1, (\operatorname{rank}(\mathbf{Y}))^{1-\frac{1}{q}}], \forall \mathbf{Y} \in \mathbb{R}^{m\times n}, q \in [1, \infty],$$

where $\|\cdot\|_{S_q}$ denotes the Schatten q-norm (the ℓ_q-norm of the singular values). Notably, $\|\cdot\|_{S_1}$, $\|\cdot\|_{S_2}$, and $\|\cdot\|_{S_\infty}$ correspond to the nuclear norm $\|\cdot\|_*$, Frobenius norm $\|\cdot\|_{\mathrm{F}}$, and operator norm $\|\cdot\|_{\mathrm{op}}$, respectively. Intuitively, the matrix density function $\psi_q(\cdot)$ measures the distribution of singular values, capturing the "flatness" of the matrix spectrum. We define the matrix trajectory density as $\psi_2(\boldsymbol{\nabla}_T) := \min_{t\in[T]} \psi_2(\nabla f(\mathbf{X}_t))$. To simplify the comparison, we focus on the case where $\mathbf{l}_1 = \mathbf{L}_1 = \boldsymbol{\sigma}_1 = \mathbf{V}_1 = \mathbf{0}$. The complete versions of normalized gradient descent for vectors and matrices are listed in Algorithms 5 and 6, respectively.

Vector Optimization Under the conditions of Assumptions 2a and 4a, the objective f is effectively $\|\mathbf{l}_0\|_\infty$-smooth with $\|\boldsymbol{\sigma}_0\|_2$-heavy-tailed noise[2]. As established in Liu and Zhou [57], Algorithm 5 achieves a complexity of $O(\Delta_f \|\mathbf{l}_0\|_\infty \|\boldsymbol{\sigma}_0\|_2^{\frac{p}{p-1}} \epsilon^{-\frac{3p-2}{p-1}})$ for finding ℓ_2-stationary points, i.e., $\mathbb{E}[\|\nabla f(\mathbf{x})\|_2] \le \epsilon$. Using density functions to convert the complexity notion, we obtain the

[2]Technically speaking, we reparameterize the original σ_0 by $\|\boldsymbol{\sigma}_0\|_2$, which shares the same range as σ_0 in $\phi_2(\boldsymbol{\sigma}_0)$ and does not affect our later discussions.

following bounds for identifying ℓ_2-stationary points.

$$\text{SignSGD \& Lion (Theorems 1 and 2):} \quad O\left(\frac{\Delta_f \left\|\mathbf{l}_0\right\|_\infty \left\|\boldsymbol{\sigma}_0\right\|_2^{\frac{p}{p-1}} \phi_\infty(\mathbf{l}_0) \left(\phi_2(\boldsymbol{\sigma}_0)\right)^{\frac{p}{p-1}}}{\epsilon^{\frac{3p-2}{p-1}} \left(\phi_2(\nabla_T)\right)^{\frac{3p-2}{p-1}}}\right)$$

$$\text{NSGD [57]:} \quad O\left(\frac{\Delta_f \left\|\mathbf{l}_0\right\|_\infty \left\|\boldsymbol{\sigma}_0\right\|_2^{\frac{p}{p-1}}}{\epsilon^{\frac{3p-2}{p-1}}}\right)$$

Evidently, the relative complexity is governed by the ratio

$$R = R_1 (R_2)^{\frac{p}{2(p-1)}}, \quad \text{where } R_1 := \frac{\phi_\infty(\mathbf{l}_0)}{\phi_2^2(\nabla_T)} \text{ and } R_2 = \frac{\phi_2^2(\boldsymbol{\sigma}_0)}{\phi_2^2(\nabla_T)}.$$

Based on the empirical evidence in Bernstein et al. [38, Figure 1], we shall see that R_2 is a mild constant ($R_2 \leq 5$). For the dominant factor R_1, the extensive experiments on language modeling as well as computer vision tasks in Bernstein et al. [38], Dong et al. [125], Li et al. [114] show that the gradients $\{\nabla f(\mathbf{x}_t)\}_{t\in[T]}$ along the optimization trajectory remain dense, and the ratio $\phi_2(\nabla_T)$ is close to $\Theta(\sqrt{d})$. Since $\phi_\infty(\mathbf{l}_0) \in [1, d]$, so we have that $R_1 = O(1)$ in the worst case. When the curvature vector $\mathbf{l}_0$ exhibits axis-alignment properties [50] such that $\phi_\infty(\mathbf{l}_0) \approx 1$, SignSGD and Lion achieve a remarkable dimension-wise speedup of factor d over NSGD.

Matrix Optimization Consider the MNSGD method shown in Algorithm 6. The trajectory of Algorithm 6 is equivalent to Algorithm 5 since $\|\mathbf{Y}\|_{\mathrm{F}} = \|\mathrm{vec}(\mathbf{Y})\|_2$. Under Assumptions 2b and 4b and by the same reparameterization trick as in the vector case, we can deduce that the objective f is $\|\mathbf{L}_0\|_{\mathrm{op}}$-smooth and $\|\mathbf{V}_0\|_{\mathrm{F}}$-heavy-tailed. The results in Liu and Zhou [57] suggest that MNSGD require $O(\Delta_f \|\mathbf{L}_0\|_{\mathrm{op}} \|\mathbf{V}_0\|_{\mathrm{F}}^{\frac{p}{p-1}} \epsilon^{-\frac{3p-2}{p-1}})$ iterations to reach $\mathbb{E}\left[\|\nabla f(\mathbf{X})\|_{\mathrm{F}}\right] \leq \epsilon$. Leveraging matrix density functions, we obtain the following complexity comparison:

$$\text{Muon \& Muonlight (Theorems 3 and 4):} \quad O\left(\frac{\Delta_f \|\mathbf{L}_0\|_{\mathrm{op}} \|\mathbf{V}_0\|_{\mathrm{F}}^{\frac{p}{p-1}} \psi_\infty(\mathbf{L}_0) \left(\psi_2(\mathbf{V}_0)\right)^{\frac{p}{p-1}}}{\epsilon^{\frac{3p-2}{p-1}} \left(\psi_2(\boldsymbol{\nabla}_T)\right)^{\frac{3p-2}{p-1}}}\right)$$

$$\text{MNSGD [57]:} \quad O\left(\frac{\Delta_f \|\mathbf{L}_0\|_{\mathrm{op}} \|\mathbf{V}_0\|_{\mathrm{F}}^{\frac{p}{p-1}}}{\epsilon^{\frac{3p-2}{p-1}}}\right)$$

As in the vector case, the comparison is dominated by the ratios below:

$$R = R_1 (R_2)^{\frac{p}{2(p-1)}}, \quad \text{where } R_1 := \frac{\psi_\infty(\mathbf{L}_0)}{\psi_2^2(\boldsymbol{\nabla}_T)} \text{ and } R_2 = \frac{\psi_2^2(\mathbf{V}_0)}{\psi_2^2(\boldsymbol{\nabla}_T)}.$$

The Hessians in modern deep neural networks (DNN) are typically low-rank [162–164, 64], implying $\psi_\infty(\mathbf{L}_0)$ is close to 1. Furthermore, while gradients in DNNs frequently exhibit low-rank structures [165–168], recent evidence [169] suggests that gradient components aligning with the low-rank eigenspace may not effectively reduce training loss. When the variance and gradient matrices share a similar effective rank—particularly under Muon or Muonlight, which balance singular values via orthogonalization—the gradients can maintain a high effective rank [70]. In this scenario, $R_1 \approx 1/\min\{m, n\}$ and R_2 remains a mild constant, highlighting a significant $\min\{m, n\}$ complexity improvement of Muon over MNSGD.

Lower Bounds The comparisons above follow the approach of Bernstein et al. [38] by evaluating the upper bounds of both method classes. However, this argument can be significantly strengthened by comparing the *upper bounds* of sign-based optimizers directly against the *lower bounds* of gradient normalization. Given that the complexity results for NSGD and MNSGD are tight [57], this comparison highlights a fundamental, provable advantage of sign-descent methods. Similar to the approach in Jiang et al. [161], evaluating our upper bounds against established lower bounds rigorously demonstrates the superiority of sign-based optimization in these specific geometric settings.

We consolidate these findings in Table 3. This table explicitly illustrates how sign-based algorithms leverage problem geometry to achieve superior convergence rates compared to the established lower bounds of normalized gradient methods.

Table 3: Complexity comparison to find an ϵ-stationary point.

Setting	Algorithm	Criterion	Complexity	Improvement
Vector	NSGD	$\mathbb{E}[\|\nabla\|_2] \le \epsilon$	$O\left(\frac{\Delta_f \|\mathbf{l}_0\|_\infty \|\boldsymbol{\sigma}_0\|_2^{\frac{p}{p-1}}}{\epsilon^{\frac{3p-2}{p-1}}}\right)$	$=$ lower bound
	SignSGD & Lion	$\mathbb{E}[\|\nabla\|_1] \le \epsilon$	$O\left(\frac{\Delta_f \|\mathbf{l}_0\|_1 \|\boldsymbol{\sigma}_0\|_1^{\frac{p}{p-1}}}{\epsilon^{\frac{3p-2}{p-1}}}\right)$	Up to d
Matrix	MNSGD	$\mathbb{E}[\|\boldsymbol{\nabla}\|_{\mathrm{F}}] \le \epsilon$	$O\left(\frac{\Delta_f \|\mathbf{L}_0\|_{\mathrm{op}} \|\mathbf{V}_0\|_{\mathrm{F}}^{\frac{p}{p-1}}}{\epsilon^{\frac{3p-2}{p-1}}}\right)$	$=$ lower bound
	Muon & Muonlight	$\mathbb{E}[\|\boldsymbol{\nabla}\|_*] \le \epsilon$	$O\left(\frac{\Delta_f \|\mathbf{L}_0\|_* \|\mathbf{V}_0\|_*^{\frac{p}{p-1}}}{\epsilon^{\frac{3p-2}{p-1}}}\right)$	Up to $\min\{m, n\}$

B.1 Finite-Variance Lower Bounds and Tightness

Tao et al. [59, Theorems 2 and 5] give algorithm-specific lower bounds for SignSGD and Muon without momentum. These results complement our upper bounds by resolving the geometry dependence in a finite-variance subclass. Let $N = BT$ denote the total number of stochastic-gradient queries, and set $p = 2$ and $\mathbf{l}_1 = \boldsymbol{\sigma}_1 = \mathbf{L}_1 = \mathbf{V}_1 = \mathbf{0}$.

For the vector construction, separable smoothness gives $L_\infty = \|\mathbf{l}_0\|_1$, and the coordinate-wise variance parameters are $\boldsymbol{\sigma}_0$. The hard instance in their proof is separable, with the curvature allocated across coordinates under this aggregate smoothness budget. It therefore satisfies our vector assumptions and yields

$$\mathbb{E}\left[\min_{t<T} \|\nabla f(\mathbf{x}_t)\|_1\right] = \Omega\left(\sqrt{\frac{\Delta_f \|\mathbf{l}_0\|_1}{N}} + \left(\frac{\Delta_f \|\mathbf{l}_0\|_1 \|\boldsymbol{\sigma}_0\|_1^2}{N}\right)^{1/4}\right).$$

The corresponding matrix construction lifts this separable instance through orthogonal transformations and diagonal embedding. Its smoothness parameter is $L_* = \|\mathbf{L}_0\|_*$ and its conditional covariance satisfies $\mathbb{E}[\mathbf{N}\mathbf{N}^\top \mid \mathcal{F}] \preceq \boldsymbol{\Sigma}^2$, where $\mathbf{N}$ is centered gradient noise. Taking $|\mathbf{V}_0|_{\mathrm{m}} = \boldsymbol{\Sigma} \succ \mathbf{0}$ implies

$$\mathbb{E}[\|\mathbf{N}\|_{\boldsymbol{\Sigma}^{-1}}^2 \mid \mathcal{F}] = \mathrm{tr}\big(\boldsymbol{\Sigma}^{-1}\mathbb{E}[\mathbf{N}\mathbf{N}^\top \mid \mathcal{F}]\big) \le \mathrm{tr}(\boldsymbol{\Sigma}) = \|\mathbf{V}_0\|_*,$$

which is precisely our weighted noise condition at $p = 2$, $\mathbf{V}_1 = \mathbf{0}$. Singular covariance factors can be approximated by positive-definite factors to respect our full-row-rank convention. The lifted hard instance gives

$$\mathbb{E}\left[\min_{t<T} \|\nabla f(\mathbf{X}_t)\|_*\right] = \Omega\left(\sqrt{\frac{\Delta_f \|\mathbf{L}_0\|_*}{N}} + \left(\frac{\Delta_f \|\mathbf{L}_0\|_* \|\mathbf{V}_0\|_*^2}{N}\right)^{1/4}\right).$$

The $N^{-1/4}$ terms match the curvature, noise, and oracle complexity dependence of our finite-variance stochastic upper bounds. A lower bound on the expected minimum also bounds the expected average along the same trajectory, so the stationarity criteria are compatible. This match concerns the subclass without momentum.

C Further Discussions on Assumptions

In this section, we complete the omitted discussions in Sections 2 and 3.

C.1 Generalized Smoothness: Assumption 2a

Below, we briefly compare Assumption 2a with other $(\mathbf{l}_0, \mathbf{l}_1)$-*coordinate-wise smooth* conditions. Consider the generalized smoothness model proposed in Crawshaw et al. [51, Assumption 2]:

$$|\nabla_i f(\mathbf{x}') - \nabla_i f(\mathbf{x})| \le (\mathbf{l}_{0,i} + \mathbf{l}_{1,i} |\nabla_i f(\mathbf{x})|) \|\mathbf{x}' - \mathbf{x}\|_2, \forall\, i \in [d], \|\mathbf{x}' - \mathbf{x}\|_2 \le \frac{1}{\|\mathbf{l}_1\|_\infty}. \quad (6)$$

Evidently, the above assumption is stronger than Assumption 2a in the sense that it requires Equation (6) to hold for all coordinates, and that the requested domain is larger as $\|\mathbf{x}' - \mathbf{x}\|_\infty \leq \|\mathbf{x}' - \mathbf{x}\|_2$. Also, Equation (6) will incur an explicit dimensional factor of d, which is unfavorable for LLMs. Liu et al. [52] further refined the condition in Equation (6) into

$$\|\nabla f(\mathbf{x}') - \nabla f(\mathbf{x})\|_{(\mathbf{l}_0 + \mathbf{l}_1 \odot |\nabla f(\mathbf{x})|)^{-1}} \leq \|\mathbf{x}' - \mathbf{x}\|_{\mathbf{l}_0 + \mathbf{l}_1 \odot |\nabla f(\mathbf{x})|}, \quad \forall \|\mathbf{x}' - \mathbf{x}\|_{\mathbf{l}_1^2} \leq \sqrt{d}. \tag{7}$$

By standard calculus, it's easy to see that Equation (7) implies Assumption 2c, a variant of Assumption 2a to be introduced later. Note that all of our theoretical results also hold under Assumption 2c (see discussions below Assumption 2c). We deduce that our smoothness model based on the quadratic bound Equation (1) is generally weaker than the gradient curvature bound in Equation (6) and Equation (7). On the other hand, Equation (6) leverages ℓ_2-norm while Equation (7) utilizes $\mathbf{l}_1^2$-weighted norm, both of which fail to align with the geometry of sign gradient descent [160, 170]. On the contrary, our ℓ_∞-norm formulation perfectly matches the interpretation of sign descent from the perspective of linear minimization oracle [149]. Lastly, the empirical validations in Crawshaw et al. [51], Liu et al. [52], Yu et al. [53] confirms the practical value of Assumption 2a.

The following condition is a variant of Assumption 2a, and can be inferred from Equation (7).

Assumption 2c (Variant of Assumption 2a)**.** *There exists non-negative vectors* $\mathbf{l}_0 = [\mathbf{l}_{0,1}, \cdots, \mathbf{l}_{0,d}] \in \mathbb{R}_+^d$ *and* $\mathbf{l}_1 = [\mathbf{l}_{1,1}, \cdots, \mathbf{l}_{1,d}] \in \mathbb{R}_+^d$ *such that for all* $\|\mathbf{x}' - \mathbf{x}\|_{\mathbf{l}_1^2} \leq \sqrt{d}$[3]*:*

$$|f(\mathbf{x}') - (f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle)| \leq \frac{1}{2} \|\mathbf{x}' - \mathbf{x}\|^2_{\mathbf{l}_0 + \mathbf{l}_1 \odot |\nabla f(\mathbf{x})|}. \tag{8}$$

Ignoring the absolute value on the LHS of Equation (8), then Assumption 2c admits the exact same form as Lemma C.3 in Liu et al. [52], which is derived under condition Equation (7) [52, Assumption 5.1]. According to the textbook analysis [171], we can expand their proof to derive Equation (8), which indicates that the type of gradient curvature bound in Equation (6) and Equation (7) is *stronger* than the function value quadratic bound in Assumptions 2a and 2c. To illustrate that the theoretical guarantee in Section 2 still holds under Assumption 2c, note that the only difference between Assumptions 2a and 2c lies in the condition $\|\mathbf{x}' - \mathbf{x}\|_\infty \leq 1/\|\mathbf{l}_1\|_\infty$ versus $\|\mathbf{x}' - \mathbf{x}\|_{\mathbf{l}_1^2} \leq \sqrt{d}$. Delving into the analysis of SignSGD in Appendix D.3, Assumption 2a is used under

$$\eta \leq 1/\|\mathbf{l}_1\|_\infty \Longrightarrow \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_\infty = \eta \|\mathrm{sign}(\mathbf{m}_t)\|_\infty \leq 1/\|\mathbf{l}_1\|_\infty.$$

Since we have

$$\eta \leq 1/\|\mathbf{l}_1\|_\infty \Longrightarrow \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{l}_1^2} = \eta \sqrt{\sum_{i=1}^d \left(\mathrm{sign}(\mathbf{m}_{t,i})^2 \cdot \mathbf{l}_{1,i}^2\right)} \leq \frac{\|\mathbf{l}_1\|_2}{\|\mathbf{l}_1\|_\infty} \leq \sqrt{d},$$

so we can replace Assumption 2a by Assumption 2c. The same arguments are still valid for Lion, which we omit here for brevity. Last but not least, it is clear that our generalized smoothness model can be viewed as an instantiation of non-Euclidean generalized smoothness proposed in Yu et al. [53].

C.2 Generalized Smoothness: Assumption 2b

For completeness, we also provide the following assumption for Muon and Muonlight.

Assumption 2d (Variant of Assumption 2b)**.** *There exists non-negative constants* $\|\mathbf{L}_0\|_*, \|\mathbf{L}_1\|_{\mathrm{op}} \in \mathbb{R}$ *such that for all* $\|\mathbf{X}' - \mathbf{X}\|_{\mathrm{op}} \leq 1/\|\mathbf{L}_1\|_{\mathrm{op}}$*:*

$$\|\nabla f(\mathbf{X}') - \nabla f(\mathbf{X})\|_* \leq \left(\|\mathbf{L}_0\|_* + \|\mathbf{L}_1\|_{\mathrm{op}} \|\nabla f(\mathbf{X})\|_*\right) \|\mathbf{X}' - \mathbf{X}\|_{\mathrm{op}}. \tag{9}$$

Assumption 2d mirrors the original (L_0, L_1)-smoothness formulation in Zhang et al. [39] by utilizing scalar constants $(\|\mathbf{L}_0\|_*, \|\mathbf{L}_1\|_{\mathrm{op}})$. This is a milder condition than Assumption 2b, as the latter necessitates the specification of two full matrices $\mathbf{L}_0$ and $\mathbf{L}_1$. To demonstrate that Assumption 2d is a weaker requirement, consider the case where Assumption 2b holds. We have

$$\|\nabla f(\mathbf{X}') - \nabla f(\mathbf{X})\|_* \overset{\text{Lemma E.4}}{\leq} \sqrt{\|\mathbf{L}(\mathbf{X})\|_*} \|\nabla f(\mathbf{X}') - \nabla f(\mathbf{X})\|_{(\mathbf{L}(\mathbf{X}))^{-1}}$$

[3]The requirement $\|\mathbf{x}' - \mathbf{x}\|_{\mathbf{l}_1^2} \leq \sqrt{d}$ here as well as in Equation (7) is stated as $\|\mathbf{x}' - \mathbf{x}\|_{\mathbf{l}_1} \leq \sqrt{d}$ in Liu et al. [52], which is not correct after private communications with the authors of Liu et al. [52]. Thus, we discuss under the right condition here.

$$\overset{\text{Assumption 2b}}{\leq} \sqrt{\|\mathbf{L}(\mathbf{X})\|_*}\,\|\mathbf{X}'-\mathbf{X}\|_{\mathbf{L}(\mathbf{X})} = \sqrt{\|\mathbf{L}(\mathbf{X})\|_* \operatorname{tr}\left((\mathbf{X}'-\mathbf{X})^\top \mathbf{L}(\mathbf{X})\,(\mathbf{X}'-\mathbf{X})\right)}$$

$$\leq \sqrt{\|\mathbf{L}(\mathbf{X})\|_* \cdot \left(\|\mathbf{L}(\mathbf{X})\|_* \left\|(\mathbf{X}'-\mathbf{X})\,(\mathbf{X}'-\mathbf{X})^\top\right\|_{\text{op}}\right)}$$

$$\overset{\text{Lemma E.1}}{\leq} \|\mathbf{L}(\mathbf{X})\|_* \,\|\mathbf{X}'-\mathbf{X}\|_{\text{op}} \leq \left(\|\mathbf{L}_0\|_* + \|\mathbf{L}_1\|_{\text{op}}\,\|\nabla f(\mathbf{X})\|_*\right)\|\mathbf{X}'-\mathbf{X}\|_{\text{op}},$$

suggesting that Assumption 2b implies Assumption 2d. In order to be more consistent with prior work [64, 65] and address the matrix structures, we adopt the formulation in the main text.

C.3 Heavy-Tailed Noise: Assumption 4a

Our theoretical justification mainly follows from Liu and Zhou [57]. The following lemma shows that Assumption 4a is meaningful for non-zero $\boldsymbol{\sigma}_1$.

Lemma C.1 (Based on Example A.1 in Liu and Zhou [57]). *Given $\mathbf{x}_* = [x_{*,1}, \cdots, x_{*,d}] \in \mathbb{R}^d$, define the separable function $h(\mathbf{x}) := \sum_{i=1}^d h_i(x_i), \forall \mathbf{x} = [x_1, \cdots, x_d] \in \mathbb{R}^d$ with*

$$h_i(x_i) := \frac{1}{2}\mathbb{E}\left[(a_i x_i - b_i)^2\right], \quad a_i \sim \text{Bernoulli}(0.5), b = a_i x_{*,i} + \xi_i,$$

where ξ_i is a centered random variable independent of a_i and further satisfy $\mathbb{E}\left[|\xi_i|^p\right] \leq \sigma_i^p$ for some $\sigma_i \geq 0$. Then, for any h defined above, it holds that

1. *$\boldsymbol{\sigma}_1 = \mathbf{0}$: Assumption 4a can not be satisfied.*
2. *$\boldsymbol{\sigma}_1 > \mathbf{0}$: Assumption 4a is satisfied with $\boldsymbol{\sigma}_{0,i} = \sigma 2^{1-\frac{2}{p}}, \boldsymbol{\sigma}_{1,i} = 2^{-\frac{1}{p}} + 2^{1-\frac{2}{p}}, \forall i \in [d]$.*

Proof. The stochastic gradient, as well as the true gradient at point $\mathbf{x}$ is given by

$$\mathbf{g}_i = a_i\,(a_i x_i - b_i), \quad \nabla_i f(\mathbf{x}) = \mathbb{E}[a_i^2]x_i - \mathbb{E}[a_i b_i].$$

Then, the function h_i is reduced to the one-dimensional case in Liu and Zhou [57, Example A.1]. Thus, we apply their results to deduce that when $\boldsymbol{\sigma}_1 = \mathbf{0}$, for all $h_i, i \in [d]$, Assumption 4a can not be satisfied. Also, their results indicate that for any $i \in [d]$, we can select $\boldsymbol{\sigma}_{0,i}^p = 2^{p-2}\sigma^p$ and $\boldsymbol{\sigma}_{1,i}^p = 0.5 + 2^{p-2}$. Due to our separate construction of h and the independence between all coordinates, we can extend this conclusion to finish the second part of the proof. □

C.4 Heavy-Tailed Noise: Assumption 4b

The following assumption is a relaxation of Assumption 4b.

Assumption 4c (Variant of Assumption 4b). *There exists $p \in (1, 2]$, $\mathbf{V}_0 \in \mathbb{R}^{m\times n}$ with full row rank, and $\|\mathbf{V}_1\|_{\text{op}} \in \mathbb{R}_+$ such that*

$$\mathbb{E}\left[\left\|\mathbf{G}_t^b - \nabla f(\mathbf{X}_t)\right\|_{|\mathbf{V}_0|_{\text{m}}^{-1}}^p \middle| \mathcal{F}_{t-1}\right] \leq \|\mathbf{V}_0\|_*^{p/2} + \frac{\|\mathbf{V}_1\|_{\text{op}}^p\,\|\nabla f(\mathbf{X}_t)\|_*^p}{\|\mathbf{V}_0\|_*^{p/2}}, \quad \forall b \in [B]. \tag{10}$$

By the Cauchy-Schwarz inequality, $|\langle \mathbf{V}_1, \nabla f(\mathbf{X}_t)\rangle|^p \leq \|\mathbf{V}_1\|_{\text{op}}^p\,\|\nabla f(\mathbf{X}_t)\|_*^p$, which implies that Assumption 4b is a stronger condition than Assumption 4c. The latter is more amenable to empirical verification, as it replaces the full matrix $\mathbf{V}_1$ with a scalar coefficient $\|\mathbf{V}_1\|_{\text{op}}$, an approach we adopt in Section 4. We also underline that in both assumptions, the term $\|\mathbf{V}_0\|_*^{p/2}$ is introduced to maintain the homogeneity of the pth moment, ensuring that the scaling of the $|\mathbf{V}_0|_{\text{m}}^{-1}$-weighted metric remains consistent with the noise magnitude.

Crucially, the use of the $\|\cdot\|_{|\mathbf{V}_0|_{\text{m}}^{-1}}$ weighted norm allows our model to capture the anisotropic structural properties of the noise [69, 64]. Specifically, An et al. [64, Assumption 3] and Pan et al. [70, Assumption 3] assume a PSD constraint on the noise covariance: $\mathbb{E}[\left(\mathbf{G}_t^b - \nabla f(\mathbf{X}_t)\right)\left(\mathbf{G}_t^b - \nabla f(\mathbf{X}_t)\right)^\top | \mathcal{F}_{t-1}] \preceq \mathbf{V}_0\mathbf{V}_0^\top$. This condition directly implies $\mathbb{E}\left[\left\|\mathbf{G}_t^b - \nabla f(\mathbf{X}_t)\right\|_{|\mathbf{V}_0|_{\text{m}}^{-1}}^2 \middle| \mathcal{F}_{t-1}\right] \leq \|\mathbf{V}_0\|_*$ [69, Proposition A.10], illustrating how our framework encapsulates the geometry of the gradient's principal directions.

C.5 Exact Newton–Schulz Oracle

In practice, Algorithms 3 and 4 often runs in $q = 5$ Newton–Schulz iterations [71, 72]. The numerical error of the Newton–Schulz algorithm has been observed to exert little error on the optimization trajectory [13, 18]. Kim and Oh [172] investigate this phenomenon theoretically and show that Muon with inexact Newton–Schulz converges at the same rate compared to the exact SVD realization, up to a constant factor which converges to 1 *double exponentially* in q. After reading their proof, it should be evident that their analysis can serve as a black box compatible with ours. It suffices to adapt Equation (30) in Appendix E.3 to equation (5) in Kim and Oh [172]. Ultimately, we only need to pay the same constant factor for the convergence of Muon without assuming exact Newton–Schulz oracle. The same technique can be extended to other numerical algorithms that compute $\operatorname{msign}(\cdot)$ in Muon, e.g., the recent hit `PolarExpress` [173].

D Analysis for SignSGD and Lion

To facilitate the theoretical analysis, we define

$$\boldsymbol{\epsilon}_t := \mathbf{m}_t - \nabla f(\mathbf{x}_t), \quad \mathbf{n}_t := \mathbf{g}_t - \nabla f(\mathbf{x}_t), \quad \mathbf{s}_t := \nabla f(\mathbf{x}_{t-1}) - \nabla f(\mathbf{x}_t). \tag{11}$$

D.1 Vector Calculus

Lemma D.1 (Minkowski's inequality)**.** *For any $p \geq 1, x, y \in \mathbb{R}_+$, it holds that $(x+y)^{1/p} \leq x^{1/p} + y^{1/p}$.*

Proof. Applying Minkowski's inequality $\|\mathbf{x}+\mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$ for two-dimensional vectors $\mathbf{x} = (x^{1/p}, 0), \mathbf{y} = (0, y^{1/p})$ yields the result. □

Lemma D.2 (Jensen's inequality for ℓ_p-means)**.** *Let $p \in [1, +\infty]$ and let $X_1, \cdots, X_n$ be non-negative scalar random variables on some probability space, then*

$$\mathbb{E}\left[\sum_{i=1}^n X_i^p\right]^{1/p} \leq \left(\sum_{i=1}^n \mathbb{E}\left[X_i^p\right]\right)^{1/p}$$

Proof. Let $Y = (X_1, \ldots, X_n)$ be a random vector in $\mathbb{R}_+^n$. By Jensen's inequality as well as the convexity of $x \mapsto x^p, p \geq 1$:

$$\left(\mathbb{E}\left[\|Y\|_p\right]\right)^p \leq \mathbb{E}\left[\|Y\|_p^p\right] = \mathbb{E}\left[\sum_{i=1}^n X_i^p\right] = \sum_{i=1}^n \mathbb{E}[X_i^p].$$

Taking the $1/p$-th power on both sides yields the result. □

Lemma D.3 (Sign-difference Bound)**.** *For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, it holds that $\langle \mathbf{x}, \operatorname{sign}(\mathbf{x}) - \operatorname{sign}(\mathbf{y})\rangle \leq 2\|\mathbf{x}-\mathbf{y}\|_1$.*

Proof. Let $\mathbf{x} = [\mathbf{x}_1, \cdots, \mathbf{x}_d] \in \mathbb{R}^d, \mathbf{y} = [\mathbf{y}_1, \cdots, \mathbf{y}_d] \in \mathbb{R}^d$, then

$$\begin{aligned}
&\langle \mathbf{x}, \operatorname{sign}(\mathbf{x}) - \operatorname{sign}(\mathbf{y})\rangle = \sum_{i=1}^d \mathbf{x}_i \cdot [\operatorname{sign}(\mathbf{x}_i) - \operatorname{sign}(\mathbf{y}_i)] \\
\leq &\sum_{i=1}^d 2|\mathbf{x}_i| \cdot \mathbb{I}(\operatorname{sign}(\mathbf{x}_i) \neq \operatorname{sign}(\mathbf{y}_i)) \\
\leq &\sum_{i=1}^d 2|\mathbf{x}_i - \mathbf{y}_i| \cdot \mathbb{I}(\operatorname{sign}(\mathbf{x}_i) \neq \operatorname{sign}(\mathbf{y}_i)) \leq \sum_{i=1}^d 2|\mathbf{x}_i - \mathbf{y}_i| = 2\|\mathbf{x}-\mathbf{y}\|_1.
\end{aligned}$$

□

D.2 Technical Lemmas

The following lemma with $\mathbf{l}_1 = \mathbf{0}$ is originally proven in Bernstein et al. [38, Lemma F.3] with some inaccuracies, which we fix below with a clean pathwise analysis and extend to general $\mathbf{l}_1$ at the same time.

Lemma D.4. *Under Assumption 2a, for any sign vector* $\mathbf{s} \in \{-1, 1\}^d$ *and any* $\eta \le 1/\left\|\mathbf{l}_1\right\|_\infty$,

$$\left\|\nabla f(\mathbf{x} + \eta\mathbf{s}) - \nabla f(\mathbf{x})\right\|_1 \le 2\eta\left(\left\|\mathbf{l}_0\right\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x})|\rangle\right).$$

Proof. If $\eta = 0$, the claim is immediate. Hence, assume $\eta > 0$. Denote

$$\mathbf{z}_t := \mathbf{x} + t\eta\mathbf{s}, \qquad \mathbf{q}(t) := \nabla f(\mathbf{z}_t), \qquad Q(t) := \left\|\mathbf{q}(t) - \mathbf{q}(0)\right\|_1,$$

and define

$$\mathbf{D}_t := \operatorname{diag}\left(\mathbf{l}_0 + \mathbf{l}_1 \odot |\mathbf{q}(t)|\right), \qquad t \in [0, 1].$$

By Assumption 2a, at every point on the segment where f is twice-differentiable,

$$-\mathbf{D}_t \preceq \nabla^2 f(\mathbf{z}_t) \preceq \mathbf{D}_t.$$

Since $\mathbf{q}'(t) = \eta\nabla^2 f(\mathbf{z}_t)\mathbf{s}$ for almost every $t \in [0, 1]$, we have

$$\frac{1}{\eta}\left\|\mathbf{q}'(t)\right\|_1 = \left\|\nabla^2 f(\mathbf{z}_t)\mathbf{s}\right\|_1 = \max_{\mathbf{u}\in\{-1,1\}^d} \left\langle \mathbf{u}, \nabla^2 f(\mathbf{z}_t)\mathbf{s}\right\rangle.$$

We next show that the last quantity is bounded by $\operatorname{tr}(\mathbf{D}_t)$. Fix any $\mathbf{u}, \mathbf{s} \in \{-1, 1\}^d$, and set

$$\mathbf{a} := \frac{\mathbf{u} + \mathbf{s}}{2}, \qquad \mathbf{b} := \frac{\mathbf{u} - \mathbf{s}}{2}.$$

Then $\mathbf{u} = \mathbf{a} + \mathbf{b}$, $\mathbf{s} = \mathbf{a} - \mathbf{b}$, and $\mathbf{a}_i^2 + \mathbf{b}_i^2 = 1$ for every coordinate i. Using the symmetry of $\nabla^2 f(\mathbf{z}_t)$, we obtain

$$\begin{aligned}
\left\langle \mathbf{u}, \nabla^2 f(\mathbf{z}_t)\mathbf{s}\right\rangle &= \left\langle \mathbf{a} + \mathbf{b}, \nabla^2 f(\mathbf{z}_t)(\mathbf{a} - \mathbf{b})\right\rangle \\
&= \left\langle \mathbf{a}, \nabla^2 f(\mathbf{z}_t)\mathbf{a}\right\rangle - \left\langle \mathbf{b}, \nabla^2 f(\mathbf{z}_t)\mathbf{b}\right\rangle \\
&\le \langle \mathbf{a}, \mathbf{D}_t\mathbf{a}\rangle + \langle \mathbf{b}, \mathbf{D}_t\mathbf{b}\rangle \\
&= \sum_{i=1}^d (\mathbf{D}_t)_{ii}\left(\mathbf{a}_i^2 + \mathbf{b}_i^2\right) = \operatorname{tr}(\mathbf{D}_t).
\end{aligned}$$

Therefore, for almost every $t \in [0, 1]$,

$$\frac{1}{\eta}\left\|\mathbf{q}'(t)\right\|_1 \le \operatorname{tr}(\mathbf{D}_t).$$

Moreover,

$$\begin{aligned}
\operatorname{tr}(\mathbf{D}_t) &= \left\|\mathbf{l}_0\right\|_1 + \langle \mathbf{l}_1, |\mathbf{q}(t)|\rangle \\
&\le \left\|\mathbf{l}_0\right\|_1 + \langle \mathbf{l}_1, |\mathbf{q}(0)|\rangle + \left\|\mathbf{l}_1\right\|_\infty \left\|\mathbf{q}(t) - \mathbf{q}(0)\right\|_1 \\
&= \left\|\mathbf{l}_0\right\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x})|\rangle + \left\|\mathbf{l}_1\right\|_\infty Q(t),
\end{aligned}$$

where we used $||a| - |b|| \le |a - b|$ coordinatewise. Hence, for every $t \in [0, 1]$,

$$Q(t) \le \int_0^t \left\|\mathbf{q}'(\tau)\right\|_1 \mathrm{d}\tau \le \eta t\left(\left\|\mathbf{l}_0\right\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x})|\rangle\right) + \eta\left\|\mathbf{l}_1\right\|_\infty \int_0^t Q(\tau)\mathrm{d}\tau.$$

By Grönwall's inequality,

$$Q(t) \le \eta\left(\left\|\mathbf{l}_0\right\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x})|\rangle\right) \int_0^t e^{\eta\left\|\mathbf{l}_1\right\|_\infty (t-\tau)}\mathrm{d}\tau.$$

If $\left\|\mathbf{l}_1\right\|_\infty > 0$, then taking $t = 1$ gives

$$\begin{aligned}
Q(1) &\le \left(\left\|\mathbf{l}_0\right\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x})|\rangle\right) \frac{e^{\eta\left\|\mathbf{l}_1\right\|_\infty} - 1}{\left\|\mathbf{l}_1\right\|_\infty} \\
&\le \eta(e - 1)\left(\left\|\mathbf{l}_0\right\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x})|\rangle\right) \le 2\eta\left(\left\|\mathbf{l}_0\right\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x})|\rangle\right),
\end{aligned}$$

where we used $\eta\left\|\mathbf{l}_1\right\|_\infty \le 1$ and $e - 1 \le 2$. If $\left\|\mathbf{l}_1\right\|_\infty = 0$, the same conclusion follows directly from the preceding integral inequality with the last term equal to zero. Since

$$Q(1) = \left\|\nabla f(\mathbf{x} + \eta\mathbf{s}) - \nabla f(\mathbf{x})\right\|_1,$$

the proof is now complete. □

Lemma D.5 (pth moment of mini-batch noise)**.** *Under Assumptions 3a and 4a, the following holds for any $i \in [d], t \in [T]$:*

$$\mathbb{E}\left[\,|\mathbf{n}_t|^p \,|\, \mathcal{F}_{t-1}\right] \leq 2B^{1-p}\left(\boldsymbol{\sigma}_{0,i}^p + \boldsymbol{\sigma}_{1,i}^p\,|\nabla_i f(\mathbf{x}_t)|^p\right),$$

where $\mathbf{n}_t$ is defined in Equation (11)*.*

Proof. We denote $\mathbf{n}_{t,i}^b := \mathbf{g}_{t,i}^b - \nabla_i f(\mathbf{x}_t)$. So $\mathbf{n}_{t,i} = \frac{1}{B}\sum_{b=1}^B \mathbf{n}_{t,i}^b$. Under Assumption 4a and Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}\left[\,|\mathbf{n}_{t,i}|^p \,|\, \mathcal{F}_{t-1}\right] =& \mathbb{E}\left[\left|\frac{1}{B}\sum_{b=1}^B \mathbf{n}_{t,i}^b\right|^p \middle| \mathcal{F}_{t-1}\right] \overset{\#}{\leq} 2\mathbb{E}\left[\frac{1}{B^p}\sum_{b=1}^B \left|\mathbf{n}_{t,i}^b\right|^p \middle| \mathcal{F}_{t-1}\right] \\ &\leq \frac{2}{B^p}\sum_{b=1}^B \left(\boldsymbol{\sigma}_{0,i}^p + \boldsymbol{\sigma}_{1,i}^p\,|\nabla_i f(\mathbf{x}_t)|^p\right) = 2B^{1-p}\left(\boldsymbol{\sigma}_{0,i}^p + \boldsymbol{\sigma}_{1,i}^p\,|\nabla_i f(\mathbf{x}_t)|^p\right),\end{aligned}$$

where # leverages von Bahr-Esseen inequality [58, Lemma 10]. □

Lemma D.6 (Regret Analysis of Diagonal AdaGrad)**.** *Given an arbitrary sequence of vectors $\{\mathbf{v}_t\}_{t=1}^T \subset \mathbb{R}^d, T \in \mathbb{N}$, there exists a sequence of vectors $\{\mathbf{w}_t\}_{t=1}^T \subset \mathbb{R}^d$ such that (i) $\|\mathbf{w}_t\|_\infty \leq 1$, (ii) every $\mathbf{w}_t$ only depends on $\mathbf{v}_1, \cdots, \mathbf{v}_{t-1}$, and (iii) they further satisfy*

$$\sum_{t=1}^T \langle \mathbf{v}_t, \mathbf{w}_t\rangle \leq \sum_{i=1}^d 2\sqrt{2\sum_{t=1}^T \mathbf{v}_{t,i}^2} - \left\|\sum_{t=1}^T \mathbf{v}_t\right\|_1.$$

Proof. The proof of this lemma stems from the regret analysis of AdaGrad-Norm [174, 9] in Liu and Zhou [57], which is based on Rakhlin and Sridharan [175, Lemma 2]. Here, we adapt their analysis to the diagonal version of AdaGrad [176, 9]. WLOG, we assume $\mathbf{v}_{1,i}^2 > 0$. Otherwise, define $\tau = \operatorname{argmin}\{t \in \mathbb{N} | \mathbf{v}_{t,i}^2 > 0\}$ and set $\mathbf{w}_{t,i} = 0, \forall t \in [\tau - 1]$, then start the proof at τ. Define

$$\mathbf{w}_1 := \mathbf{0}, \quad \mathbf{w}_{t+1,i} := \Pi_{[-1,1]}\left[\mathbf{w}_{t,i} - \gamma_t \mathbf{v}_{t,i}\right], \quad \gamma_{t,i} = \begin{cases}\sqrt{\frac{2}{\sum_{s=1}^t \mathbf{v}_{s,i}^2}}, & t \geq 1 \\ +\infty, & t = 0\end{cases}, \quad \forall i \in [d], \tag{12}$$

which ensures (i) and (ii) in Lemma D.6. By standard analysis of projected online subgradient descent [177, Lemma 2.30], for any $\mathbf{u}_i \in [-1, 1]$,

$$\mathbf{v}_{t,i} \cdot (\mathbf{w}_{t,i} - \mathbf{u}_i) \leq \frac{(\mathbf{u}_i - \mathbf{w}_{t,i})^2}{2\gamma_{t,i}} - \frac{(\mathbf{u}_i - \mathbf{w}_{t+1,i})^2}{2\gamma_{t,i}} + \frac{\gamma_{t,i}\mathbf{v}_{t,i}^2}{2}.$$

Summing from 1 to T:

$$\begin{aligned}&\sum_{t=1}^T \mathbf{v}_{t,i} \cdot (\mathbf{w}_{t,i} - \mathbf{u}_i) \leq \sum_{t=1}^T \left(\frac{(\mathbf{u}_i - \mathbf{w}_{t,i})^2}{2\gamma_{t,i}} - \frac{(\mathbf{u}_i - \mathbf{w}_{t+1,i})^2}{2\gamma_{t,i}} + \frac{\gamma_{t,i}\mathbf{v}_{t,i}^2}{2}\right) \\ =&\frac{(\mathbf{u}_i - \mathbf{w}_{1,i})^2}{2\gamma_{1,i}} - \frac{(\mathbf{u}_i - \mathbf{w}_{T+1,i})^2}{2\gamma_{T,i}} + \sum_{t=2}^T (\mathbf{u}_i - \mathbf{w}_{t,i})^2\left(\frac{1}{2\gamma_{t,i}} - \frac{1}{2\gamma_{t-1,i}}\right) + \sum_{t=1}^T \frac{\gamma_{t,i}\mathbf{v}_{t,i}^2}{2} \\ \leq&\frac{1}{2\gamma_{1,i}} + \sum_{t=2}^T \left(\frac{2}{\gamma_{t,i}} - \frac{2}{\gamma_{t-1,i}}\right) + \sum_{t=1}^T \gamma_{t,i}\left(\frac{1}{2\gamma_{t,i}^2} - \frac{1}{2\gamma_{t-1,i}^2}\right) \\ \leq&\frac{2}{\gamma_{T,i}} + \sum_{t=1}^T 2\left(\frac{1}{\gamma_{t,i}} - \frac{1}{\gamma_{t-1,i}}\right) \leq \frac{4}{\gamma_{T,i}} = 2\sqrt{2\sum_{t=1}^T \mathbf{v}_{t,i}^2},\end{aligned}$$

where we utilize $\mathbf{u}_i \in [-1, 1], \mathbf{w}_{1,i} = 0, \gamma_t \leq \gamma_{t-1}, \mathbf{u}_i - \mathbf{w}_{t,i} \in [-2, 2], 1/\gamma_{0,i} = 0$ in the second inequality. Rearranging the above relation and taking minimum over $\mathbf{u}_i \in [-1, 1]$,

$$\sum_{t=1}^T \mathbf{v}_{t,i} \cdot \mathbf{w}_{t,i} \leq 2\sqrt{2\sum_{t=1}^T \mathbf{v}_{t,i}^2} + \min_{\mathbf{u}_i \in [-1,1]} \sum_{t=1}^T \mathbf{v}_{t,i} \cdot \mathbf{u}_i = 2\sqrt{2\sum_{t=1}^T \mathbf{v}_{t,i}^2} - \left|\sum_{t=1}^T \mathbf{v}_{t,i}\right|.$$

Summing the regret across all coordinates $i \in [d]$ finishes the proof. □

The following ℓ_1 martingale bound follows by applying the sharp scalar square-function inequality of Osękowski [178] coordinate-wise. We keep the regret-based derivation below because it gives a direct online-learning interpretation through diagonal AdaGrad.

Lemma D.7 (ℓ_1-norm vector martingale concentration)**.** *Given a sequence of integrable random vectors $\mathbf{v}_t \in \mathbb{R}^d, \forall t \in \mathbb{N}$ such that $\mathbb{E}[\mathbf{v}_t|\mathcal{F}_{t-1}] = \mathbf{0}$ where $\mathcal{F}_t = \sigma(\mathbf{v}_1, \cdots, \mathbf{v}_t)$ is the natural filtration, then for any $p \in [1, 2]$, there is*

$$\mathbb{E}\left[\left\|\sum_{t=1}^{T} \mathbf{v}_t\right\|_1\right] \leq 2\sqrt{2}\sum_{i=1}^{d} \mathbb{E}\left[\left\|\mathbf{v}_{1:T,i}\right\|_p\right], \quad \forall T \in \mathbb{N},$$

where $\mathbf{v}_{1:t,i} := [\mathbf{v}_{1,i}, \cdots, \mathbf{v}_{t,i}] \in \mathbb{R}^t$.

Proof. According to Lemma D.6, there exists a sequence of random vectors $\{\mathbf{w}_t\}_{t=1}^T \subset \mathbb{R}^d$ such that (i) $\|\mathbf{w}_t\|_\infty \leq 1$, (ii) every $\mathbf{w}_t \in \mathcal{F}_{t-1}$, and (iii) they further satisfy

$$\sum_{t=1}^{T} \langle \mathbf{v}_t, \mathbf{w}_t \rangle \leq \sum_{i=1}^{d} 2\sqrt{2} \left\|\mathbf{v}_{1:T,i}\right\|_2 - \left\|\sum_{t=1}^{T} \mathbf{v}_t\right\|_1.$$

Rearranging this inequality and taking expectations yields

$$\mathbb{E}\left[\left\|\sum_{t=1}^{T} \mathbf{v}_t\right\|_1\right] \leq 2\sqrt{2}\sum_{i=1}^{d} \mathbb{E}\left[\left\|\mathbf{v}_{1:T,i}\right\|_2\right] - \sum_{t=1}^{T} \mathbb{E}\left[\langle \mathbf{v}_t, \mathbf{w}_t \rangle\right] \leq 2\sqrt{2}\sum_{i=1}^{d} \mathbb{E}\left[\left\|\mathbf{v}_{1:T,i}\right\|_p\right],$$

where the last step is due to $\|\cdot\|_2 \leq \|\cdot\|_p$ and the tower rule:

$$\mathbb{E}\left[\langle \mathbf{v}_t, \mathbf{w}_t \rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\langle \mathbf{v}_t, \mathbf{w}_t \rangle | \mathcal{F}_{t-1}\right]\right] = \mathbb{E}\left[\langle \mathbb{E}\left[\mathbf{v}_t|\mathcal{F}_{t-1}\right], \mathbf{w}_t \rangle\right] = 0.$$

□

Lemma D.8 (Stability with weight decay)**.** *Running Algorithm 2 with*

$$\lambda \leq \frac{1}{\eta}\left(1 - \frac{1}{2^{1/T}}\right), \text{ and } \|\mathbf{x}_1\|_\infty \leq \frac{a}{\lambda}, \quad \forall a \in (0, 1),$$

ensures that

$$\|\mathbf{x}_t\|_\infty \leq \frac{a+1}{2\lambda}, \quad \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_\infty \leq \frac{a+3}{2}\eta, \quad \forall t \in [T].$$

Proof. We first show that the trajectory is bounded if the model is appropriately initialized with $\|\mathbf{x}_1\|_\infty \leq a/\lambda$. Denote $1 - \eta\lambda$ by q, suggesting $q \in (1/2^{1/T}, 1)$ by the choice of λ. By the update rule of Algorithm 2, we have

$$\begin{aligned}
&\|\mathbf{x}_t\|_\infty = \|(1-\eta\lambda)\mathbf{x}_{t-1} - \eta \operatorname{sign}(\mathbf{v}_{t-1})\|_\infty \leq q\|\mathbf{x}_{t-1}\|_\infty + \eta \\
\leq& q^2\|\mathbf{x}_{t-2}\|_\infty + q\eta + \eta \leq \cdots \leq q^{t-1}\|\mathbf{x}_1\|_\infty + \eta\sum_{i=0}^{t-2} q^i \leq \frac{aq^{t-1}}{\lambda} + \frac{\eta(1-q^{t-1})}{1-q} \\
=& \frac{1-(1-a)q^{t-1}}{\lambda} \leq \frac{1-(1-a)\left(\frac{1}{2}\right)^{\frac{t-1}{T}}}{\lambda} \leq \frac{a+1}{2\lambda},
\end{aligned}$$

which holds for any $t \in [T]$. Therefore, we can deduce

$$\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_\infty = \|\eta\lambda\mathbf{x}_t + \eta \operatorname{sign}(\mathbf{v}_t)\|_\infty \leq \eta\lambda\|\mathbf{x}_t\|_\infty + \eta \leq \frac{a+3}{2}\eta, \quad \forall t \in [T].$$

□

Lemma D.9 (Lemma D.4 with weight decay)**.** *For any sign vector $\mathbf{s} \in \{-1, 1\}^d$ and any $\eta \leq 1/((c+1)\|\mathbf{l}_1\|_\infty), c > 0$, consider the sign-based update with weight decay:*

$$\mathbf{x}' := \mathbf{x} + \eta\mathbf{s}', \quad \mathbf{s}' = -\mathbf{s} - \lambda\mathbf{x}, \quad \forall \lambda \geq 0, \mathbf{x} \in \mathbb{R}^d \text{ such that } \lambda\|\mathbf{x}\|_\infty \leq c.$$

Under Assumption 2a, it holds that

$$\|\nabla f(\mathbf{x}') - \nabla f(\mathbf{x})\|_1 \leq (2 + 2c)\eta\left(\|\mathbf{l}_0\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x})| \rangle\right).$$

Proof. The constraint on $\mathbf{x}$ implies $\lambda \|\mathbf{x}\|_\infty \le c$, where the case $\lambda = 0$ is immediate. Hence

$$\|\mathbf{s}'\|_\infty = \|-\mathbf{s} - \lambda\mathbf{x}\|_\infty \le \|\mathbf{s}\|_\infty + \lambda \|\mathbf{x}\|_\infty \le 1 + c.$$

If $\eta = 0$ or $\mathbf{s}' = \mathbf{0}$, the claim is immediate. Thus, assume $\eta > 0$ and $\mathbf{s}' \neq \mathbf{0}$.

Denote

$$\mathbf{z}_t := \mathbf{x} + t\eta\mathbf{s}', \qquad \mathbf{q}(t) := \nabla f(\mathbf{z}_t), \qquad Q(t) := \|\mathbf{q}(t) - \mathbf{q}(0)\|_1,$$

and define

$$\mathbf{D}_t := \operatorname{diag}(\mathbf{l}_0 + \mathbf{l}_1 \odot |\mathbf{q}(t)|), \qquad t \in [0,1].$$

Since

$$\|\mathbf{z}_t - \mathbf{x}\|_\infty \le t\eta \|\mathbf{s}'\|_\infty \le \eta(1+c) \le \frac{1}{\|\mathbf{l}_1\|_\infty},$$

whenever $\|\mathbf{l}_1\|_\infty > 0$, Assumption 2a gives, at every point on the segment where f is twice-differentiable,

$$-\mathbf{D}_t \preceq \nabla^2 f(\mathbf{z}_t) \preceq \mathbf{D}_t.$$

The same display is also the corresponding bound when $\|\mathbf{l}_1\|_\infty = 0$.

Since $\mathbf{q}'(t) = \eta\nabla^2 f(\mathbf{z}_t)\mathbf{s}'$ for almost every $t \in [0,1]$, we bound $\left\|\nabla^2 f(\mathbf{z}_t)\mathbf{s}'\right\|_1$ as follows. Because $\mathbf{s}'/\|\mathbf{s}'\|_\infty \in [-1,1]^d$, and because a linear function over $[-1,1]^d$ is maximized at a sign vector,

$$\begin{aligned}\left\|\nabla^2 f(\mathbf{z}_t)\mathbf{s}'\right\|_1 &= \|\mathbf{s}'\|_\infty \max_{\mathbf{u}\in\{-1,1\}^d} \left\langle \mathbf{u}, \nabla^2 f(\mathbf{z}_t)\frac{\mathbf{s}'}{\|\mathbf{s}'\|_\infty}\right\rangle \\ &\le \|\mathbf{s}'\|_\infty \max_{\mathbf{u},\tilde{\mathbf{s}}\in\{-1,1\}^d} \left\langle \mathbf{u}, \nabla^2 f(\mathbf{z}_t)\tilde{\mathbf{s}}\right\rangle.\end{aligned}$$

We next show that, for any $\mathbf{u}, \tilde{\mathbf{s}} \in \{-1,1\}^d$,

$$\left\langle \mathbf{u}, \nabla^2 f(\mathbf{z}_t)\tilde{\mathbf{s}}\right\rangle \le \operatorname{tr}(\mathbf{D}_t).$$

Indeed, set

$$\mathbf{a} := \frac{\mathbf{u} + \tilde{\mathbf{s}}}{2}, \qquad \mathbf{b} := \frac{\mathbf{u} - \tilde{\mathbf{s}}}{2}.$$

Then $\mathbf{u} = \mathbf{a} + \mathbf{b}$, $\tilde{\mathbf{s}} = \mathbf{a} - \mathbf{b}$, and $\mathbf{a}_i^2 + \mathbf{b}_i^2 = 1$ for every coordinate i. By symmetry of the Hessian,

$$\begin{aligned}\left\langle \mathbf{u}, \nabla^2 f(\mathbf{z}_t)\tilde{\mathbf{s}}\right\rangle &= \left\langle \mathbf{a}+\mathbf{b}, \nabla^2 f(\mathbf{z}_t)(\mathbf{a}-\mathbf{b})\right\rangle \\ &= \left\langle \mathbf{a}, \nabla^2 f(\mathbf{z}_t)\mathbf{a}\right\rangle - \left\langle \mathbf{b}, \nabla^2 f(\mathbf{z}_t)\mathbf{b}\right\rangle \\ &\le \langle \mathbf{a}, \mathbf{D}_t\mathbf{a}\rangle + \langle \mathbf{b}, \mathbf{D}_t\mathbf{b}\rangle \\ &= \sum_{i=1}^d (\mathbf{D}_t)_{ii}\left(\mathbf{a}_i^2 + \mathbf{b}_i^2\right) = \operatorname{tr}(\mathbf{D}_t).\end{aligned}$$

Therefore, for almost every $t \in [0,1]$,

$$\|\mathbf{q}'(t)\|_1 \le \eta \|\mathbf{s}'\|_\infty \operatorname{tr}(\mathbf{D}_t) \le \eta(1+c)\operatorname{tr}(\mathbf{D}_t).$$

Moreover,

$$\begin{aligned}\operatorname{tr}(\mathbf{D}_t) &= \|\mathbf{l}_0\|_1 + \langle \mathbf{l}_1, |\mathbf{q}(t)|\rangle \\ &\le \|\mathbf{l}_0\|_1 + \langle \mathbf{l}_1, |\mathbf{q}(0)|\rangle + \|\mathbf{l}_1\|_\infty \|\mathbf{q}(t) - \mathbf{q}(0)\|_1 \\ &= \|\mathbf{l}_0\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x})|\rangle + \|\mathbf{l}_1\|_\infty Q(t),\end{aligned}$$

where we used $||a| - |b|| \le |a - b|$ coordinatewise. Hence, for every $t \in [0,1]$,

$$Q(t) \le \int_0^t \|\mathbf{q}'(\tau)\|_1 \,\mathrm{d}\tau \le \eta(1+c)t\left(\|\mathbf{l}_0\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x})|\rangle\right) + \eta(1+c)\|\mathbf{l}_1\|_\infty \int_0^t Q(\tau)\mathrm{d}\tau.$$

Applying Grönwall's inequality yields

$$Q(t) \le \eta(1+c)\left(\|\mathbf{l}_0\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x})|\rangle\right) \int_0^t e^{\eta(1+c)\|\mathbf{l}_1\|_\infty (t-\tau)}\mathrm{d}\tau.$$

If $\|\mathbf{l}_1\|_\infty > 0$, then taking $t = 1$ gives

$$\begin{aligned} Q(1) &\leq (\|\mathbf{l}_0\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x})| \rangle) \frac{e^{\eta(1+c)\|\mathbf{l}_1\|_\infty} - 1}{\|\mathbf{l}_1\|_\infty} \\ &\leq \eta(1+c)(e-1)\left(\|\mathbf{l}_0\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x})| \rangle\right) \\ &\leq (2+2c)\eta\left(\|\mathbf{l}_0\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x})| \rangle\right), \end{aligned}$$

where we used $\eta(1+c)\|\mathbf{l}_1\|_\infty \leq 1$ and $e - 1 \leq 2$. If $\|\mathbf{l}_1\|_\infty = 0$, the same conclusion follows directly from the preceding integral inequality with the last term equal to zero. Since

$$Q(1) = \|\nabla f(\mathbf{x}') - \nabla f(\mathbf{x})\|_1,$$

the proof is complete. □

D.3 Proof of Theorem 1

For any $t \in [T]$, we have $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_\infty = \eta \|\text{sign}(\mathbf{m}_t)\|_\infty \leq 1/\|\mathbf{l}_1\|_\infty$ by the choice of $\eta \leq 1/\|\mathbf{l}_1\|_\infty$. Thus, under Assumption 2a, it holds that

$$\begin{aligned} f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq& \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2_{\mathbf{l}_0 + \mathbf{l}_1 \odot |\nabla f(\mathbf{x}_t)|} \\ =& \langle \nabla f(\mathbf{x}_t), -\eta \,\text{sign}(\mathbf{m}_t) \rangle + \frac{1}{2}\sum_{i=1}^d \left(\mathbf{l}_0 + \mathbf{l}_1 |\nabla_i f(\mathbf{x}_t)|\right)\left(\eta \,\text{sign}(\mathbf{m}_{t,i})\right)^2 \\ =& \eta \langle \nabla f(\mathbf{x}_t), -\text{sign}(\nabla f(\mathbf{x}_t)) \rangle + \eta \langle \nabla f(\mathbf{x}_t), \text{sign}(\nabla f(\mathbf{x}_t)) - \text{sign}(\mathbf{m}_t) \rangle \\ &+ \frac{\eta^2 \|\mathbf{l}_0\|_1}{2} + \frac{\eta^2 \langle \mathbf{l}_1, |\nabla f(\mathbf{x}_t)| \rangle}{2} \\ \leq& -\eta \|\nabla f(\mathbf{x}_t)\|_1 + 2\eta \|\mathbf{m}_t - \nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 \|\mathbf{l}_0\|_1}{2} + \frac{\eta^2 \|\mathbf{l}_1\|_\infty \|\nabla f(\mathbf{x}_t)\|_1}{2} \\ \leq& -\frac{\eta}{2}\|\nabla f(\mathbf{x}_t)\|_1 + 2\eta \|\mathbf{m}_t - \nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 \|\mathbf{l}_0\|_1}{2}, \end{aligned}$$

where the second inequality is due to Lemma D.3; the last step utilizes $\eta \leq 1/\|\mathbf{l}_1\|_\infty$. Rearranging the above relation and summing up yields

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1\right] \leq \frac{2\Delta_f}{\eta T} + 4\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \|\mathbf{m}_t - \nabla f(\mathbf{x}_t)\|_1\right] + \eta \|\mathbf{l}_0\|_1, \tag{13}$$

where we use $\Delta_f = f(\mathbf{x}_1) - f_* \geq f(\mathbf{x}_1) - f(\mathbf{x}_{T+1})$. Next, we proceed to bound the deviation between the momentum and the true gradient [159]. By Equation (11), we have

$$\begin{aligned} \boldsymbol{\epsilon}_t &= \mathbf{m}_t - \nabla f(\mathbf{x}_t) = \beta \mathbf{m}_{t-1} + (1-\beta)\mathbf{g}_t - \nabla f(\mathbf{x}_t) \\ &= \beta\left(\mathbf{m}_{t-1} - \nabla f(\mathbf{x}_{t-1})\right) + \beta\left(\nabla f(\mathbf{x}_{t-1}) - \nabla f(\mathbf{x}_t)\right) + (1-\beta)\left(\mathbf{g}_t - \nabla f(\mathbf{x}_t)\right) \\ &= \beta \boldsymbol{\epsilon}_{t-1} + \beta \mathbf{s}_t + (1-\beta)\mathbf{n}_t. \end{aligned} \tag{14}$$

Applying the above relation recursively yields

$$\boldsymbol{\epsilon}_t = \beta^{t-1}\mathbf{n}_1 + (1-\beta)\sum_{k=2}^t \beta^{t-k}\mathbf{n}_k + \sum_{k=2}^t \beta^{t-k+1}\mathbf{s}_k,$$

where we utilize $\boldsymbol{\epsilon}_1 = \mathbf{m}_1 - \nabla f(\mathbf{x}_1) = \mathbf{g}_1 - \nabla f(\mathbf{x}_1) = \mathbf{n}_1$. Next, we decompose $\boldsymbol{\epsilon}_t$ into

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|_1\right] \leq \underbrace{\mathbb{E}\left[\left\|\beta^{t-1}\mathbf{n}_1\right\|_1\right]}_{\mathtt{A}_t} + \underbrace{\mathbb{E}\left[\left\|(1-\beta)\sum_{k=2}^t \beta^{t-k}\mathbf{n}_k\right\|_1\right]}_{\mathtt{B}_t} + \underbrace{\mathbb{E}\left[\left\|\sum_{k=2}^t \beta^{t-k+1}\mathbf{s}_k\right\|_1\right]}_{\mathtt{C}_t}, \tag{15}$$

and bound these terms separately.

Initial noise $\mathtt{A}_t$ We have

$$\begin{aligned}
\mathtt{A}_t =&\beta^{t-1}\sum_{i=1}^{d}\mathbb{E}\left[|\mathbf{n}_{1,i}|\right] \overset{\text{Lemma D.2}}{\leq} \beta^{t-1}\sum_{i=1}^{d}\left(\mathbb{E}\left[|\mathbf{n}_{1,i}|^p\right]\right)^{1/p}\\
\overset{\text{Lemma D.5}}{\leq}& 2\beta^{t-1}\sum_{i=1}^{d}\left(B^{1-p}\left(\boldsymbol{\sigma}_{0,i}^p+\boldsymbol{\sigma}_{1,i}^p\left|\nabla_i f(\mathbf{x}_1)\right|^p\right)\right)^{1/p}\\
\overset{\text{Lemma D.1}}{\leq}& 2\beta^{t-1}\sum_{i=1}^{d}B^{\frac{1-p}{p}}\left(\boldsymbol{\sigma}_{0,i}+\boldsymbol{\sigma}_{1,i}\left|\nabla_i f(\mathbf{x}_1)\right|\right)\\
\leq& 2\beta^{t-1}B^{\frac{1-p}{p}}\left(\left\|\boldsymbol{\sigma}_0\right\|_1+\left\|\boldsymbol{\sigma}_1\right\|_\infty\left\|\nabla f(\mathbf{x}_1)\right\|_1\right).
\end{aligned} \tag{16}$$

Cumulative noise $\mathtt{B}_t$ Define $\mathbf{v}_k:=\beta^{t-k}\mathbf{n}_k, k\in[2,t]$. It holds that

$$\frac{\mathtt{B}_t}{1-\beta}\leq\mathbb{E}\left[\left\|\sum_{k=2}^{t}\mathbf{v}_k\right\|_1\right]\overset{\text{Lemma D.7}}{\leq}2\sqrt{2}\sum_{i=1}^{d}\mathbb{E}\left[\sum_{k=2}^{t}|\mathbf{v}_{k,i}|^p\right]^{1/p}. \tag{17}$$

Here, we cannot directly bound the above noise terms using similar procedures as in $\mathtt{A}_t$ due to technical reasons. As an alternative, we recursively expand the conditional expectation.

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{k=2}^{t}|\mathbf{v}_{k,i}|^p\right)^{1/p}\middle|\mathcal{F}_{t-1}\right]\overset{\text{Lemma D.2}}{\leq}\left(\sum_{k=2}^{t}\mathbb{E}\left[\left|\mathbf{v}_{t,i}\right|^p\middle|\mathcal{F}_{t-1}\right]\right)^{1/p}\\
=&\left(\sum_{k=2}^{t-1}|\mathbf{v}_{k,i}|^p+\mathbb{E}\left[\left|\mathbf{v}_{t,i}\right|^p\middle|\mathcal{F}_{t-1}\right]\right)^{1/p}\\
\overset{\text{Lemma D.5}}{\leq}&\left(\sum_{k=2}^{t-1}|\mathbf{v}_{k,i}|^p+2\beta^{p(t-t)}B^{1-p}\left(\boldsymbol{\sigma}_{0,i}^p+\boldsymbol{\sigma}_{1,i}^p\left|\nabla_i f(\mathbf{x}_t)\right|^p\right)\right)^{1/p}\\
\overset{\text{Lemma D.1}}{\leq}&\left(\sum_{k=2}^{t-1}|\mathbf{v}_{k,i}|^p+2\beta^{p(t-t)}B^{1-p}\boldsymbol{\sigma}_{0,i}^p\right)^{1/p}+2^{\frac{1}{p}}\beta^{t-t}B^{\frac{1-p}{p}}\boldsymbol{\sigma}_{1,i}\left|\nabla_i f(\mathbf{x}_t)\right|.
\end{aligned} \tag{18}$$

Taking total expectations on both sides and applying the relation in Equation (18) recursively:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=2}^{t}|\mathbf{v}_{k,i}|^p\right]^{1/p}\\
\leq&\mathbb{E}\left[\sum_{k=2}^{t-1}|\mathbf{v}_{k,i}|^p+2\beta^{p(t-t)}B^{1-p}\boldsymbol{\sigma}_{0,i}^p\right]^{1/p}+2^{\frac{1}{p}}\beta^{t-t}B^{\frac{1-p}{p}}\boldsymbol{\sigma}_{1,i}\mathbb{E}\left[\left|\nabla_i f(\mathbf{x}_t)\right|\right]\\
\overset{\text{Lemma D.2}}{\leq}&\left(\mathbb{E}\left[\sum_{k=2}^{t-1}|\mathbf{v}_{k,i}|^p+2\beta^{p(t-t)}B^{1-p}\boldsymbol{\sigma}_{0,i}^p\right]\right)^{1/p}+2^{\frac{1}{p}}\beta^{t-t}B^{\frac{1-p}{p}}\boldsymbol{\sigma}_{1,i}\mathbb{E}\left[\left|\nabla_i f(\mathbf{x}_t)\right|\right]\\
=&\left(\mathbb{E}\left[\mathbb{E}\left[\sum_{k=2}^{t-1}|\mathbf{v}_{k,i}|^p\middle|\mathcal{F}_{t-2}\right]\right]+2\beta^{p(t-t)}B^{1-p}\boldsymbol{\sigma}_{0,i}^p\right)^{1/p}+2^{\frac{1}{p}}\beta^{t-t}B^{\frac{1-p}{p}}\boldsymbol{\sigma}_{1,i}\mathbb{E}\left[\left|\nabla_i f(\mathbf{x}_t)\right|\right]\\
\leq&\left(\mathbb{E}\left[\mathbb{E}\left[\sum_{k=2}^{t-2}|\mathbf{v}_{k,i}|^p\middle|\mathcal{F}_{t-3}\right]\right]+2\beta^{p(t-(t-1))}B^{1-p}\boldsymbol{\sigma}_{0,i}^p+2\beta^{p(t-t)}B^{1-p}\boldsymbol{\sigma}_{0,i}^p\right)^{1/p}\\
&+2^{\frac{1}{p}}\beta^{t-(t-1)}B^{\frac{1-p}{p}}\boldsymbol{\sigma}_{1,i}\mathbb{E}\left[\left|\nabla_i f(\mathbf{x}_{t-1})\right|\right]+2^{\frac{1}{p}}\beta^{t-t}B^{\frac{1-p}{p}}\boldsymbol{\sigma}_{1,i}\mathbb{E}\left[\left|\nabla_i f(\mathbf{x}_t)\right|\right]\\
\leq&\cdots\\
\leq&\left(\sum_{k=2}^{t}2\beta^{p(t-k)}B^{1-p}\boldsymbol{\sigma}_{0,i}^p\right)^{1/p}+\sum_{k=2}^{t}2^{\frac{1}{p}}\beta^{t-k}B^{\frac{1-p}{p}}\boldsymbol{\sigma}_{1,i}\mathbb{E}\left[\left|\nabla_i f(\mathbf{x}_k)\right|\right]
\end{aligned}$$

$$\leq \frac{2^{\frac{1}{p}} B^{\frac{1-p}{p}} \boldsymbol{\sigma}_{0,i}}{(1-\beta^p)^{1/p}} + 2^{\frac{1}{p}} B^{\frac{1-p}{p}} \sum_{k=2}^{t} \beta^{t-k} \boldsymbol{\sigma}_{1,i} \mathbb{E}\left[|\nabla_i f(\mathbf{x}_k)|\right].$$

Thus, combining with Equation (17), we conclude that

$$\begin{aligned}
\mathtt{B_t} \leq & 4\sqrt{2}(1-\beta) B^{\frac{1-p}{p}} \sum_{i=1}^{d} \left(\frac{\boldsymbol{\sigma}_{0,i}}{(1-\beta^p)^{1/p}} + \sum_{k=2}^{t} \beta^{t-k} \boldsymbol{\sigma}_{1,i} \mathbb{E}\left[|\nabla_i f(\mathbf{x}_k)|\right] \right) \\
\leq & 4\sqrt{2}(1-\beta) B^{\frac{1-p}{p}} \left(\frac{\|\boldsymbol{\sigma}_0\|_1}{(1-\beta)^{1/p}} + \sum_{k=2}^{t} \beta^{t-k} \|\boldsymbol{\sigma}_1\|_\infty \mathbb{E}\left[\|\nabla f(\mathbf{x}_k)\|_1\right] \right) \\
\leq & 4\sqrt{2} B^{\frac{1-p}{p}} (1-\beta)^{\frac{p-1}{p}} \|\boldsymbol{\sigma}_0\|_1 + 4\sqrt{2} B^{\frac{1-p}{p}} (1-\beta) \|\boldsymbol{\sigma}_1\|_\infty \sum_{k=2}^{t} \beta^{t-k} \mathbb{E}\left[\|\nabla f(\mathbf{x}_k)\|_1\right].
\end{aligned} \tag{19}$$

Trajectory curvature $\mathtt{C_t}$ Since $\{\mathbf{x}_t\}_{t\in[T]}$ are generated by sign-based updates, we can apply Lemma D.4 to bound $\|\mathbf{s}_k\|_1$:

$$\begin{aligned}
\mathtt{C_t} \leq & \mathbb{E}\left[\sum_{k=2}^{t} \beta^{t-k+1} \|\mathbf{s}_k\|_1 \right] \leq \sum_{k=2}^{t} \beta^{t-k+1} \cdot \mathbb{E}\left[2\eta \left(\|\mathbf{l}_0\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x}_k)| \rangle \right)\right] \\
\leq & \frac{2\eta\beta \|\mathbf{l}_0\|_1}{1-\beta} + \sum_{k=2}^{t} 2\eta\beta^{t-k+1} \|\mathbf{l}_1\|_\infty \mathbb{E}\left[\|\nabla f(\mathbf{x}_k)\|_1\right].
\end{aligned} \tag{20}$$

Combining the bounds for $\mathtt{A_t}, \mathtt{B_t}, \mathtt{C_t}$ (adding Equation (16), Equation (19) and Equation (20) together), we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|_1\right] \leq & \beta^{t-1} B^{\frac{1-p}{p}} \left(\|\boldsymbol{\sigma}_0\|_1 + \|\boldsymbol{\sigma}_1\|_\infty \|\nabla f(\mathbf{x}_1)\|_1 \right) \\
& + 4\sqrt{2} B^{\frac{1-p}{p}} (1-\beta)^{\frac{p-1}{p}} \|\boldsymbol{\sigma}_0\|_1 + 4\sqrt{2} B^{\frac{1-p}{p}} (1-\beta) \|\boldsymbol{\sigma}_1\|_\infty \sum_{k=2}^{t} \beta^{t-k} \mathbb{E}\left[\|\nabla f(\mathbf{x}_k)\|_1\right] \\
& + \frac{2\eta\beta \|\mathbf{l}_0\|_1}{1-\beta} + \sum_{k=2}^{t} 2\eta\beta^{t-k+1} \|\mathbf{l}_1\|_\infty \mathbb{E}\left[\|\nabla f(\mathbf{x}_k)\|_1\right].
\end{aligned}$$

Summing from 1 to T:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t\|_1\right] \leq & \frac{B^{\frac{1-p}{p}}}{T(1-\beta)} \left(\|\boldsymbol{\sigma}_0\|_1 + \|\boldsymbol{\sigma}_1\|_\infty \|\nabla f(\mathbf{x}_1)\|_1 \right) + 4\sqrt{2} B^{\frac{1-p}{p}} (1-\beta)^{\frac{p-1}{p}} \|\boldsymbol{\sigma}_0\|_1 \\
& + 4\sqrt{2} B^{\frac{1-p}{p}} (1-\beta) \|\boldsymbol{\sigma}_1\|_\infty \sum_{t=1}^{T} \sum_{k=2}^{t} \frac{\beta^{t-k}}{T} \mathbb{E}\left[\|\nabla f(\mathbf{x}_k)\|_1\right] \\
& + \frac{2\eta\beta \|\mathbf{l}_0\|_1}{1-\beta} + 2\eta \|\mathbf{l}_1\|_\infty \sum_{t=1}^{T} \sum_{k=2}^{t} \frac{\beta^{t-k+1}}{T} \mathbb{E}\left[\|\nabla f(\mathbf{x}_k)\|_1\right].
\end{aligned}$$

Note that

$$\sum_{t=1}^{T} \sum_{k=2}^{t} \beta^{t-k} \mathbb{E}\left[\|\nabla f(\mathbf{x}_k)\|_1\right] = \sum_{t=2}^{T} \left(\sum_{k=0}^{T-t} \beta^k \right) \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|_1\right] \leq \sum_{t=2}^{T} \frac{\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|_1\right]}{(1-\beta)}. \tag{21}$$

Hence,

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|\boldsymbol{\epsilon}_t\right\|_1\right] \\
\leq&\frac{B^{\frac{1-p}{p}}}{T(1-\beta)}\left(\left\|\boldsymbol{\sigma}_0\right\|_1+\left\|\boldsymbol{\sigma}_1\right\|_\infty\left\|\nabla f(\mathbf{x}_1)\right\|_1\right)+4\sqrt{2}B^{\frac{1-p}{p}}(1-\beta)^{\frac{p-1}{p}}\left\|\boldsymbol{\sigma}_0\right\|_1+\frac{2\eta\beta\left\|\mathbf{l}_0\right\|_1}{1-\beta} \\
&+4\sqrt{2}B^{\frac{1-p}{p}}\left\|\boldsymbol{\sigma}_1\right\|_\infty\sum_{t=2}^{T}\frac{1}{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t)\right\|_1\right]+2\eta\left\|\mathbf{l}_1\right\|_\infty\sum_{t=2}^{T}\frac{\beta}{T(1-\beta)}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t)\right\|_1\right] \qquad (22)\\
=&4\sqrt{2}(1-\beta)^{\frac{p-1}{p}}B^{\frac{1-p}{p}}\left\|\boldsymbol{\sigma}_0\right\|_1+\frac{2B^{\frac{1-p}{p}}\left\|\boldsymbol{\sigma}_0\right\|_1+2B^{\frac{1-p}{p}}\left\|\boldsymbol{\sigma}_1\right\|_\infty\left\|\nabla f(\mathbf{x}_1)\right\|_1}{T(1-\beta)}+\frac{2\eta\beta\left\|\mathbf{l}_0\right\|_1}{1-\beta} \\
&+\sum_{t=2}^{T}\left(4\sqrt{2}B^{\frac{1-p}{p}}\left\|\boldsymbol{\sigma}_1\right\|_\infty+\frac{2\eta\left\|\mathbf{l}_1\right\|_\infty\beta}{1-\beta}\right)\frac{1}{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t)\right\|_1\right].
\end{aligned}
$$

By Equation (2), we deduce that

$$
4\sqrt{2}B^{\frac{1-p}{p}}\left\|\boldsymbol{\sigma}_1\right\|_\infty+\frac{2\eta\left\|\mathbf{l}_1\right\|_\infty\beta}{1-\beta}\leq\frac{1}{16}+\frac{1}{16}=\frac{1}{8}.
$$

Plugging the above relation into the last line of Equation (22) and recall Equation (13), we get

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla f(\mathbf{x}_t)\right\|_1\right]\leq&\frac{2\Delta_f}{\eta T}+\eta\left\|\mathbf{l}_0\right\|_1+\mathbb{E}\left[\frac{1}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t)\right\|_1\right]\right] \\
&+16\sqrt{2}(1-\beta)^{\frac{p-1}{p}}B^{\frac{1-p}{p}}\left\|\boldsymbol{\sigma}_0\right\|_1+\frac{8B^{\frac{1-p}{p}}\left(\left\|\boldsymbol{\sigma}_0\right\|_1+\left\|\boldsymbol{\sigma}_1\right\|_\infty\left\|\nabla f(\mathbf{x}_1)\right\|_1\right)}{T(1-\beta)}+\frac{8\eta\beta\left\|\mathbf{l}_0\right\|_1}{1-\beta},
\end{aligned}
$$

Rearranging the above relation yields

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla f(\mathbf{x}_t)\right\|_1\right]\leq&\frac{4\Delta_f}{\eta T}+\frac{18\eta\beta\left\|\mathbf{l}_0\right\|_1}{1-\beta}+32\sqrt{2}(1-\beta)^{\frac{p-1}{p}}B^{\frac{1-p}{p}}\left\|\boldsymbol{\sigma}_0\right\|_1 \\
&+\frac{16B^{\frac{1-p}{p}}\left(\left\|\boldsymbol{\sigma}_0\right\|_1+\left\|\boldsymbol{\sigma}_1\right\|_\infty\left\|\nabla f(\mathbf{x}_1)\right\|_1\right)}{T(1-\beta)}.
\end{aligned}
$$

In view of η in Equation (2), it holds that

$$
\begin{aligned}
\frac{4\Delta_f}{\eta T}+\frac{18\eta\beta\left\|\mathbf{l}_0\right\|_1}{1-\beta}\leq&12\sqrt{\frac{2\Delta_f\left\|\mathbf{l}_0\right\|_1}{T(1-\beta)}}+\frac{128\beta\Delta_f\left\|\mathbf{l}_1\right\|_\infty}{T(1-\beta)}+\frac{9\left\|\mathbf{l}_0\right\|_1}{16\left\|\mathbf{l}_1\right\|_\infty} \\
\leq&12\sqrt{\frac{2\Delta_f\left\|\mathbf{l}_0\right\|_1}{T(1-\beta)}}+\frac{256\beta\Delta_f\left\|\mathbf{l}_1\right\|_\infty}{T(1-\beta)} \qquad (23)\\
\leq&\frac{12\sqrt{2}(\Delta_f\left\|\mathbf{l}_0\right\|_1)^{\frac{p-1}{3p-2}}\left\|\boldsymbol{\sigma}_0\right\|_1^{\frac{p}{3p-2}}}{(BT)^{\frac{p-1}{3p-2}}}+\frac{256\Delta_f^{\frac{2p-2}{3p-2}}\left\|\boldsymbol{\sigma}_0\right\|_1^{\frac{2p}{3p-2}}\left\|\mathbf{l}_1\right\|_\infty}{\left\|\mathbf{l}_0\right\|_1^{\frac{p}{3p-2}}(BT)^{\frac{2p-2}{3p-2}}},
\end{aligned}
$$

where the second inequality holds when $\sqrt{\frac{2\Delta_f(1-\beta)}{9\beta\|\mathbf{l}_0\|_1 T}}\geq\frac{1-\beta}{32\beta\|\mathbf{l}_1\|_\infty}$. Similarly, the rest terms are bounded by

$$
\begin{aligned}
&32\sqrt{2}(1-\beta)^{\frac{p-1}{p}}B^{\frac{1-p}{p}}\left\|\boldsymbol{\sigma}_0\right\|_1+\frac{16B^{\frac{1-p}{p}}\left(\left\|\boldsymbol{\sigma}_0\right\|_1+\left\|\boldsymbol{\sigma}_1\right\|_\infty\left\|\nabla f(\mathbf{x}_1)\right\|_1\right)}{T(1-\beta)} \\
\leq&\frac{32\sqrt{2}(\Delta_f\left\|\mathbf{l}_0\right\|_1)^{\frac{p-1}{3p-2}}\left\|\boldsymbol{\sigma}_0\right\|_1^{\frac{p}{3p-2}}}{(BT)^{\frac{p-1}{3p-2}}}+\frac{16B^{\frac{1-p}{p}}\left(\left\|\boldsymbol{\sigma}_0\right\|_1+\left\|\boldsymbol{\sigma}_1\right\|_\infty\left\|\nabla f(\mathbf{x}_1)\right\|_1\right)\left\|\boldsymbol{\sigma}_0\right\|_1^{\frac{2p}{3p-2}}}{(\Delta_f\left\|\mathbf{l}_0\right\|_1)^{\frac{p}{3p-2}}(BT)^{\frac{2p-2}{3p-2}}} \qquad (24)
\end{aligned}
$$

Combining Equation (24) with Equation (23), we finally obtain

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla f(\mathbf{x}_t)\right\|_1\right] \leq \frac{44\sqrt{2}(\Delta_f \left\|\mathbf{l}_0\right\|_1)^{\frac{p-1}{3p-2}} \left\|\boldsymbol{\sigma}_0\right\|_1^{\frac{p}{3p-2}}}{(BT)^{\frac{p-1}{3p-2}}}$$
$$+ \frac{16B^{\frac{1-p}{p}}\left(\left\|\boldsymbol{\sigma}_0\right\|_1 + \left\|\boldsymbol{\sigma}_1\right\|_\infty \left\|\nabla f(\mathbf{x}_1)\right\|_1\right)\left\|\boldsymbol{\sigma}_0\right\|_1^{\frac{2p}{3p-2}}}{(\Delta_f \left\|\mathbf{l}_0\right\|_1)^{\frac{p}{3p-2}}(BT)^{\frac{2p-2}{3p-2}}} + \frac{256\Delta_f^{\frac{2p-2}{3p-2}}\left\|\boldsymbol{\sigma}_0\right\|_1^{\frac{2p}{3p-2}}\left\|\mathbf{l}_1\right\|_\infty}{\left\|\mathbf{l}_0\right\|_1^{\frac{p}{3p-2}}(BT)^{\frac{2p-2}{3p-2}}},$$

which implies a convergence rate of $O\left((\Delta_f \left\|\mathbf{l}_0\right\|_1)^{\frac{p-1}{3p-2}} \left\|\boldsymbol{\sigma}_0\right\|_1^{\frac{p}{3p-2}} (BT)^{\frac{1-p}{3p-2}}\right)$.

D.4 Proof of Theorem 2

By Lemma D.8, $\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_\infty \leq 5\eta/3$ holds for any $t \in [T]$ provided that $\left\|\mathbf{x}_1\right\|_\infty \leq 1/(3\lambda)$, which immediately implies $\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_\infty \leq 1/\left\|\mathbf{l}_1\right\|_\infty$ by the choice of η in Equation (3). Thus, under Assumption 2a, it holds that

$$\begin{aligned}
f(\mathbf{x}_{t+1}) \leq & f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1}-\mathbf{x}_t\rangle + \frac{1}{2}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|^2_{\mathbf{l}_0+\mathbf{l}_1\odot|\nabla f(\mathbf{x}_t)|} \\
\leq & f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), -\eta \operatorname{sign}(\mathbf{v}_t)\rangle + \langle \nabla f(\mathbf{x}_t), -\eta\lambda\mathbf{x}_t\rangle \\
& + \frac{1}{2}\left\|\mathbf{l}_0+\mathbf{l}_1\odot|\nabla f(\mathbf{x}_t)|\right\|_1\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_\infty^2 \\
\overset{\text{Lemma D.8}}{\leq} & f(\mathbf{x}_t) + \eta\left\langle \nabla f(\mathbf{x}_t), -\operatorname{sign}(\nabla f(\mathbf{x}_t))\right\rangle + \eta\left\langle \nabla f(\mathbf{x}_t), \operatorname{sign}(\nabla f(\mathbf{x}_t)) - \operatorname{sign}(\mathbf{v}_t)\right\rangle \\
& + \eta\lambda\left\|\mathbf{x}_t\right\|_\infty\left\|\nabla f(\mathbf{x}_t)\right\|_1 + \frac{25\eta^2}{9}\left(\left\|\mathbf{l}_0\right\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x}_t)|\rangle\right) \\
\overset{\text{Lemmas D.3 and D.8}}{\leq} & f(\mathbf{x}_t) - \eta\left\|\nabla f(\mathbf{x}_t)\right\|_1 + 2\eta\left\|\mathbf{v}_t-\nabla f(\mathbf{x}_t)\right\|_1 + \frac{2\eta}{3}\left\|\nabla f(\mathbf{x}_t)\right\|_1 \\
& + \frac{25\eta^2\left\|\mathbf{l}_0\right\|_1}{9} + \frac{25\eta^2\left\|\mathbf{l}_1\right\|_\infty\left\|\nabla f(\mathbf{x}_t)\right\|_1}{9} \\
\leq & f(\mathbf{x}_t) - \frac{\eta}{6}\left\|\nabla f(\mathbf{x}_t)\right\|_1 + 2\eta\left\|\mathbf{v}_t-\nabla f(\mathbf{x}_t)\right\|_1 + \frac{25\eta^2\left\|\mathbf{l}_0\right\|_1}{9},
\end{aligned}$$

where the last step utilizes $\eta \leq 3/(50\left\|\mathbf{l}_1\right\|_\infty)$. Similar to Equation (13), we arrive at

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t)\right\|_1\right] \leq \frac{6\Delta_f}{\eta T} + \frac{12}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\mathbf{v}_t-\nabla f(\mathbf{x}_t)\right\|_1\right] + \frac{50\eta\left\|\mathbf{l}_0\right\|_1}{3}. \tag{25}$$

Recall the notations in Equation (11) and apply the same decomposition as in Equation (14) to bound $\mathbb{E}\left[\left\|\mathbf{v}_t-\nabla f(\mathbf{x}_t)\right\|_1\right]$:

$$\begin{aligned}
& \mathbb{E}\left[\left\|\mathbf{v}_t-\nabla f(\mathbf{x}_t)\right\|_1\right] = \mathbb{E}\left[\left\|\beta_1\mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t - \nabla f(\mathbf{x}_t)\right\|_1\right] \\
= & \mathbb{E}\left[\left\|\beta_1(\mathbf{m}_{t-1}-\nabla f(\mathbf{x}_{t-1})) + (1-\beta_1)(\mathbf{g}_t-\nabla f(\mathbf{x}_t)) + \beta_1(\nabla f(\mathbf{x}_{t-1})-\nabla f(\mathbf{x}_t))\right\|_1\right] \\
\leq & \beta_1\mathbb{E}\left[\left\|\boldsymbol{\epsilon}_{t-1}\right\|_1\right] + (1-\beta_1)\mathbb{E}\left[\left\|\mathbf{n}_t\right\|_1\right] + \beta_1\mathbb{E}\left[\left\|\mathbf{s}_t\right\|_1\right],
\end{aligned} \tag{26}$$

which holds for all $t \geq 2$. The noise term $\mathbf{n}_t$ can be bounded by

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{n}_t\right\|_1\right] = & \sum_{i=1}^{d}\mathbb{E}\left[\mathbb{E}\left[\left|\mathbf{n}_{t,i}\right| \mid \mathcal{F}_{t-1}\right]\right] \overset{\text{Lemma D.2}}{\leq} \sum_{i=1}^{d}\mathbb{E}\left[\left(\mathbb{E}\left[\left|\mathbf{n}_{t,i}\right|^p \mid \mathcal{F}_{t-1}\right]\right)^{1/p}\right] \\
\overset{\text{Lemma D.5}}{\leq} & \sum_{i=1}^{d}\mathbb{E}\left[\left(2B^{1-p}\left(\boldsymbol{\sigma}_{0,i}^p + \boldsymbol{\sigma}_{1,i}^p\left|\nabla_i f(\mathbf{x}_t)\right|^p\right)\right)^{1/p}\right] \\
\overset{\text{Lemma D.1}}{\leq} & \sum_{i=1}^{d}\mathbb{E}\left[2^{\frac{1}{p}}B^{\frac{1-p}{p}}\left(\boldsymbol{\sigma}_{0,i} + \boldsymbol{\sigma}_{1,i}\left|\nabla_i f(\mathbf{x}_t)\right|\right)\right] \\
\leq & 2^{\frac{1}{p}}B^{\frac{1-p}{p}}\left(\left\|\boldsymbol{\sigma}_0\right\|_1 + \left\|\boldsymbol{\sigma}_1\right\|_\infty\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t)\right\|_1\right]\right).
\end{aligned}$$

To bound the trajectory curvature term $\mathbf{s}_t$, note that the trajectory of Algorithm 2 satisfies the conditions in Lemma D.9 with $c = 2/3$ by Lemma D.8. Therefore, we deduce that

$$\mathbb{E}\left[\|\mathbf{s}_t\|_1\right] \leq \mathbb{E}\left[\frac{10\eta}{3}\left(\|\mathbf{l}_0\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x})|\rangle\right)\right] \leq \frac{10\eta}{3}\|\mathbf{l}_0\|_1 + \frac{10\eta}{3}\|\mathbf{l}_1\|_\infty \mathbb{E}\left[\|\nabla f(\mathbf{x})\|_1\right].$$

Plugging the bounds for $\mathbb{E}\left[\|\mathbf{n}_t\|_1\right]$ and $\mathbb{E}\left[\|\mathbf{s}_t\|_1\right]$, then summing from 2 to T, we get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|_1\right] = \frac{1}{T}\mathbb{E}\left[\|\mathbf{v}_1 - \nabla f(\mathbf{x}_1)\|_1\right] + \frac{1}{T}\sum_{t=2}^{T}\mathbb{E}\left[\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|_1\right] \\
\leq &\frac{1}{T}\mathbb{E}\left[\|\mathbf{n}_1\|_1\right] + \frac{\beta_1}{T}\sum_{t=2}^{T}\mathbb{E}\left[\|\boldsymbol{\epsilon}_{t-1}\|_1\right] + \frac{(1-\beta_1)}{T}\sum_{t=2}^{T}\mathbb{E}\left[\|\mathbf{n}_t\|_1\right] + \frac{\beta_1}{T}\sum_{t=2}^{T}\mathbb{E}\left[\|\mathbf{s}_t\|_1\right] \\
\leq &\frac{\beta_1}{T}\sum_{t=1}^{T-1}\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|_1\right] + \frac{\beta_1}{T}\mathbb{E}\left[\|\mathbf{n}_1\|_1\right] + \frac{2(1-\beta_1)}{T}\sum_{t=1}^{T}B^{\frac{1-p}{p}}\left(\|\boldsymbol{\sigma}_0\|_1 + \|\boldsymbol{\sigma}_1\|_\infty \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|_1\right]\right) \\
&+ \frac{\beta_1}{T}\sum_{t=2}^{T}\left(\frac{10\eta}{3}\|\mathbf{l}_0\|_1 + \frac{10\eta}{3}\|\mathbf{l}_1\|_\infty \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|_1\right]\right) \\
\leq &\frac{\beta_1}{T}\sum_{t=1}^{T-1}\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|_1\right] + \frac{2\beta_1\left(\|\boldsymbol{\sigma}_0\|_1 + \|\boldsymbol{\sigma}_1\|_\infty \|\nabla f(\mathbf{x}_1)\|_1\right)}{B^{\frac{p-1}{p}}T} + \frac{2(1-\beta_1)\|\boldsymbol{\sigma}_0\|_1}{B^{\frac{p-1}{p}}} + \frac{10\eta\beta_1}{3}\|\mathbf{l}_0\|_1 \\
&+ \frac{2(1-\beta_1)\|\boldsymbol{\sigma}_1\|_\infty}{B^{\frac{p-1}{p}}T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|_1\right] + \frac{10\beta_1\eta\|\mathbf{l}_1\|_\infty}{3T}\sum_{t=2}^{T}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|_1\right],
\end{aligned} \tag{27}$$

where the first inequality is due to $\mathbf{v}_1 - \nabla f(\mathbf{x}_1) = \mathbf{g}_1 - \nabla f(\mathbf{x}_1) = \mathbf{n}_1$. Next, we proceed by bounding the error term $\frac{\beta_1}{T}\sum_{t=1}^{T-1}\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|_1\right]$. Following the same recipe as in Equation (14), we immediately obtain $\boldsymbol{\epsilon}_t = (1-\beta_2)\boldsymbol{\epsilon}_{t-1} + (1-\beta_2)\mathbf{s}_t + \beta_2\mathbf{n}_t$. Expanding this relation recursively yields

$$\boldsymbol{\epsilon}_t = \beta_2^{t-1}\mathbf{n}_1 + (1-\beta_2)\sum_{k=2}^{t}\beta_2^{t-k}\mathbf{n}_k + \sum_{k=2}^{t}\beta_2^{t-k+1}\mathbf{s}_k,$$

where we utilize $\boldsymbol{\epsilon}_1 = \mathbf{m}_1 - \nabla f(\mathbf{x}_1) = \mathbf{g}_1 - \nabla f(\mathbf{x}_1) = \mathbf{n}_1$. Next, we decompose $\boldsymbol{\epsilon}_t$ into

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|_1\right] \leq \underbrace{\mathbb{E}\left[\left\|\beta_2^{t-1}\mathbf{n}_1\right\|_1\right]}_{\mathtt{A}_t} + \underbrace{\mathbb{E}\left[\left\|(1-\beta_2)\sum_{k=2}^{t}\beta_2^{t-k}\mathbf{n}_k\right\|_1\right]}_{\mathtt{B}_t} + \underbrace{\mathbb{E}\left[\left\|\sum_{k=2}^{t}\beta_2^{t-k+1}\mathbf{s}_k\right\|_1\right]}_{\mathtt{C}_t},$$

which exhibits the almost identical form to that in Equation (15), with the only difference in β. Thus, we can safely follow the derivations in Appendix D.3 to bound $\mathtt{A}_t, \mathtt{B}_t, \mathtt{C}_t$. For $\mathtt{A}_t$ and $\mathtt{B}_t$, we can simply replace the β in Equation (16) and Equation (19) by β_2 to obtain

$$\begin{aligned}
&\mathtt{A}_t \leq \beta_2^{t-1}B^{\frac{1-p}{p}}\left(\|\boldsymbol{\sigma}_0\|_1 + \|\boldsymbol{\sigma}_1\|_\infty \|\nabla f(\mathbf{x}_1)\|_1\right), \\
&\mathtt{B}_t \leq 4\sqrt{2}B^{\frac{1-p}{p}}(1-\beta_2)^{\frac{p-1}{p}}\|\boldsymbol{\sigma}_0\|_1 + 4\sqrt{2}B^{\frac{1-p}{p}}(1-\beta_2)\|\boldsymbol{\sigma}_1\|_\infty \sum_{k=2}^{t}\beta_2^{t-k}\mathbb{E}\left[\|\nabla f(\mathbf{x}_k)\|_1\right].
\end{aligned}$$

For $\mathtt{C}_t$, we can not directly substitute β in Equation (20) with β_2 due to the presence of weight decay. Instead, we invoke Lemma D.9 to deduce

$$\begin{aligned}
\mathtt{C}_t \leq &\mathbb{E}\left[\sum_{k=2}^{t}\beta_2^{t-k+1}\|\mathbf{s}_k\|_1\right] \leq \sum_{k=2}^{t}\beta_2^{t-k+1}\cdot\mathbb{E}\left[\frac{10}{3}\eta\left(\|\mathbf{l}_0\|_1 + \langle \mathbf{l}_1, |\nabla f(\mathbf{x}_k)|\rangle\right)\right] \\
\leq &\frac{10\eta\beta_2\|\mathbf{l}_0\|_1}{3(1-\beta_2)} + \sum_{k=2}^{t}\frac{10}{3}\eta\beta_2^{t-k+1}\|\mathbf{l}_1\|_\infty \mathbb{E}\left[\|\nabla f(\mathbf{x}_k)\|_1\right].
\end{aligned}$$

Combining the bounds for $\mathtt{A_t}, \mathtt{B_t}, \mathtt{C_t}$ as below:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|_1\right] \\
\leq& 4\sqrt{2}(1-\beta_2)^{\frac{p-1}{p}}B^{\frac{1-p}{p}}\|\boldsymbol{\sigma}_0\|_1 + \frac{2B^{\frac{1-p}{p}}\|\boldsymbol{\sigma}_0\|_1 + 2B^{\frac{1-p}{p}}\|\boldsymbol{\sigma}_1\|_\infty\|\nabla f(\mathbf{x}_1)\|_1}{T(1-\beta_2)} + \frac{10\eta\beta_2\|\mathbf{l}_0\|_1}{3(1-\beta_2)} \\
&+ \sum_{t=2}^{T}\left(4\sqrt{2}B^{\frac{1-p}{p}}\|\boldsymbol{\sigma}_1\|_\infty + \frac{10\eta\|\mathbf{l}_1\|_\infty\beta_2}{3(1-\beta_2)}\right)\frac{1}{T}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|_1\right].
\end{aligned}
$$

Plugging the above relation into Equation (27) and simplify:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|_1\right] \\
\leq& 4\sqrt{2}\beta_1(1-\beta_2)^{\frac{p-1}{p}}B^{\frac{1-p}{p}}\|\boldsymbol{\sigma}_0\|_1 + \frac{2\beta_1\left(\|\boldsymbol{\sigma}_0\|_1 + \|\boldsymbol{\sigma}_1\|_\infty\|\nabla f(\mathbf{x}_1)\|_1\right)}{B^{\frac{p-1}{p}}T(1-\beta_2)} + \frac{10\eta\beta_1\beta_2\|\mathbf{l}_0\|_1}{3(1-\beta_2)} \\
&+ \sum_{t=2}^{T}\left(4\sqrt{2}B^{\frac{1-p}{p}}\|\boldsymbol{\sigma}_1\|_\infty + \frac{10\eta\|\mathbf{l}_1\|_\infty\beta_2}{3(1-\beta_2)}\right)\frac{\beta_1}{T}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|_1\right] \\
&+ \frac{2\beta_1\left(\|\boldsymbol{\sigma}_0\|_1 + \|\boldsymbol{\sigma}_1\|_\infty\|\nabla f(\mathbf{x}_1)\|_1\right)}{B^{\frac{p-1}{p}}T} + \frac{2(1-\beta_1)\|\boldsymbol{\sigma}_0\|_1}{B^{\frac{p-1}{p}}} + \frac{10\eta\beta_1}{3}\|\mathbf{l}_0\|_1 \\
&+ \frac{2(1-\beta_1)\|\boldsymbol{\sigma}_1\|_\infty}{B^{\frac{p-1}{p}}T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|_1\right] + \frac{10\beta_1\eta\|\mathbf{l}_1\|_\infty}{3T}\sum_{t=2}^{T}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|_1\right] \\
\leq& \frac{4\sqrt{2}\beta_1\|\boldsymbol{\sigma}_0\|_1}{(1-\beta_2)^{\frac{1-p}{p}}B^{\frac{p-1}{p}}} + \frac{2(1-\beta_1)\|\boldsymbol{\sigma}_0\|_1}{B^{\frac{p-1}{p}}} + \frac{4\beta_1\left(\|\boldsymbol{\sigma}_0\|_1 + \|\boldsymbol{\sigma}_1\|_\infty\|\nabla f(\mathbf{x}_1)\|_1\right)}{B^{\frac{p-1}{p}}T(1-\beta_2)} + \frac{10\eta\beta_1\|\mathbf{l}_0\|_1}{3(1-\beta_2)} \\
&+ \frac{1}{T}\sum_{t=1}^{T}\left(4\sqrt{2}B^{\frac{1-p}{p}}\|\boldsymbol{\sigma}_1\|_\infty + \frac{10\eta\beta_1\|\mathbf{l}_1\|_\infty}{3(1-\beta_2)}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|_1\right] \\
\leq& \frac{4\sqrt{2}\|\boldsymbol{\sigma}_0\|_1}{(1-\beta_2)^{\frac{1-p}{p}}B^{\frac{p-1}{p}}} + \frac{2(1-\beta_1)\|\boldsymbol{\sigma}_0\|_1}{B^{\frac{p-1}{p}}} + \frac{4\left(\|\boldsymbol{\sigma}_0\|_1 + \|\boldsymbol{\sigma}_1\|_\infty\|\nabla f(\mathbf{x}_1)\|_1\right)}{B^{\frac{p-1}{p}}T(1-\beta_2)} + \frac{10\eta\|\mathbf{l}_0\|_1}{3(1-\beta_2)} \\
&+ \sum_{t=1}^{T}\frac{1}{24T}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|_1\right],
\end{aligned}
$$

where the last inequality uses the hyperparameters in Equation (3), as depicted below:

$$
4\sqrt{2}B^{\frac{1-p}{p}}\|\boldsymbol{\sigma}_1\|_\infty + \frac{10\eta\beta_1\|\mathbf{l}_1\|_\infty}{3(1-\beta_2)} \leq \frac{1}{36} + \frac{1}{36} = \frac{1}{18}.
$$

Now, it suffices to combine the above error bound for $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|_1\right]$ with Equation (25), which gives

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|_1\right] \leq& \frac{12\Delta_f}{\eta T} + \frac{170\eta\|\mathbf{l}_0\|_1}{1-\beta_2} + \frac{144\sqrt{2}\|\boldsymbol{\sigma}_0\|_1}{(1-\beta_2)^{\frac{1-p}{p}}B^{\frac{p-1}{p}}} \\
&+ \frac{72(1-\beta_1)\|\boldsymbol{\sigma}_0\|_1}{B^{\frac{p-1}{p}}} + \frac{144\left(\|\boldsymbol{\sigma}_0\|_1 + \|\boldsymbol{\sigma}_1\|_\infty\|\nabla f(\mathbf{x}_1)\|_1\right)}{B^{\frac{p-1}{p}}T(1-\beta_2)}.
\end{aligned}
$$

Choosing $B, \beta_1, \beta_2, \eta$ according to Equation (3) yields the desired bound.

E Analysis for Muon and Muonlight

As established in Appendix C.4, Assumption 4b implies Assumption 4c. Accordingly, throughout this appendix, we maintain Assumption 4b as our operative assumption while utilizing the relation Equation (10) specified in Assumption 4c.

E.1 Matrix Calculus

Lemmas E.1 and E.2 are standard textbook results [179–181], which we omit the proof here for simplicity.

Lemma E.1 (Properties of trace and matrix norms)**.** *For any* $\mathbf{X} \in \mathbb{R}^{m\times n}, \mathbf{Y} \in \mathbb{R}^{n\times r}, \mathbf{L}, \mathbf{L}' \in \mathbb{S}^n$

1. $\operatorname{tr}(\mathbf{L}) = \|\mathbf{L}\|_*, \operatorname{tr}(|\mathbf{X}|_{\mathrm{m}}) = \operatorname{tr}\left(\left(\mathbf{X}\mathbf{X}^\top\right)^{1/2}\right) = \|\mathbf{X}\|_*$.
2. $\|\mathbf{X}\mathbf{Y}\|_* \le \min\left\{\|\mathbf{X}\|_{\mathrm{op}} \|\mathbf{Y}\|_*, \|\mathbf{X}\|_* \|\mathbf{Y}\|_{\mathrm{op}}\right\}$.

Lemma E.2 (Properties of the $\operatorname{msign}(\cdot)$ operator)**.** *For any* $\mathbf{X} \in \mathbb{R}^{m\times n}, \mathbf{L} \in \mathbb{S}^m, a > 0$

1. $\langle \mathbf{X}, \operatorname{msign}(\mathbf{X})\rangle = \|\mathbf{X}\|_*$.
2. $\operatorname{msign}(a\mathbf{X}) = \operatorname{msign}(\mathbf{X})$.
3. $\|\operatorname{msign}(\mathbf{X})\|_{\mathbf{L}}^2 = \operatorname{tr}\left(\left(\operatorname{msign}(\mathbf{X})\right)^\top \mathbf{L} \cdot \operatorname{msign}(\mathbf{X})\right) \le \operatorname{tr}(\mathbf{L})$.

Lemma E.3 (Matrix Extension of Lemma D.3)**.** *For any* $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}$*, it holds that*

$$\langle \mathbf{X}, \operatorname{msign}(\mathbf{X}) - \operatorname{msign}(\mathbf{Y})\rangle \le 2\|\mathbf{X} - \mathbf{Y}\|_*.$$

Proof. By Lemma E.2, we compute

$$\begin{aligned}\langle \mathbf{X}, \operatorname{msign}(\mathbf{X}) - \operatorname{msign}(\mathbf{Y})\rangle &= \langle \mathbf{X}, \operatorname{msign}(\mathbf{X})\rangle - \langle \mathbf{X}, \operatorname{msign}(\mathbf{Y})\rangle \\ &= \|\mathbf{X}\|_* - \|\mathbf{Y}\|_* + \langle \mathbf{Y} - \mathbf{X}, \operatorname{msign}(\mathbf{Y})\rangle \\ &\le \|\mathbf{X} - \mathbf{Y}\|_* + \|\mathbf{Y} - \mathbf{X}\|_* \|\operatorname{msign}(\mathbf{Y})\|_{\mathrm{op}} \le 2\|\mathbf{X} - \mathbf{Y}\|_*,\end{aligned}$$

where the first inequality uses the triangle inequality as well as the duality between nuclear norm and operator norm. □

Lemma E.4 (Matrix Cauchy-Schwarz Inequality, Lemma 8 in An et al. [64])**.** *For any* $\mathbf{X} \in \mathbb{R}^{m\times n}$ *and* $\mathbf{L} \in \mathbb{S}^m, \mathbf{L} \succ \mathbf{0}$*, it holds that*

$$\|\mathbf{X}\|_* \le \sqrt{\|\mathbf{L}\|_* \operatorname{tr}(\mathbf{X}^\top \mathbf{L}^{-1}\mathbf{X})} = \sqrt{\|\mathbf{L}\|_* \|\mathbf{X}\|_{\mathbf{L}^{-1}}^2}.$$

E.2 Technical Lemmas

Lemma E.5 (Descent Lemma)**.** *Under Assumption 2b, if* $\|\mathbf{X}' - \mathbf{X}\|_{\mathrm{op}} \le 1/\|\mathbf{L}_1\|_{\mathrm{op}}$*, then:*

$$f(\mathbf{X}') \le f(\mathbf{X}) + \langle \nabla f(\mathbf{X}), \mathbf{X}' - \mathbf{X}\rangle + \frac{1}{2}\|\mathbf{X}' - \mathbf{X}\|_{\mathbf{L}(\mathbf{X})}^2$$

Proof. Define $\mathbf{X}(s) = s\mathbf{X}' + (1-s)\mathbf{X}, \forall s \in [0,1]$, which implies $\|\mathbf{X}(s) - \mathbf{X}\|_{\mathrm{op}} = s\|\mathbf{X}' - \mathbf{X}\|_{\mathrm{op}} \le 1/\|\mathbf{L}_1\|_{\mathrm{op}}$. By Taylor expansion, we have

$$\begin{aligned}f(\mathbf{X}') =& f(\mathbf{X}) + \langle \nabla f(\mathbf{X}), \mathbf{X}' - \mathbf{X}\rangle + \int_0^1 \langle \nabla f(\mathbf{X}(s)) - \nabla f(\mathbf{X}), \mathbf{X}' - \mathbf{X}\rangle \,\mathrm{d}s \\ \le& f(\mathbf{X}) + \langle \nabla f(\mathbf{X}), \mathbf{X}' - \mathbf{X}\rangle + \int_0^1 \|\nabla f(\mathbf{X}(s)) - \nabla f(\mathbf{X})\|_{(\mathbf{L}(\mathbf{X}))^{-1}} \|\mathbf{X}' - \mathbf{X}\|_{\mathbf{L}(\mathbf{X})} \,\mathrm{d}s \\ \overset{\text{Assumption 2b}}{\le}& f(\mathbf{X}) + \langle \nabla f(\mathbf{X}), \mathbf{X}' - \mathbf{X}\rangle + \int_0^1 \|\mathbf{X}(s) - \mathbf{X}\|_{\mathbf{L}(\mathbf{X})} \|\mathbf{X}' - \mathbf{X}\|_{\mathbf{L}(\mathbf{X})} \,\mathrm{d}s \\ =& f(\mathbf{X}) + \langle \nabla f(\mathbf{X}), \mathbf{X}' - \mathbf{X}\rangle + \int_0^1 s\|\mathbf{X}' - \mathbf{X}\|_{\mathbf{L}(\mathbf{X})}^2 \,\mathrm{d}s \\ =& f(\mathbf{X}) + \langle \nabla f(\mathbf{X}), \mathbf{X}' - \mathbf{X}\rangle + \frac{1}{2}\|\mathbf{X}' - \mathbf{X}\|_{\mathbf{L}(\mathbf{X})}^2,\end{aligned}$$

where the first inequality is due to Cauchy-Schwarz inequality, and the fact that $\|\cdot\|_{\mathbf{L}(\mathbf{X})}$ and $\|\cdot\|_{(\mathbf{L}(\mathbf{X}))^{-1}}$ are primal-dual norm pairs [171, 64]. □

Lemma E.6 (Matrix Version of Lemma D.5)**.** *Under Assumptions 3b and 4b, the following holds for any $t \in [T]$:*

$$\mathbb{E}\left[\left\|\mathbf{N}_t\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}^p \middle| \mathcal{F}_{t-1}\right] \le 2B^{1-p}\left(\left\|\mathbf{V}_0\right\|_*^{p/2} + \frac{\left\|\mathbf{V}_1\right\|_{\mathrm{op}}^p \left\|\boldsymbol{\nabla}_t\right\|_*^p}{\left\|\mathbf{V}_0\right\|_*^{p/2}}\right),$$

where $\mathbf{N}_t$ is defined in Equation (29)*.*

Proof. Denote

$$\mathbf{N}_t^b := \mathbf{G}_t^b - \nabla f(\mathbf{X}_t), \qquad b \in [B].$$

Then

$$\mathbf{N}_t = \frac{1}{B}\sum_{b=1}^B \mathbf{N}_t^b.$$

By Assumption 3b, conditionally on $\mathcal{F}_{t-1}$, the mini-batch samples $\mathbf{G}_t^1, \ldots, \mathbf{G}_t^B$ are independent and unbiased estimators of $\nabla f(\mathbf{X}_t)$. Therefore, conditionally on $\mathcal{F}_{t-1}$, the centered noises $\mathbf{N}_t^1, \ldots, \mathbf{N}_t^B$ are independent and satisfy

$$\mathbb{E}\left[\mathbf{N}_t^b \middle| \mathcal{F}_{t-1}\right] = \mathbf{0}, \qquad b \in [B].$$

Equivalently, if we define the within-batch filtration

$$\mathcal{F}_{t,b} := \sigma\left(\mathcal{F}_{t-1}, \mathbf{N}_t^1, \ldots, \mathbf{N}_t^b\right), \qquad b = 0, 1, \ldots, B,$$

with $\mathcal{F}_{t,0} := \mathcal{F}_{t-1}$, then

$$\mathbb{E}\left[\mathbf{N}_t^b \middle| \mathcal{F}_{t,b-1}\right] = \mathbf{0}, \qquad b \in [B],$$

where we used the conditional independence and conditional unbiasedness from Assumption 3b. Thus $\left\{\mathbf{N}_t^b\right\}_{b=1}^B$ forms a martingale difference sequence with respect to the within-batch filtration $\{\mathcal{F}_{t,b}\}_{b=0}^B$.

Since $\mathbf{V}_0$ has full row rank by Assumption 4b, $|\mathbf{V}_0|_{\mathrm{m}}$ is positive definite, and $|\mathbf{V}_0|_{\mathrm{m}}^{-1/2}$ is well-defined. Define

$$\mathbf{Z}_t^b := |\mathbf{V}_0|_{\mathrm{m}}^{-1/2}\mathbf{N}_t^b, \qquad b \in [B].$$

Then

$$\left\|\mathbf{Z}_t^b\right\|_{\mathrm{F}} = \left\|\mathbf{N}_t^b\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}, \qquad \left\|\frac{1}{B}\sum_{b=1}^B \mathbf{Z}_t^b\right\|_{\mathrm{F}} = \left\|\mathbf{N}_t\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}.$$

After vectorization, let

$$\mathbf{z}_t^b := \mathrm{vec}(\mathbf{Z}_t^b) \in \mathbb{R}^{mn}.$$

Then $\left\{\mathbf{z}_t^b\right\}_{b=1}^B$ is also a martingale difference sequence with respect to $\{\mathcal{F}_{t,b}\}_{b=0}^B$, and

$$\left\|\mathbf{z}_t^b\right\|_2 = \left\|\mathbf{N}_t^b\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}, \qquad \left\|\frac{1}{B}\sum_{b=1}^B \mathbf{z}_t^b\right\|_2 = \left\|\mathbf{N}_t\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}.$$

Applying the vectorized von Bahr–Esseen inequality [58, Lemma 10] conditionally on $\mathcal{F}_{t-1}$ gives

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{N}_t\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}^p \middle| \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\left\|\frac{1}{B}\sum_{b=1}^B \mathbf{z}_t^b\right\|_2^p \middle| \mathcal{F}_{t-1}\right] \\
&\le \frac{2}{B^p}\sum_{b=1}^B \mathbb{E}\left[\left\|\mathbf{z}_t^b\right\|_2^p \middle| \mathcal{F}_{t-1}\right] \\
&= \frac{2}{B^p}\sum_{b=1}^B \mathbb{E}\left[\left\|\mathbf{N}_t^b\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}^p \middle| \mathcal{F}_{t-1}\right].
\end{aligned}$$

On the other hand, by Assumption 4b,

$$\mathbb{E}\left[\left\|\mathbf{N}_t^b\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}^p \middle| \mathcal{F}_{t-1}\right] \le \left\|\mathbf{V}_0\right\|_*^{p/2} + \frac{\left|\langle \mathbf{V}_1, \nabla f(\mathbf{X}_t)\rangle\right|^p}{\left\|\mathbf{V}_0\right\|_*^{p/2}}$$

$$\leq \|\mathbf{V}_0\|_*^{p/2} + \frac{\|\mathbf{V}_1\|_{\mathrm{op}}^p \|\nabla f(\mathbf{X}_t)\|_*^p}{\|\mathbf{V}_0\|_*^{p/2}}$$
$$= \|\mathbf{V}_0\|_*^{p/2} + \frac{\|\mathbf{V}_1\|_{\mathrm{op}}^p \|\boldsymbol{\nabla}_t\|_*^p}{\|\mathbf{V}_0\|_*^{p/2}},$$

where the second inequality uses the duality between the operator norm and the nuclear norm. Combining the preceding two displays yields

$$\mathbb{E}\left[\|\mathbf{N}_t\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}^p \middle| \mathcal{F}_{t-1}\right] \leq \frac{2}{B^p} \sum_{b=1}^{B} \left(\|\mathbf{V}_0\|_*^{p/2} + \frac{\|\mathbf{V}_1\|_{\mathrm{op}}^p \|\boldsymbol{\nabla}_t\|_*^p}{\|\mathbf{V}_0\|_*^{p/2}}\right)$$
$$= 2B^{1-p} \left(\|\mathbf{V}_0\|_*^{p/2} + \frac{\|\mathbf{V}_1\|_{\mathrm{op}}^p \|\boldsymbol{\nabla}_t\|_*^p}{\|\mathbf{V}_0\|_*^{p/2}}\right),$$

which completes the proof. □

The nuclear-norm martingale inequality used below can also be obtained from the non-commutative martingale framework of Pisier and Xu [182] via the standard block embedding. We retain the one-sided-Shampoo derivation as a self-contained proof tailored to the matrix geometry of Muon.

Lemma E.7 (Nuclear-norm martingale concentration)**.** *Let $\{\mathbf{G}_t\}_{t=1}^T \subset \mathbb{R}^{m\times n}$ be a matrix martingale difference sequence, i.e.,*

$$\mathbb{E}\left[\mathbf{G}_t \mid \mathcal{F}_{t-1}\right] = \mathbf{0}, \qquad \mathcal{F}_t = \sigma\left(\mathbf{G}_1, \ldots, \mathbf{G}_t\right).$$

Then

$$\mathbb{E}\left[\left\|\sum_{t=1}^T \mathbf{G}_t\right\|_*\right] \leq 2\sqrt{2}\,\mathbb{E}\left[\left\|\left(\sum_{t=1}^T \mathbf{G}_t\mathbf{G}_t^\top\right)^{1/2}\right\|_*\right].$$

Proof. This lemma is motivated by the regret analysis of the ASGO optimizer [64, 183]. Consider the closed convex set

$$\mathcal{W} := \left\{\mathbf{W} \in \mathbb{R}^{m\times n} : \|\mathbf{W}\|_{\mathrm{op}} \leq 1\right\}.$$

To avoid the possible singularity of the adaptive matrix, for any $\epsilon > 0$, define

$$\boldsymbol{\Lambda}_t^{(\epsilon)} := \left(\sum_{s=1}^t \mathbf{G}_s\mathbf{G}_s^\top\right)^{1/2} + \epsilon\mathbf{I}, \qquad \boldsymbol{\Lambda}_0^{(\epsilon)} := \epsilon\mathbf{I}.$$

Then $\boldsymbol{\Lambda}_t^{(\epsilon)} \succ \mathbf{0}$ for every $t \in [T]$. Imagine there is an online algorithm given by

$$\mathbf{W}_1^{(\epsilon)} = \mathbf{0}, \qquad \mathbf{W}_{t+1}^{(\epsilon)} = \Pi_{\mathcal{W}}^{\boldsymbol{\Lambda}_t^{(\epsilon)}}\left(\mathbf{W}_t^{(\epsilon)} - \eta\left(\boldsymbol{\Lambda}_t^{(\epsilon)}\right)^{-1}\mathbf{G}_t\right), \qquad t \in [T],$$

where the generalized projection operator [184, 9] is defined as

$$\Pi_{\mathcal{W}}^{\boldsymbol{\Lambda}}(\mathbf{W}) := \operatorname*{argmin}_{\mathbf{W}'\in\mathcal{W}} \|\mathbf{W} - \mathbf{W}'\|_{\boldsymbol{\Lambda}}^2.$$

Clearly, $\left\|\mathbf{W}_t^{(\epsilon)}\right\|_{\mathrm{op}} \leq 1$, and $\mathbf{W}_t^{(\epsilon)}$ only depends on $\mathbf{G}_1, \ldots, \mathbf{G}_{t-1}$. For any $\mathbf{W}_* \in \mathcal{W}$, the non-expansiveness of the generalized projection operator [184, Lemma 8] implies

$$\left\|\mathbf{W}_{t+1}^{(\epsilon)} - \mathbf{W}_*\right\|_{\boldsymbol{\Lambda}_t^{(\epsilon)}}^2 \leq \left\|\mathbf{W}_t^{(\epsilon)} - \eta\left(\boldsymbol{\Lambda}_t^{(\epsilon)}\right)^{-1}\mathbf{G}_t - \mathbf{W}_*\right\|_{\boldsymbol{\Lambda}_t^{(\epsilon)}}^2$$
$$= \left\|\mathbf{W}_t^{(\epsilon)} - \mathbf{W}_*\right\|_{\boldsymbol{\Lambda}_t^{(\epsilon)}}^2 + \eta^2 \left\|\mathbf{G}_t\right\|_{\left(\boldsymbol{\Lambda}_t^{(\epsilon)}\right)^{-1}}^2 - 2\eta\left\langle \mathbf{G}_t, \mathbf{W}_t^{(\epsilon)} - \mathbf{W}_*\right\rangle.$$

Rearranging and summing from 1 to T gives

$$\begin{aligned}
&\sum_{t=1}^{T}\left\langle \mathbf{G}_t, \mathbf{W}_t^{(\epsilon)}\right\rangle \\
\leq&\sum_{t=1}^{T}\left(\frac{\left\|\mathbf{W}_t^{(\epsilon)}-\mathbf{W}_*\right\|_{\mathbf{\Lambda}_t^{(\epsilon)}}^2-\left\|\mathbf{W}_{t+1}^{(\epsilon)}-\mathbf{W}_*\right\|_{\mathbf{\Lambda}_t^{(\epsilon)}}^2}{2\eta}+\frac{\eta}{2}\left\|\mathbf{G}_t\right\|_{\left(\mathbf{\Lambda}_t^{(\epsilon)}\right)^{-1}}^2+\left\langle \mathbf{G}_t, \mathbf{W}_*\right\rangle\right) \\
\leq&\sum_{t=1}^{T}\frac{\left\|\mathbf{W}_t^{(\epsilon)}-\mathbf{W}_*\right\|_{\mathbf{\Lambda}_t^{(\epsilon)}-\mathbf{\Lambda}_{t-1}^{(\epsilon)}}^2}{2\eta}+\frac{\left\|\mathbf{W}_1^{(\epsilon)}-\mathbf{W}_*\right\|_{\mathbf{\Lambda}_0^{(\epsilon)}}^2}{2\eta}+\frac{\eta}{2}\sum_{t=1}^{T}\left\|\mathbf{G}_t\right\|_{\left(\mathbf{\Lambda}_t^{(\epsilon)}\right)^{-1}}^2+\sum_{t=1}^{T}\left\langle \mathbf{G}_t, \mathbf{W}_*\right\rangle.
\end{aligned}$$

Compared with the unregularized proof, the only additional term is

$$\frac{\left\|\mathbf{W}_1^{(\epsilon)}-\mathbf{W}_*\right\|_{\mathbf{\Lambda}_0^{(\epsilon)}}^2}{2\eta}=\frac{\epsilon\left\|\mathbf{W}_*\right\|_{\mathrm{F}}^2}{2\eta},$$

where we used $\mathbf{W}_1^{(\epsilon)}=\mathbf{0}$ and $\mathbf{\Lambda}_0^{(\epsilon)}=\epsilon\mathbf{I}$.

We proceed by bounding the remaining two terms as in the original regret argument. First, since the matrix square-root map is operator monotone,

$$\mathbf{\Lambda}_t^{(\epsilon)}-\mathbf{\Lambda}_{t-1}^{(\epsilon)}=\left(\sum_{s=1}^{t}\mathbf{G}_s\mathbf{G}_s^\top\right)^{1/2}-\left(\sum_{s=1}^{t-1}\mathbf{G}_s\mathbf{G}_s^\top\right)^{1/2}\succeq\mathbf{0}.$$

Therefore,

$$\begin{aligned}
\sum_{t=1}^{T}\frac{\left\|\mathbf{W}_t^{(\epsilon)}-\mathbf{W}_*\right\|_{\mathbf{\Lambda}_t^{(\epsilon)}-\mathbf{\Lambda}_{t-1}^{(\epsilon)}}^2}{2\eta}=&\frac{1}{2\eta}\sum_{t=1}^{T}\left\langle\left(\mathbf{W}_t^{(\epsilon)}-\mathbf{W}_*\right)\left(\mathbf{W}_t^{(\epsilon)}-\mathbf{W}_*\right)^\top, \mathbf{\Lambda}_t^{(\epsilon)}-\mathbf{\Lambda}_{t-1}^{(\epsilon)}\right\rangle \\
\leq&\frac{1}{2\eta}\sum_{t=1}^{T}\left\|\left(\mathbf{W}_t^{(\epsilon)}-\mathbf{W}_*\right)\left(\mathbf{W}_t^{(\epsilon)}-\mathbf{W}_*\right)^\top\right\|_{\mathrm{op}}\left\|\mathbf{\Lambda}_t^{(\epsilon)}-\mathbf{\Lambda}_{t-1}^{(\epsilon)}\right\|_* \\
\leq&\frac{1}{2\eta}\sum_{t=1}^{T}\left\|\mathbf{W}_t^{(\epsilon)}-\mathbf{W}_*\right\|_{\mathrm{op}}^2\operatorname{tr}\left(\mathbf{\Lambda}_t^{(\epsilon)}-\mathbf{\Lambda}_{t-1}^{(\epsilon)}\right) \\
\leq&\frac{2}{\eta}\sum_{t=1}^{T}\operatorname{tr}\left(\mathbf{\Lambda}_t^{(\epsilon)}-\mathbf{\Lambda}_{t-1}^{(\epsilon)}\right) \\
=&\frac{2}{\eta}\left\|\left(\sum_{t=1}^{T}\mathbf{G}_t\mathbf{G}_t^\top\right)^{1/2}\right\|_*,
\end{aligned}$$

where we used $\left\|\mathbf{W}_t^{(\epsilon)}-\mathbf{W}_*\right\|_{\mathrm{op}}\leq\left\|\mathbf{W}_t^{(\epsilon)}\right\|_{\mathrm{op}}+\left\|\mathbf{W}_*\right\|_{\mathrm{op}}\leq 2$ and the fact that $\mathbf{\Lambda}_t^{(\epsilon)}-\mathbf{\Lambda}_{t-1}^{(\epsilon)}\succeq\mathbf{0}$.

For the second term, the matrix AdaGrad trace inequality [64, Eq. (9)] gives

$$\frac{\eta}{2}\sum_{t=1}^{T}\left\|\mathbf{G}_t\right\|_{\left(\mathbf{\Lambda}_t^{(\epsilon)}\right)^{-1}}^2\leq\eta\sum_{t=1}^{T}\operatorname{tr}\left(\mathbf{\Lambda}_t^{(\epsilon)}-\mathbf{\Lambda}_{t-1}^{(\epsilon)}\right)=\eta\left\|\left(\sum_{t=1}^{T}\mathbf{G}_t\mathbf{G}_t^\top\right)^{1/2}\right\|_*.$$

Combining the preceding bounds, we obtain

$$\sum_{t=1}^{T}\left\langle \mathbf{G}_t, \mathbf{W}_t^{(\epsilon)}\right\rangle\leq\left(\frac{2}{\eta}+\eta\right)\left\|\left(\sum_{t=1}^{T}\mathbf{G}_t\mathbf{G}_t^\top\right)^{1/2}\right\|_*+\sum_{t=1}^{T}\left\langle \mathbf{G}_t, \mathbf{W}_*\right\rangle+\frac{\epsilon\left\|\mathbf{W}_*\right\|_{\mathrm{F}}^2}{2\eta}.$$

Now choose

$$\mathbf{W}_* \in \operatorname*{argmin}_{\mathbf{W}\in\mathcal{W}} \left\langle \sum_{t=1}^{T} \mathbf{G}_t, \mathbf{W} \right\rangle.$$

By nuclear/operator norm duality,

$$\sum_{t=1}^{T} \langle \mathbf{G}_t, \mathbf{W}_* \rangle = -\left\| \sum_{t=1}^{T} \mathbf{G}_t \right\|_*.$$

Taking $\eta = \sqrt{2}$ yields

$$\sum_{t=1}^{T} \left\langle \mathbf{G}_t, \mathbf{W}_t^{(\epsilon)} \right\rangle \le 2\sqrt{2} \left\| \left(\sum_{t=1}^{T} \mathbf{G}_t \mathbf{G}_t^\top \right)^{1/2} \right\|_* - \left\| \sum_{t=1}^{T} \mathbf{G}_t \right\|_* + \frac{\epsilon \left\| \mathbf{W}_* \right\|_{\mathrm{F}}^2}{2\sqrt{2}}. \tag{28}$$

If $\{\mathbf{G}_t\}_{t=1}^T$ is a martingale difference sequence, then $\mathbf{W}_t^{(\epsilon)}$ is $\mathcal{F}_{t-1}$-measurable, and thus

$$\mathbb{E}\left[\left\langle \mathbf{G}_t, \mathbf{W}_t^{(\epsilon)} \right\rangle \right] = \mathbb{E}\left[\left\langle \mathbb{E}\left[\mathbf{G}_t | \mathcal{F}_{t-1} \right], \mathbf{W}_t^{(\epsilon)} \right\rangle \right] = 0.$$

Taking expectations in Equation (28) and rearranging gives

$$\mathbb{E}\left[\left\| \sum_{t=1}^{T} \mathbf{G}_t \right\|_* \right] \le 2\sqrt{2}\mathbb{E}\left[\left\| \left(\sum_{t=1}^{T} \mathbf{G}_t \mathbf{G}_t^\top \right)^{1/2} \right\|_* \right] + \frac{\epsilon}{2\sqrt{2}} \mathbb{E}\left[\left\| \mathbf{W}_* \right\|_{\mathrm{F}}^2 \right].$$

Since $\left\| \mathbf{W}_* \right\|_{\mathrm{op}} \le 1$, we have

$$\left\| \mathbf{W}_* \right\|_{\mathrm{F}}^2 \le \min\left\{ m, n \right\}.$$

Thus the last term vanishes by taking $\epsilon \downarrow 0$, and we conclude that

$$\mathbb{E}\left[\left\| \sum_{t=1}^{T} \mathbf{G}_t \right\|_* \right] \le 2\sqrt{2}\mathbb{E}\left[\left\| \left(\sum_{t=1}^{T} \mathbf{G}_t \mathbf{G}_t^\top \right)^{1/2} \right\|_* \right].$$

□

Lemma E.8 (Stability with weight decay, matrix version of Lemma D.8)**.** *Running Algorithm 4*

$$\lambda \le \frac{1}{\eta}\left(1 - \frac{1}{2^{1/T}} \right), \text{ and } \left\| \mathbf{X}_1 \right\|_{\mathrm{op}} \le \frac{a}{\lambda}, \quad \forall a \in (0,1),$$

ensures that

$$\left\| \mathbf{X}_t \right\|_{\mathrm{op}} \le \frac{a+1}{2\lambda}, \quad \left\| \mathbf{X}_{t+1} - \mathbf{X}_t \right\|_{\mathrm{op}} \le \frac{a+3}{2}\eta, \quad \forall t \in [T].$$

Proof. The proof of this lemma is analogous to that of Lemma D.8. Denote $1 - \eta\lambda$ by q, which implies $q \in (1/2^{1/T}, 1)$ as $\lambda \le \frac{1}{\eta}\left(1 - \frac{1}{2^{1/T}}\right)$. By the update rule of Algorithm 4, we have

$$\begin{aligned}
\left\| \mathbf{X}_t \right\|_{\mathrm{op}} &= \left\| (1-\eta\lambda)\mathbf{X}_{t-1} - \eta \operatorname{msign}\left(\widetilde{\mathbf{B}}_{t-1} \right) \right\|_{\mathrm{op}} \le q \left\| \mathbf{X}_{t-1} \right\|_{\mathrm{op}} + \eta \\
&\le q^2 \left\| \mathbf{X}_{t-2} \right\|_{\mathrm{op}} + q\eta + \eta \le \cdots \le q^{t-1} \left\| \mathbf{X}_1 \right\|_{\mathrm{op}} + \eta \sum_{i=0}^{t-2} q^i \le \frac{aq^{t-1}}{\lambda} + \frac{\eta(1-q^{t-1})}{1-q} \\
&= \frac{1-(1-a)q^{t-1}}{\lambda} \le \frac{1-(1-a)\left(\frac{1}{2}\right)^{\frac{t-1}{T}}}{\lambda} \le \frac{a+1}{2\lambda},
\end{aligned}$$

which holds for any $t \in [T]$. Therefore, we can deduce

$$\left\| \mathbf{X}_{t+1} - \mathbf{X}_t \right\|_{\mathrm{op}} = \left\| \eta\lambda\mathbf{X}_t + \eta \operatorname{msign}\left(\widetilde{\mathbf{B}}_t \right) \right\|_{\mathrm{op}} \le \eta\lambda \left\| \mathbf{X}_t \right\|_{\mathrm{op}} + \eta \le \frac{a+3}{2}\eta, \quad \forall t \in [T].$$

□

E.3 Proof of Theorem 3

Preliminaries Throughout this section, we define

$$\begin{aligned}&\boldsymbol{\nabla}_t := \nabla f(\mathbf{X}_t), \quad \mathbf{M}_t := (1-\beta)\mathbf{B}_t,\\ &\mathbf{E}_t := \mathbf{M}_t - \boldsymbol{\nabla}_t, \quad \mathbf{N}_t := \mathbf{G}_t - \boldsymbol{\nabla}_t, \quad \mathbf{S}_t := \boldsymbol{\nabla}_{t-1} - \boldsymbol{\nabla}_t.\end{aligned} \tag{29}$$

Under the above notation, the momentum term $\mathbf{B}_t = \beta\mathbf{B}_{t-1} + \mathbf{G}_t$ in Algorithm 3 can be rewritten as the one with a damping factor, i.e., $\mathbf{M}_t = \beta\mathbf{M}_{t-1} + (1-\beta)\mathbf{G}_t$, allowing us to borrow the similar derivations for Theorems 1 and 2. Also, under the exact Newton–Schulz oracle assumption, we have $\mathbf{O}_t = \texttt{NewtonSchulz}(\mathbf{B}_t) = \operatorname{msign}(\mathbf{B}_t)$. Since $\operatorname{msign}(\cdot)$ is invariant to input scale according to Lemma E.2, so it further implies $\mathbf{O}_t = \operatorname{msign}(\mathbf{M}_t)$. With the above preparations, we formally begin the proof.

By choosing $\eta \le 1/\left\|\mathbf{L}_1\right\|_{\text{op}}$, we can ensure $\left\|\mathbf{X}_{t+1} - \mathbf{X}_t\right\|_{\text{op}} = \eta\left\|\mathbf{O}_t\right\|_{\text{op}} \le 1/\left\|\mathbf{L}_1\right\|_{\text{op}}$. Recall the definitions and notations in Equation (29), according to Lemma E.5, we have

$$\begin{aligned}f(\mathbf{X}_{t+1}) \le& f(\mathbf{X}_t) - \eta\left\langle\boldsymbol{\nabla}_t, \operatorname{msign}(\mathbf{B}_t)\right\rangle + \frac{\eta^2}{2}\left\|\mathbf{O}_t\right\|_{\mathbf{L}(\mathbf{X}_t)}^2\\ \overset{\text{Lemma E.2}}{\le}& f(\mathbf{X}_t) - \eta\left\langle\boldsymbol{\nabla}_t, \operatorname{msign}(\boldsymbol{\nabla}_t)\right\rangle + \eta\left\langle\boldsymbol{\nabla}_t, \operatorname{msign}(\boldsymbol{\nabla}_t) - \operatorname{msign}(\mathbf{M}_t)\right\rangle\\ &+ \frac{\eta^2}{2}\operatorname{tr}\left(\left|\mathbf{L}_0\right|_{\text{m}} + \left|\boldsymbol{\nabla}_t\mathbf{L}_1^\top\right|_{\text{m}}\right)\\ \overset{\text{Lemmas E.1 to E.3}}{\le}& f(\mathbf{X}_t) - \eta\left\|\boldsymbol{\nabla}_t\right\|_* + 2\eta\left\|\boldsymbol{\nabla}_t - \mathbf{M}_t\right\|_* + \frac{\eta^2\left\|\mathbf{L}_0\right\|_*}{2} + \frac{\eta^2}{2}\left\|\mathbf{L}_1\right\|_{\text{op}}\left\|\boldsymbol{\nabla}_t\right\|_*\\ \le& f(\mathbf{X}_t) - \frac{\eta}{2}\left\|\boldsymbol{\nabla}_t\right\|_* + 2\eta\left\|\boldsymbol{\nabla}_t - \mathbf{M}_t\right\|_* + \frac{\eta^2\left\|\mathbf{L}_0\right\|_*}{2}.\end{aligned} \tag{30}$$

Rearranging the above relation and summing from 1 to T yields

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|\boldsymbol{\nabla}_t\right\|_*\right] \le \frac{2\Delta_f}{\eta T} + \eta\left\|\mathbf{L}_0\right\|_* + 4\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|\mathbf{M}_t - \boldsymbol{\nabla}_t\right\|_*\right], \tag{31}$$

where we use $\Delta_f = f(\mathbf{X}_1) - f_* \ge f(\mathbf{X}_1) - f(\mathbf{X}_{T+1})$. By Equation (29), we have that $\mathbf{M}_t = \beta\mathbf{M}_{t-1} + (1-\beta)\mathbf{G}_t$. Hence, we follow a similar expansion as in Equation (14) to get

$$\begin{aligned}\mathbf{E}_t = \mathbf{M}_t - \boldsymbol{\nabla}_t =& \beta\mathbf{M}_{t-1} + (1-\beta)\mathbf{G}_t - \boldsymbol{\nabla}_t\\ =& \beta\left(\mathbf{M}_{t-1} - \boldsymbol{\nabla}_{t-1}\right) + \beta\left(\boldsymbol{\nabla}_{t-1} - \boldsymbol{\nabla}_t\right) + (1-\beta)\left(\mathbf{G}_t - \boldsymbol{\nabla}_t\right)\\ =& \beta\mathbf{E}_{t-1} + \beta\mathbf{S}_t + (1-\beta)\mathbf{N}_t.\end{aligned}$$

which implies

$$\mathbf{E}_t = -\beta^t\boldsymbol{\nabla}_1 + (1-\beta)\sum_{k=1}^{t}\beta^{t-k}\mathbf{N}_k + \sum_{k=2}^{t}\beta^{t-k+1}\mathbf{S}_k,$$

where we utilize $\mathbf{E}_1 = \mathbf{M}_1 - \boldsymbol{\nabla}_1 = (1-\beta)\mathbf{G}_1 - \boldsymbol{\nabla}_1 = (1-\beta)\mathbf{N}_1 - \beta\boldsymbol{\nabla}_1$. We then decompose $\mathbf{E}_t$ into

$$\mathbb{E}\left[\left\|\mathbf{E}_t\right\|_*\right] \le \underbrace{\left\|\beta^t\boldsymbol{\nabla}_1\right\|_*}_{\mathtt{A}_t} + \underbrace{\mathbb{E}\left[\left\|(1-\beta)\sum_{k=1}^{t}\beta^{t-k}\mathbf{N}_k\right\|_*\right]}_{\mathtt{B}_t} + \underbrace{\mathbb{E}\left[\left\|\sum_{k=2}^{t}\beta^{t-k+1}\mathbf{S}_k\right\|_*\right]}_{\mathtt{C}_t},$$

and bound these terms separately.

Cumulative noise $\mathtt{B_t}$ Define $\widetilde{\mathbf{N}}_k := \beta^{t-k}\mathbf{N}_k, k \in [1,t]$. Under Assumption 3b, we can show that

$$\begin{aligned}
\frac{\mathtt{B_t}}{1-\beta} &\overset{\text{Lemma E.7}}{\leq} 2\sqrt{2}\mathbb{E}\left[\left\|\left(\sum_{k=1}^{t}\widetilde{\mathbf{N}}_k\widetilde{\mathbf{N}}_k^\top\right)^{1/2}\right\|_*\right] \\
&\overset{\text{Lemma E.4}}{\leq} 2\sqrt{2}\mathbb{E}\left[\sqrt{\left\|\left(\mathbf{V}_0\mathbf{V}_0^\top\right)^{1/2}\right\|_* \operatorname{tr}\left(\left(\sum_{k=1}^{t}\widetilde{\mathbf{N}}_k\widetilde{\mathbf{N}}_k^\top\right)^{1/2}|\mathbf{V}_0|_{\mathrm{m}}^{-1}\left(\sum_{k=1}^{t}\widetilde{\mathbf{N}}_k\widetilde{\mathbf{N}}_k^\top\right)^{1/2}\right)}\right] \qquad (32) \\
&= 2\sqrt{2\left\|\mathbf{V}_0\right\|_*}\cdot\mathbb{E}\left[\sqrt{\operatorname{tr}\left(\left(\sum_{k=1}^{t}\widetilde{\mathbf{N}}_k\widetilde{\mathbf{N}}_k^\top\right)|\mathbf{V}_0|_{\mathrm{m}}^{-1}\right)}\right].
\end{aligned}$$

We proceed by following the conditional expectation recursion in Equation (18):

$$\begin{aligned}
&\mathbb{E}\left[\sqrt{\operatorname{tr}\left(\left(\sum_{k=1}^{t}\widetilde{\mathbf{N}}_k\widetilde{\mathbf{N}}_k^\top\right)|\mathbf{V}_0|_{\mathrm{m}}^{-1}\right)}\,\middle|\,\mathcal{F}_{t-1}\right] \\
=&\mathbb{E}\left[\left(\left(\sum_{k=1}^{t}\operatorname{tr}\left(\widetilde{\mathbf{N}}_k\widetilde{\mathbf{N}}_k^\top|\mathbf{V}_0|_{\mathrm{m}}^{-1}\right)\right)^{p/2}\right)^{1/p}\,\middle|\,\mathcal{F}_{t-1}\right] \\
\overset{\text{Lemma D.2}}{\leq}&\left(\mathbb{E}\left[\left(\sum_{k=1}^{t}\left\|\widetilde{\mathbf{N}}_k\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}^2\right)^{p/2}\,\middle|\,\mathcal{F}_{t-1}\right]\right)^{1/p} \\
\overset{\text{Lemma D.1}}{\leq}&\left(\left(\sum_{k=1}^{t-1}\left\|\widetilde{\mathbf{N}}_k\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}^2\right)^{p/2}+\mathbb{E}\left[\left(\operatorname{tr}\left(\widetilde{\mathbf{N}}_t\widetilde{\mathbf{N}}_t^\top|\mathbf{V}_0|_{\mathrm{m}}^{-1}\right)\right)^{p/2}\,\middle|\,\mathcal{F}_{t-1}\right]\right)^{1/p} \\
\overset{\text{Lemma E.6}}{\leq}&\left(\left(\sum_{k=1}^{t-1}\left\|\widetilde{\mathbf{N}}_k\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}^2\right)^{p/2}+2B^{1-p}\beta^{p(t-t)}\left(\|\mathbf{V}_0\|_*^{p/2}+\frac{\|\mathbf{V}_1\|_{\mathrm{op}}^p\|\boldsymbol{\nabla}_t\|_*^p}{\|\mathbf{V}_0\|_*^{p/2}}\right)\right)^{1/p} \\
\overset{\text{Lemma D.1}}{\leq}&\left(\left(\sum_{k=1}^{t-1}\left\|\widetilde{\mathbf{N}}_k\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}^2\right)^{p/2}+2B^{1-p}\beta^{p(t-t)}\|\mathbf{V}_0\|_*^{p/2}\right)^{1/p}+\frac{2^{\frac{1}{p}}\beta^{(t-t)}\|\mathbf{V}_1\|_{\mathrm{op}}\|\boldsymbol{\nabla}_t\|_*}{B^{\frac{p-1}{p}}\|\mathbf{V}_0\|_*^{1/2}}.
\end{aligned}$$

Taking total expectations on both sides and applying the above relation recursively:

$$\begin{aligned}
&\mathbb{E}\left[\sqrt{\operatorname{tr}\left(\left(\sum_{k=1}^{t}\widetilde{\mathbf{N}}_k\widetilde{\mathbf{N}}_k^\top\right)|\mathbf{V}_0|_{\mathrm{m}}^{-1}\right)}\right] \\
&\leq \underbrace{\mathbb{E}\left[\frac{2\beta^{(t-k)}\|\mathbf{V}_1\|_{\mathrm{op}}\|\boldsymbol{\nabla}_k\|_*}{B^{\frac{p-1}{p}}\|\mathbf{V}_0\|_*^{1/2}}\right]\Bigg|_{k=t}}_{\Phi_k|_{k=t}} \\
&\quad+\mathbb{E}\left[\left(\left(\sum_{k=1}^{t-1}\left\|\widetilde{\mathbf{N}}_k\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}^2\right)^{p/2}+2B^{1-p}\beta^{p(t-t)}\|\mathbf{V}_0\|_*^{p/2}\right)^{1/p}\right] \\
&\overset{\text{Lemma D.1}}{\leq}\left(\mathbb{E}\left[\left(\sum_{k=1}^{t-1}\left\|\widetilde{\mathbf{N}}_k\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}^2\right)^{p/2}+2B^{1-p}\beta^{p(t-t)}\|\mathbf{V}_0\|_*^{p/2}\right]\right)^{1/p}+\Phi_t \\
&=\left(\mathbb{E}\left[\mathbb{E}\left[\left(\sum_{k=1}^{t-1}\left\|\widetilde{\mathbf{N}}_k\right\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}^2\right)^{p/2}\,\middle|\,\mathcal{F}_{t-2}\right]\right]+2B^{1-p}\beta^{p(t-t)}\|\mathbf{V}_0\|_*^{p/2}\right)^{1/p}+\Phi_t
\end{aligned}$$

$$
\begin{aligned}
&\leq \left(\mathbb{E}\left[\mathbb{E}\left[\left(\sum_{k=1}^{t-2} \left\| \widetilde{\mathbf{N}}_k \right\|^2_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}} \right)^{p/2} \middle| \mathcal{F}_{t-3} \right] \right] + 2B^{1-p} \left(\beta^{p(t-(t-1))} + \beta^{p(t-t)} \right) \|\mathbf{V}_0\|_*^{p/2} \right)^{1/p} \\
&\quad + \Phi_{t-1} + \Phi_t \leq \cdots \\
&\leq \left(2B^{1-p} \|\mathbf{V}_0\|_*^{p/2} \sum_{k=1}^{t} \beta^{p(t-k)} \right)^{1/p} + \sum_{k=1}^{t} \Phi_k \\
&\leq \frac{2\|\mathbf{V}_0\|_*^{p/2}}{B^{\frac{p-1}{p}}(1-\beta^p)^{1/p}} + \sum_{k=1}^{t} \mathbb{E}\left[\frac{2\beta^{t-k} \|\mathbf{V}_1\|_{\mathrm{op}} \|\boldsymbol{\nabla}_k\|_*}{B^{\frac{p-1}{p}} \|\mathbf{V}_0\|_*^{1/2}} \right].
\end{aligned}
$$

Hence, we can bound $\mathtt{B}_\mathtt{t}$ as

$$
\mathtt{B}_\mathtt{t} \leq \frac{4\sqrt{2}(1-\beta)^{\frac{p-1}{p}} \|\mathbf{V}_0\|_*}{B^{\frac{p-1}{p}}} + \frac{4\sqrt{2}(1-\beta)\|\mathbf{V}_1\|_{\mathrm{op}}}{B^{\frac{p-1}{p}}} \sum_{k=1}^{t} \mathbb{E}\left[\beta^{t-k} \|\boldsymbol{\nabla}_k\|_* \right], \tag{33}
$$

where we make use of $(1-\beta^p)^{-1/p} \leq (1-\beta)^{-1/p}$.

Trajectory curvature $\mathtt{C}_\mathtt{t}$ Under Assumption 2b with $\mathbf{L}(\mathbf{X}) = |\mathbf{L}_0|_{\mathrm{m}} + \left|\nabla f(\mathbf{X})\mathbf{L}_1^\top\right|_{\mathrm{m}}$, together with the fact that $\|\mathbf{X}_k - \mathbf{X}_{k-1}\|_{\mathrm{op}} = \eta \|\mathbf{O}_{k-1}\|_{\mathrm{op}} \leq 1/\|\mathbf{L}_1\|_{\mathrm{op}}$, we have

$$
\begin{aligned}
\mathtt{C}_\mathtt{t} &\leq \sum_{k=2}^{t} \beta^{t-k+1} \mathbb{E}\left[\|\mathbf{S}_k\|_* \right] \overset{\text{Lemma E.4}}{\leq} \sum_{k=2}^{t} \beta^{t-k+1} \mathbb{E}\left[\sqrt{\|\mathbf{L}(\mathbf{X}_k)\|_*} \|\mathbf{S}_k\|_{(\mathbf{L}(\mathbf{X}_k))^{-1}} \right] \\
&\leq \sum_{k=2}^{t} \beta^{t-k+1} \mathbb{E}\left[\sqrt{\mathrm{tr}\left(\mathbf{L}(\mathbf{X}_k)\right)} \|\eta \mathbf{O}_{k-1}\|_{\mathbf{L}(\mathbf{X}_k)} \right] \\
&\overset{\text{Lemma E.2}}{\leq} \eta \sum_{k=2}^{t} \beta^{t-k+1} \mathbb{E}\left[\mathrm{tr}\left(\mathbf{L}(\mathbf{X}_k)\right) \right] \\
&\overset{\text{Lemma E.1}}{\leq} \eta \sum_{k=2}^{t} \beta^{t-k+1} \mathrm{tr}\left(|\mathbf{L}_0|_{\mathrm{m}}\right) + \eta \sum_{k=2}^{t} \beta^{t-k+1} \mathbb{E}\left[\|\mathbf{L}_1\|_{\mathrm{op}} \|\boldsymbol{\nabla}_k\|_* \right] \\
&\leq \frac{\eta \|\mathbf{L}_0\|_*}{1-\beta} + \eta \|\mathbf{L}_1\|_{\mathrm{op}} \sum_{k=2}^{t} \beta^{t-k+1} \mathbb{E}\left[\|\boldsymbol{\nabla}_k\|_* \right].
\end{aligned}
$$

Combining the bounds for $\mathtt{A}_\mathtt{t}, \mathtt{B}_\mathtt{t}, \mathtt{C}_\mathtt{t}$, and summing from 1 to T:

$$
\begin{aligned}
&\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\|\mathbf{E}_t\|_* \right] \leq \frac{1}{T} \sum_{t=1}^{T} \beta^t \|\boldsymbol{\nabla}_1\|_* \\
&\quad + \frac{4\sqrt{2}(1-\beta)^{\frac{p-1}{p}} \|\mathbf{V}_0\|_*}{B^{\frac{p-1}{p}}} + \frac{4\sqrt{2}(1-\beta)\|\mathbf{V}_1\|_{\mathrm{op}}}{TB^{\frac{p-1}{p}}} \sum_{t=1}^{T} \sum_{k=1}^{t} \mathbb{E}\left[\beta^{(t-k)} \|\boldsymbol{\nabla}_k\|_* \right] \\
&\quad + \frac{\eta \|\mathbf{L}_0\|_*}{1-\beta} + \frac{\eta \|\mathbf{L}_1\|_{\mathrm{op}}}{T} \sum_{t=1}^{T} \sum_{k=2}^{t} \beta^{t-k+1} \mathbb{E}\left[\|\boldsymbol{\nabla}_k\|_* \right] \\
&\overset{\text{Equation (21)}}{\leq} \frac{\|\boldsymbol{\nabla}_1\|_*}{T(1-\beta)} + \frac{4\sqrt{2}(1-\beta)^{\frac{p-1}{p}} \|\mathbf{V}_0\|_*}{B^{\frac{p-1}{p}}} + \frac{\eta \|\mathbf{L}_0\|_*}{1-\beta} \\
&\quad + \frac{1}{T} \sum_{t=1}^{T} \left(\frac{4\sqrt{2}\|\mathbf{V}_1\|_{\mathrm{op}}}{B^{\frac{p-1}{p}}} + \frac{\eta\beta \|\mathbf{L}_1\|_{\mathrm{op}}}{1-\beta} \right) \mathbb{E}\left[\|\boldsymbol{\nabla}_t\|_* \right] \\
&\overset{\text{Equation (4)}}{\leq} \frac{\|\boldsymbol{\nabla}_1\|_*}{T(1-\beta)} + \frac{4\sqrt{2}(1-\beta)^{\frac{p-1}{p}} \|\mathbf{V}_0\|_*}{B^{\frac{p-1}{p}}} + \frac{\eta \|\mathbf{L}_0\|_*}{1-\beta} + \sum_{t=1}^{T} \frac{1}{8T} \mathbb{E}\left[\|\boldsymbol{\nabla}_t\|_* \right].
\end{aligned}
$$

Plugging the above relation into Equation (31), we obtain

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\|\boldsymbol{\nabla}_t\|_* \right] \leq \frac{4\Delta_f}{\eta T} + \frac{8\|\boldsymbol{\nabla}_1\|_*}{T(1-\beta)} + \frac{32\sqrt{2}(1-\beta)^{\frac{p-1}{p}} \|\mathbf{V}_0\|_*}{B^{\frac{p-1}{p}}} + \frac{10\eta \|\mathbf{L}_0\|_*}{1-\beta}.
$$

Finally, choosing B, β, η according to Equation (4) and following the similar steps as in Appendix D.3, we can obtain the convergence rate $O\left(\left(\Delta_f \left\|\mathbf{L}_0\right\|_*\right)^{\frac{p-1}{3p-2}} \left\|\mathbf{V}_0\right\|_*^{\frac{p}{3p-2}} (BT)^{\frac{1-p}{3p-2}}\right)$.

E.4 Proof of Theorem 4

Preliminaries For Algorithm 4, we define

$$\begin{aligned} &\boldsymbol{\nabla}_t := \nabla f(\mathbf{X}_t), \quad \mathbf{M}_t := (1-\beta_2)\mathbf{B}_t, \quad \widetilde{\mathbf{M}}_t := (1-\beta_2)\widetilde{\mathbf{B}}_t, \\ &\mathbf{E}_t := \mathbf{M}_t - \boldsymbol{\nabla}_t, \quad \mathbf{N}_t := \mathbf{G}_t - \boldsymbol{\nabla}_t, \quad \mathbf{S}_t := \boldsymbol{\nabla}_{t-1} - \boldsymbol{\nabla}_t. \end{aligned} \tag{34}$$

Based on the above notation, we have $\mathbf{M}_t = \beta_2 \mathbf{M}_{t-1} + (1-\beta_2)\mathbf{G}_t$ and $\widetilde{\mathbf{M}}_t = \beta_1 \mathbf{M}_t + (1-\beta_2)\mathbf{G}_t$. We also have $\mathbf{O}_t = \texttt{NewtonSchulz}(\widetilde{\mathbf{B}}_t) = \operatorname{msign}\left(\widetilde{\mathbf{M}}_t\right)$. Now, we begin the proof as follows.

By Lemma E.8 and $\left\|\mathbf{X}_1\right\|_{\text{op}} \le 1/(3\lambda)$, $\left\|\mathbf{X}_{t+1} - \mathbf{X}_t\right\|_{\text{op}} \le 5\eta/3$ holds for any $t \in [T]$, which immediately implies $\left\|\mathbf{X}_{t+1} - \mathbf{X}_t\right\|_{\text{op}} \le 1/\left\|\mathbf{L}_1\right\|_{\text{op}}$ by the choice of η in Equation (5). Thus, we invoke Lemma E.5 to deduce that

$$\begin{aligned} f(\mathbf{X}_{t+1}) \le & f(\mathbf{X}_t) - \eta \left\langle \boldsymbol{\nabla}_t, \operatorname{msign}\left(\widetilde{\mathbf{B}}_t\right) + \lambda \mathbf{X}_t \right\rangle \\ & + \frac{1}{2} \operatorname{tr}\left(\left(\mathbf{X}_{t+1} - \mathbf{X}_t\right)^\top \mathbf{L}(\mathbf{X}_t) \left(\mathbf{X}_{t+1} - \mathbf{X}_t\right) \right) \\ \overset{\text{Lemma E.2}}{\le} & f(\mathbf{X}_t) - \eta \left\langle \boldsymbol{\nabla}_t, \operatorname{msign}\left(\boldsymbol{\nabla}_t\right) \right\rangle + \eta \left\langle \boldsymbol{\nabla}_t, \operatorname{msign}\left(\boldsymbol{\nabla}_t\right) - \operatorname{msign}\left(\widetilde{\mathbf{M}}_t\right) \right\rangle \\ & + \eta\lambda \left\|\mathbf{X}_t\right\|_{\text{op}} \left\|\boldsymbol{\nabla}_t\right\|_* + \frac{1}{2} \left\|\mathbf{X}_{t+1} - \mathbf{X}_t\right\|_{\text{op}}^2 \left\|\mathbf{L}(\mathbf{X}_t)\right\|_* \\ \overset{\text{Lemmas E.1 to E.3 and E.8}}{\le} & f(\mathbf{X}_t) - \eta \left\|\boldsymbol{\nabla}_t\right\|_* + 2\eta \left\|\boldsymbol{\nabla}_t - \widetilde{\mathbf{M}}_t\right\|_* + \frac{2\eta}{3} \left\|\boldsymbol{\nabla}_t\right\|_* \\ & + \frac{25\eta^2}{9} \operatorname{tr}\left(\left|\mathbf{L}_0\right|_{\text{m}} + \left|\boldsymbol{\nabla}_t \mathbf{L}_1^\top\right|_{\text{m}} \right) \\ \overset{\text{Lemma E.1}}{\le} & f(\mathbf{X}_t) - \frac{\eta}{3} \left\|\boldsymbol{\nabla}_t\right\|_* + 2\eta \left\|\boldsymbol{\nabla}_t - \widetilde{\mathbf{M}}_t\right\|_* + \frac{25\eta^2 \left\|\mathbf{L}_0\right\|_*}{9} + \frac{25\eta^2 \left\|\mathbf{L}_1\right\|_{\text{op}} \left\|\boldsymbol{\nabla}_t\right\|_*}{9} \\ \overset{\text{Equation (5)}}{\le} & f(\mathbf{X}_t) - \frac{8\eta}{25} \left\|\boldsymbol{\nabla}_t\right\|_* + 2\eta \left\|\boldsymbol{\nabla}_t - \widetilde{\mathbf{M}}_t\right\|_* + \frac{25\eta^2 \left\|\mathbf{L}_0\right\|_*}{9}, \end{aligned}$$

where the second inequality uses Cauchy-Schwarz inequality and the fact that $\left\|\cdot\right\|_{\text{op}}$ is submultiplicative; the last step is due to $\eta \le 3/(625 \left\|\mathbf{L}_1\right\|_{\text{op}})$. Rearranging the above relation and summing from 1 to T yields

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\left\|\boldsymbol{\nabla}_t\right\|_*\right] \le \frac{25\Delta_f}{8\eta T} + \frac{625\eta}{72} \left\|\mathbf{L}_0\right\|_* + \frac{25}{4T} \sum_{t=1}^{T} \mathbb{E}\left[\left\|\widetilde{\mathbf{M}}_t - \boldsymbol{\nabla}_t\right\|_*\right], \tag{35}$$

Recall the notations in Equation (34) and decompose $\mathbb{E}\left[\left\|\widetilde{\mathbf{M}}_t - \boldsymbol{\nabla}_t\right\|_*\right]$ as follows:

$$\begin{aligned} & \mathbb{E}\left[\left\|\widetilde{\mathbf{M}}_t - \boldsymbol{\nabla}_t\right\|_*\right] = \mathbb{E}\left[\left\|\beta_1 \mathbf{M}_t + (1-\beta_2)\mathbf{G}_t - \boldsymbol{\nabla}_t\right\|_*\right] \\ = & \mathbb{E}\left[\left\|\beta_1(\mathbf{M}_t - \boldsymbol{\nabla}_t) + (1-\beta_2)(\mathbf{G}_t - \boldsymbol{\nabla}_t) + (\beta_1 - \beta_2)(\boldsymbol{\nabla}_t)\right\|_*\right] \\ \le & \beta_1 \mathbb{E}\left[\left\|\mathbf{E}_t\right\|_*\right] + (1-\beta_2)\mathbb{E}\left[\left\|\mathbf{N}_t\right\|_*\right] + \left|\beta_1 - \beta_2\right| \mathbb{E}\left[\left\|\boldsymbol{\nabla}_t\right\|_*\right], \end{aligned} \tag{36}$$

which holds for all $t \in [T]$. The noise term $\mathbf{N}_t$ can be bounded by

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{N}_t\|_*\right] \overset{\text{Lemma E.4}}{\leq} & \mathbb{E}\left[\sqrt{\||\mathbf{V}_0|_{\mathrm{m}}\|_* \operatorname{tr}\left(\mathbf{N}_t^\top |\mathbf{V}_0|_{\mathrm{m}}^{-1} \mathbf{N}_t\right)}\right] \\
= & \sqrt{\|\mathbf{V}_0\|_*}\mathbb{E}\left[\mathbb{E}\left[\|\mathbf{N}_t\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}} \middle| \mathcal{F}_{t-1}\right]\right] \\
\overset{\text{Lemma D.2}}{\leq} & \sqrt{\|\mathbf{V}_0\|_*}\mathbb{E}\left[\left(\mathbb{E}\left[\|\mathbf{N}_t\|_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}^p \middle| \mathcal{F}_{t-1}\right]\right)^{1/p}\right] \\
\overset{\text{Lemma E.6}}{\leq} & \sqrt{\|\mathbf{V}_0\|_*}\mathbb{E}\left[\left(2B^{1-p}\left(\|\mathbf{V}_0\|_*^{p/2} + \frac{\|\mathbf{V}_1\|_{\mathrm{op}}^p \|\boldsymbol{\nabla}_t\|_*^p}{\|\mathbf{V}_0\|_*^{p/2}}\right)\right)^{1/p}\right] \\
\overset{\text{Lemma D.1}}{\leq} & 2B^{\frac{1-p}{p}}\sqrt{\|\mathbf{V}_0\|_*}\mathbb{E}\left[\|\mathbf{V}_0\|_*^{1/2} + \frac{\|\mathbf{V}_1\|_{\mathrm{op}} \|\boldsymbol{\nabla}_t\|_*}{\|\mathbf{V}_0\|_*^{1/2}}\right] \\
= & 2B^{\frac{1-p}{p}} \|\mathbf{V}_0\|_* + 2B^{\frac{1-p}{p}} \|\mathbf{V}_1\|_{\mathrm{op}} \mathbb{E}\left[\|\boldsymbol{\nabla}_t\|_*\right].
\end{aligned} \tag{37}$$

As for the error term $\mathbf{E}_t$, we mainly adopt the same procedures as in Appendix E.3. Following Equation (14), $\mathbf{E}_t$ could be expressed as $\mathbf{E}_t = \beta_2 \mathbf{E}_{t-1} + \beta_2 \mathbf{S}_t + (1-\beta_2)\mathbf{N}_t$, implying

$$\mathbf{E}_t = -\beta_2^t \boldsymbol{\nabla}_1 + (1-\beta_2)\sum_{k=1}^t \beta_2^{t-k}\mathbf{N}_k + \sum_{k=2}^t \beta_2^{t-k+1}\mathbf{S}_k,$$

where we utilize $\mathbf{E}_1 = \mathbf{M}_1 - \boldsymbol{\nabla}_1 = (1-\beta)\mathbf{G}_1 - \boldsymbol{\nabla}_1 = (1-\beta)\mathbf{N}_1 - \beta\boldsymbol{\nabla}_1$. Then, it holds that

$$\mathbb{E}\left[\|\mathbf{E}_t\|_*\right] \leq \underbrace{\left\|\beta_2^t \boldsymbol{\nabla}_1\right\|_*}_{\mathtt{A}_t} + \underbrace{\mathbb{E}\left[\left\|(1-\beta_2)\sum_{k=1}^t \beta_2^{t-k}\mathbf{N}_k\right\|_*\right]}_{\mathtt{B}_t} + \underbrace{\mathbb{E}\left[\left\|\sum_{k=2}^t \beta_2^{t-k+1}\mathbf{S}_k\right\|_*\right]}_{\mathtt{C}_t}.$$

For $\mathtt{B}_t$, we can safely replace the β in Equation (33) by β_2 to derive

$$\mathtt{B}_t \leq \frac{4\sqrt{2}(1-\beta_2)^{\frac{p-1}{p}} \|\mathbf{V}_0\|_*}{B^{\frac{p-1}{p}}} + \frac{4\sqrt{2}(1-\beta_2)\|\mathbf{V}_1\|_{\mathrm{op}}}{B^{\frac{p-1}{p}}} \sum_{k=1}^t \mathbb{E}\left[\beta_2^{t-k}\|\boldsymbol{\nabla}_k\|_*\right],$$

For $\mathtt{C}_t$, we need to cope with weight decay carefully. By Lemma E.8, we have $\|\mathbf{X}_t - \mathbf{X}_{t-1}\|_{\mathrm{op}} \leq 5\eta/3 \leq 1/\|\mathbf{L}_1\|_{\mathrm{op}}$. So we can utilize Assumption 2b in the following way:

$$\begin{aligned}
\mathtt{C}_t \leq & \sum_{k=2}^t \beta_2^{t-k+1}\mathbb{E}\left[\|\mathbf{S}_k\|_*\right] \overset{\text{Lemma E.4}}{\leq} \sum_{k=2}^t \beta_2^{t-k+1}\mathbb{E}\left[\sqrt{\|\mathbf{L}(\mathbf{X}_k)\|_*}\|\mathbf{S}_k\|_{(\mathbf{L}(\mathbf{X}_k))^{-1}}\right] \\
\leq & \sum_{k=2}^t \beta_2^{t-k+1}\mathbb{E}\left[\sqrt{\|\mathbf{L}(\mathbf{X}_k)\|_*}\|\mathbf{X}_k - \mathbf{X}_{k-1}\|_{\mathbf{L}(\mathbf{X}_k)}\right] \\
\leq & \sum_{k=2}^t \beta_2^{t-k+1}\mathbb{E}\left[\sqrt{\|\mathbf{L}(\mathbf{X}_k)\|_* \cdot \|\mathbf{X}_k - \mathbf{X}_{k-1}\|_{\mathrm{op}}^2 \cdot \|\mathbf{L}(\mathbf{X}_k)\|_*}\right] \\
\overset{\text{Lemma E.8}}{\leq} & \sum_{k=2}^t \beta_2^{t-k+1}\mathbb{E}\left[\operatorname{tr}\left(\mathbf{L}(\mathbf{X}_k)\right) \cdot \frac{5\eta}{3}\right] \\
\overset{\text{Lemma E.1}}{\leq} & \frac{5\eta}{3}\sum_{k=2}^t \beta_2^{t-k+1}\operatorname{tr}\left(|\mathbf{L}_0|_{\mathrm{m}}\right) + \eta\sum_{k=2}^t \beta_2^{t-k+1}\mathbb{E}\left[\|\mathbf{L}_1\|_{\mathrm{op}}\|\boldsymbol{\nabla}_k\|_*\right] \\
\leq & \frac{5\eta\|\mathbf{L}_0\|_*}{3(1-\beta_2)} + \frac{5\eta}{3}\|\mathbf{L}_1\|_{\mathrm{op}}\sum_{k=2}^t \beta_2^{t-k+1}\mathbb{E}\left[\|\boldsymbol{\nabla}_k\|_*\right],
\end{aligned}$$

where the fourth inequality leverages Cauchy-Schwarz inequality and $\|AB\|_{\text{op}} \leq \|A\|_{\text{op}} \|B\|_{\text{op}}, \forall A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times r}$. Combining the bounds for $\mathtt{A_t}, \mathtt{B_t}, \mathtt{C_t}$, summing from 1 to T, and applying Equation (21):

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathbf{E}_t\|_*\right] \leq \frac{1}{T}\sum_{t=1}^{T}\beta_2^t\|\boldsymbol{\nabla}_1\|_* + \frac{5\eta\|\mathbf{L}_0\|_*}{3(1-\beta_2)} + \frac{5\eta\|\mathbf{L}_1\|_{\text{op}}}{3T}\sum_{t=1}^{T}\sum_{k=2}^{t}\beta_2^{t-k+1}\mathbb{E}\left[\|\boldsymbol{\nabla}_k\|_*\right] \\
&+ \frac{4\sqrt{2}(1-\beta_2)^{\frac{p-1}{p}}\|\mathbf{V}_0\|_*}{B^{\frac{p-1}{p}}} + \frac{4\sqrt{2}(1-\beta_2)\|\mathbf{V}_1\|_{\text{op}}}{TB^{\frac{p-1}{p}}}\sum_{t=1}^{T}\sum_{k=1}^{t}\mathbb{E}\left[\beta_2^{(t-k)}\|\boldsymbol{\nabla}_k\|_*\right] \\
\leq &\frac{\|\boldsymbol{\nabla}_1\|_*}{T(1-\beta_2)} + \frac{4\sqrt{2}(1-\beta_2)^{\frac{p-1}{p}}\|\mathbf{V}_0\|_*}{B^{\frac{p-1}{p}}} + \frac{5\eta\|\mathbf{L}_0\|_*}{3(1-\beta_2)} \\
&+ \frac{1}{T}\sum_{t=1}^{T}\left(\frac{4\sqrt{2}\|\mathbf{V}_1\|_{\text{op}}}{B^{\frac{p-1}{p}}} + \frac{5\eta\beta_2\|\mathbf{L}_1\|_{\text{op}}}{3(1-\beta_2)}\right)\mathbb{E}\left[\|\boldsymbol{\nabla}_t\|_*\right].
\end{aligned}$$

Plugging the above bound for $\mathbb{E}\left[\|\mathbf{E}_t\|_*\right]$, the bound for $\mathbb{E}\left[\|\mathbf{N}_t\|_*\right]$ in Equation (37) into Equation (36):

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\widetilde{\mathbf{M}}_t - \boldsymbol{\nabla}_t\right\|_*\right] \leq \frac{\beta_1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathbf{E}_t\|_*\right] + \frac{1-\beta_2}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathbf{N}_t\|_*\right] + \frac{|\beta_1-\beta_2|}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{\nabla}_t\|_*\right] \\
\leq &\frac{\beta_1\|\boldsymbol{\nabla}_1\|_*}{T(1-\beta_2)} + \frac{4\sqrt{2}\beta_1(1-\beta_2)^{\frac{p-1}{p}}\|\mathbf{V}_0\|_*}{B^{\frac{p-1}{p}}} + \frac{5\beta_1\eta\|\mathbf{L}_0\|_*}{3(1-\beta_2)} + \frac{(1-\beta_2)\|\mathbf{V}_0\|_*}{B^{\frac{p-1}{p}}} \\
&+ \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\left(4\sqrt{2}\beta_1+1-\beta_2\right)\|\mathbf{V}_1\|_{\text{op}}}{B^{\frac{p-1}{p}}} + \frac{5\eta\beta_1\beta_2\|\mathbf{L}_1\|_{\text{op}}}{3(1-\beta_2)}\right)\mathbb{E}\left[\|\boldsymbol{\nabla}_t\|_*\right] + \frac{3}{20T}\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{\nabla}_t\|_*\right] \\
\leq &\frac{\beta_1\|\boldsymbol{\nabla}_1\|_*}{T(1-\beta_2)} + \frac{(4\sqrt{2}+1)\beta_1\|\mathbf{V}_0\|_*}{(1-\beta_2)^{\frac{1-p}{p}}B^{\frac{p-1}{p}}} + \frac{5\beta_1\eta\|\mathbf{L}_0\|_*}{3(1-\beta_2)} \\
&+ \frac{159}{1000T}\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{\nabla}_t\|_*\right],
\end{aligned}$$

where the first inequality is due to $|\beta_1 - \beta_2| \leq 3/20$ as indicated by Equation (5); the last step is due to

$$\frac{\left(4\sqrt{2}\beta_1+1-\beta_2\right)\|\mathbf{V}_1\|_{\text{op}}}{B^{\frac{p-1}{p}}} + \frac{5\eta\beta_1\beta_2\|\mathbf{L}_1\|_{\text{op}}}{3(1-\beta_2)} \leq \frac{1}{1000} + \frac{5\times 3}{3\times 625} = \frac{9}{1000}.$$

Therefore, we plug in this relation into the initial bound in Equation (35) to obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{\nabla}_t\|_*\right] \leq \frac{25\Delta_f}{8\eta T} + \frac{625\eta}{72}\|\mathbf{L}_0\|_* \\
&+ \frac{25\beta_1\|\boldsymbol{\nabla}_1\|_*}{4T(1-\beta_2)} + \frac{25(4\sqrt{2}+1)\beta_1\|\mathbf{V}_0\|_*}{4(1-\beta_2)^{\frac{1-p}{p}}B^{\frac{p-1}{p}}} + \frac{125\beta_1\eta\|\mathbf{L}_0\|_*}{12(1-\beta_2)} + \frac{159}{160T}\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{\nabla}_t\|_*\right], \\
\Longrightarrow &\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{\nabla}_t\|_*\right] \leq \frac{500\Delta_f}{\eta T} + \frac{1000\|\boldsymbol{\nabla}_1\|_*}{T(1-\beta_2)} + \frac{6658\|\mathbf{V}_0\|_*}{(1-\beta_2)^{\frac{1-p}{p}}B^{\frac{p-1}{p}}} + \frac{1875\eta\|\mathbf{L}_0\|_*}{1-\beta_2}.
\end{aligned}$$

Setting $B, \beta_1, \beta_2, \eta$ according to Equation (5) completes the proof.

F Experimental Details

We provide the experimental setup and measurement protocols for Section 4.

F.1 Validation of Heavy-Tailed Noise

We train nanoGPT [185] on C4 [186] and measure centered stochastic-gradient noise over independently sampled mini-batches. The corresponding histograms and Gaussian Q-Q plots are shown in Figure 1. Figures 2 and 3 show the vector- and matrix-level noise–gradient relations near initialization. Measurements at step 4000 are shown in Figures 7 and 8.

F.2 Practical Efficiency of Sign-Based Optimizers

We compare NSGD [155, 159], AdamW [10, 11], SignSGD [38], Lion [12], Muon [13], and Muonlight [18]. The full trajectories in Figure 5 show a substantial gap between NSGD and sign-based methods, supporting the geometric comparisons in Appendix B. The matrix-sign methods also outperform AdamW, consistent with Liu et al. [18], Shah et al. [15], Wen et al. [16], Semenov et al. [17]. The algorithms and theorems distinguish Muon from Muonlight. In empirical summaries, Muon denotes the optimizer family.

F.3 Experimental Setup and Methodology

We conduct all experiments for the NanoGPT model using PyTorch and Distributed Data Parallel (DDP) across four NVIDIA Pro6000 GPUs (96GB VRAM each). Our implementation extends the codebase provided by Semenov et al. [17], incorporating additional modules for noise visualization. All models are trained for 10k steps with a global batch size of 512 sequences. We use a sequence length of 512, totaling approximately $1\times$ the Chinchilla-optimal token count as proposed by Hoffmann et al. [187]. For the AdamW, Lion, and SignSGD optimizers, we adopt the standard hyperparameters for 124M-parameter models suggested by Semenov et al. [17], specifically a learning rate $\eta = 10^{-3}$ and weight decay $\lambda = 0.1$. Meanwhile, we change the learning rates for Muon and Muonlight to $2 \times lr$ of AdamW, as suggested in Jordan et al. [13]. Finally, we employ a linear warm-up period of 10% iterations at the start of pretraining.

F.4 Estimating the Heavy-Tail Index p

Following the methodology of Zhang et al. [25], we estimate the heavy-tail index p using the estimator introduced by Simsekli et al. [24]. The corresponding histograms and Gaussian Q-Q plots are shown in Figure 1. Coordinates are randomly sampled from the model. Deviations in the tails from the Gaussian reference support the heavy-tailed noise model.

F.5 Validating the Noise Model

At each saved checkpoint, we freeze the model parameters and draw fresh independently sampled mini-batches without updating the iterate. Their stochastic gradients estimate the conditional mean gradient and centered noise moments at the fixed checkpoint. Given B such gradients, we write $\widehat{\mathbf{G}} = B^{-1}\sum_{b=1}^{B}\mathbf{G}^b$ and $\mathbf{N}^b = \mathbf{G}^b - \widehat{\mathbf{G}}$. For a selected coordinate j, we compare $|\widehat{\mathbf{g}}_j|^p$ with $B^{-1}\sum_b |\mathbf{n}_{b,j}|^p$ across sampled states. The matrix analogue compares $\|\widehat{\mathbf{G}}\|_*^p$ with the empirical weighted noise moment in Figure 3.

Diagonal noise estimator. For the matrix noise-relation plots, reshape the centered samples into $\mathbf{N}^b \in \mathbb{R}^{m\times n}$ and estimate

$$v_i = \left(\frac{1}{Bn}\sum_{b=1}^{B}\sum_{j=1}^{n}|N_{ij}^b|^p\right)^{1/p} + \epsilon, \qquad \widehat{\mathbf{V}}_0 = \operatorname{diag}(v_1,\dots,v_m), \quad \epsilon > 0.$$

This diagonal matrix defines the left weighting, with $\left|\widehat{\mathbf{V}}_0\right|_{\mathrm{m}} = \widehat{\mathbf{V}}_0$ and $\|\widehat{\mathbf{V}}_0\|_* = \sum_i v_i$. The regularizer makes the inverse well-defined. The plotted ordinate is

$$\frac{1}{B}\sum_{b=1}^{B}\left(\sum_i v_i\right)^{p/2}\left(\sum_{i,j}\frac{|N_{ij}^b|^2}{v_i}\right)^{p/2}.$$

These fixed-checkpoint measurements test the gradient-dependent moment relation, with the norm-based matrix formulation detailed in Assumption 4c.

F.6 Additional Experimental Protocols

Optimizer and off-trajectory perturbations. The 3×3 experiment uses GPT-2 124M, C4, and a common random seed across Adam, Lion, and Muon. Each optimizer is evaluated on its natural

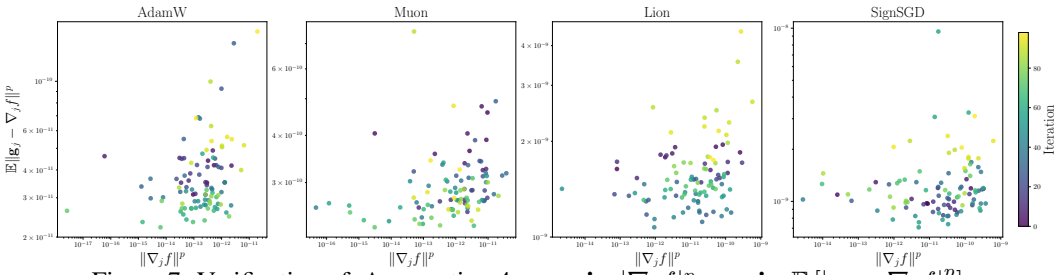


Figure 7: Verification of Assumption 4a. **x-axis**: $|\nabla_j f|^p$, **y-axis**: $\mathbb{E}\left[|\mathbf{g}_j - \nabla_j f|^p\right]$.

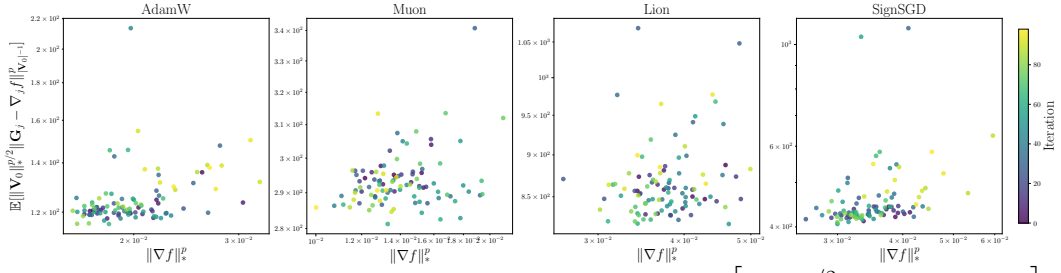


Figure 8: Verification of Assumption 4b. **x-axis**: $\|\nabla f\|_*^p$, **y-axis**: $\mathbb{E}\left[\|\mathbf{V}_0\|_*^{p/2}\,\|\mathbf{G} - \nabla f\|^p_{|\mathbf{V}_0|_{\mathrm{m}}^{-1}}\right]$.

trajectory and with Gaussian or Student-t perturbations that move iterates away from that trajectory. Table 1 reports the matrix-level linear-fit scores for the nine settings, using the noise–gradient quantities described above.

Vector geometry over the full model. Along a Lion run, we evaluate all 75 trainable tensors (123,980,544 parameters) at steps $0, 2500, 5000, 7500, 10000$. Noise scales are estimated from 256 stochastic gradients at each frozen checkpoint. Coordinate-wise Hessian row norms are estimated using HVPs on the loss averaged over eight mini-batches, with 64 Rademacher probes (128 at initialization). Using the density functions in Appendix B, the checkpoint analogues are

$$R_1 = \frac{\phi_\infty(\widehat{\mathbf{l}}_0)}{\phi_2^2(\widehat{\mathbf{g}})}, \qquad R_2 = \frac{\phi_2^2(\widehat{\boldsymbol{\sigma}}_0)}{\phi_2^2(\widehat{\mathbf{g}})}, \qquad R = R_1 R_2^{p/(2(p-1))}.$$

Here $\widehat{\mathbf{g}}$ is the mean gradient estimate. The hatted quantities are empirical curvature and noise estimates, not global theorem parameters. Table 4 gives the values plotted in Figure 4. In particular, the gradient densities at individual checkpoints differ from the trajectory minimum used in Appendix B.

Matrix Hessian and noise geometry. For `transformer.h.6.attn.c_attn.weight` in a Muon run, write $m = 2304$, $n = 768$, and $D = mn = 1{,}769{,}472$. The block Hessian $H \in \mathbb{R}^{D\times D}$ is taken with respect to the vectorized weight, with the remaining parameters fixed. We use HVPs and stochastic Lanczos quadrature without materializing H, averaging the loss over eight mini-batches and using three Rademacher probes with 24 Lanczos steps per probe. At steps 0, 5000, and 10000, $B = 256$ stochastic gradients give the sample mean $\widehat{\mathbf{G}} = B^{-1}\sum_{b=1}^{B}\mathbf{G}^b$ and centered noise samples $\mathbf{N}^b = \mathbf{G}^b - \widehat{\mathbf{G}}$. For this spectral diagnostic, we use a covariance factor whose singular values are determined by

$$\widehat{\mathbf{V}}_0\widehat{\mathbf{V}}_0^\top = \frac{1}{nB}\sum_{b=1}^{B}\mathbf{N}^b(\mathbf{N}^b)^\top.$$

Equivalently, one may take the positive semidefinite square root of this empirical row covariance. The factor $1/n$ is a scalar normalization and cancels in the effective dimension below. This diagnostic differs from the diagonal finite-p estimator used for the noise-relation plots.

The three columns of Table 2 are, respectively, the Hessian nuclear effective rank, the Hessian stable rank, and the noise effective dimension:

$$r_{\mathrm{nuc}}(H) = \frac{\|H\|_*}{\|H\|_{\mathrm{op}}}, \qquad r_{\mathrm{st}}(H) = \frac{\|H\|_{\mathrm{F}}^2}{\|H\|_{\mathrm{op}}^2}, \qquad \psi_2^2(\widehat{\mathbf{V}}_0) = \frac{\|\widehat{\mathbf{V}}_0\|_*^2}{\|\widehat{\mathbf{V}}_0\|_{\mathrm{F}}^2}. \tag{38}$$

Table 4: Geometry of the full model at five checkpoints along the Lion trajectory.

Step	R_1	R_2	R
0	0.106369	2.106431	0.245926
2.5k	0.00031698	4.505964	0.00172404
5k	0.00011942	2.162006	0.00028430
7.5k	0.00014205	2.074166	0.00032278
10k	0.00011295	1.451855	0.00017181

Table 5: Measured curvature and noise nuclear-norm scales at three checkpoints.

Step	Hessian-based nuclear norm	Noise nuclear norm
0	117.931	0.090438
5k	184.124	0.074534
10k	210.024	0.079701

Since H is symmetric, its singular values are the magnitudes of its eigenvalues, so $\|H\|_*$ sums the magnitudes of all Hessian eigenvalues, including negative ones. The nuclear effective rank sums eigenvalue magnitudes relative to the largest magnitude. The stable rank uses their squares. Both equal D for an isotropic full-rank spectrum and are small when curvature is concentrated in a few directions. For nonzero $\widehat{\mathbf{V}}_0$, $\psi_2^2(\widehat{\mathbf{V}}_0)$ lies between 1 and $\mathrm{rank}(\widehat{\mathbf{V}}_0)$, attaining the latter when the nonzero singular values are equal. The covariance factor can have rank up to $m = 2304$, whereas $r = \min\{m, n\} = 768$ is the rank limit of the weight matrix used in the theoretical norm comparison.

In Theorem 3, the bound depends on $\|\mathbf{L}_0\|_* = \psi_\infty(\mathbf{L}_0)\|\mathbf{L}_0\|_{\mathrm{op}}$ and $\|\mathbf{V}_0\|_* = \psi_2(\mathbf{V}_0)\|\mathbf{V}_0\|_{\mathrm{F}}$. Small spectral densities reduce these factors relative to their isotropic conversions. The noise effective dimension of 37.2 at 10k steps gives $\psi_2(\widehat{\mathbf{V}}_0)/\sqrt{768} \approx 0.22$. The ratio $\psi_2^2(\widehat{\mathbf{V}}_0)/\psi_2^2(\widehat{\mathbf{G}}) = 1.43$ also indicates moderate noise density relative to the gradient, as used in the comparison with matrix normalization in Appendix B. The Hessian ranks indicate curvature concentration. However, H acts on vectorized parameters and differs from $\mathbf{L}_0$ in Assumption 2b. These local measurements therefore do not determine a global smoothness matrix or the matrix complexity ratio R. Table 5 records the measured nuclear-norm scales.

F.7 Multi-Seed Stability

We repeat GPT-2 124M/C4 pretraining with multiple random seeds. Table 6 reports the mean and standard deviation of the final validation losses. Figure 5 shows individual runs.

The final losses vary little across seeds. Lion has variability comparable to AdamW, while Muon reaches a lower mean loss with a smaller standard deviation. This empirical stability is consistent with the update and trajectory bounds in Theorems 2 and 4.

Table 6: Final validation loss across random seeds (mean $\pm$ standard deviation).

Optimizer	AdamW	Lion	Muon
Final validation loss	3.467 ± 0.001	3.464 ± 0.001	3.429 ± 0.0005

F.8 Wall-Clock Efficiency

We measure end-to-end time for 10,000-step runs on the same four NVIDIA Pro6000 GPUs, holding the model, data, batch configuration, and evaluation schedule fixed. Table 7 summarizes both iteration and time efficiency, with average milliseconds per step rounded to the nearest integer. The relevant quantity is

$$\text{time-to-target} = (\text{steps-to-target}) \times (\text{wall-clock time per step}).$$

Muon incurs a modest per-step overhead from five Newton–Schulz iterations, but needs fewer steps to reach the same loss. Combining the iteration speedups with average step times gives time speedups of approximately $1.08\times$ and $1.27\times$ for Lion and Muon over AdamW. Total run time instead measures the cost of a fixed number of steps.

Table 7: Efficiency of 10k-step runs. Speedups are relative to AdamW at matched validation loss. Time speedups combine iteration speedups with measured average step times.

Optimizer	Final val. loss	Total time (s)	Time/step (ms)	Iter. speedup	Time speedup
AdamW	3.4672	4470	447	$1.00\times$	$1.00\times$
Lion	3.4642	4434	443	$1.07\times$	$1.08\times$
Muon	3.4293	4667	467	$1.32\times$	$1.27\times$

F.9 Local Geometry around the Final Checkpoints

We perturb a representative final checkpoint for each optimizer along 100 random directions, using a common normalized radius. Table 8 shows small loss increases for all three checkpoints. Lion has the smallest mean increase, while Muon has the smallest variation across directions and the lowest baseline loss. These measurements describe local sensitivity at individual checkpoints, while variability across seeds is reported separately in Table 6.

Table 8: Validation loss under 100 random perturbations around representative final checkpoints. The baseline values belong to these specific checkpoints.

Metric	AdamW	Lion	Muon
Baseline val. loss	3.467242	3.454907	3.429244
Mean perturbed loss	3.467340	3.454935	3.429352
Std. across probes	2.2×10^{-5}	1.8×10^{-5}	1.7×10^{-5}
Mean loss increase	9.8×10^{-5}	2.8×10^{-5}	1.08×10^{-4}

The Lion probing checkpoint has baseline loss 3.454907, compared with the mean 3.464 across seeds and the endpoint 3.4642 in the timing run. These measurements concern checkpoints within a finite training budget and do not establish asymptotic convergence.

G Limitations

Two-time-scale momentum. Our convergence bounds for Lion and Muonlight match those for SignSGD and Muon. They therefore do not establish a faster asymptotic rate from the additional momentum mechanisms. They show that decoupling the update and stored momentum, using Nesterov momentum, and adding weight decay preserve the rate under heavy-tailed noise. The weight decay analysis also gives a uniform trajectory bound. There exist two major viewpoints for the empirical gains. The first one is trajectory stability [63, 124], which we have demonstrated $\|\mathbf{x}_t\| \lesssim 1/\lambda$ in Theorem 2. The second one is implicit variance reduction [16, 35, 62, 76]. However, without stronger average smooth assumptions, it's impossible to prove a better $O(T^{-1/3})$ convergence.

Theory for AdamW. We do not analyze AdamW under the same generalized heavy-tailed noise model. The comparison with AdamW relies on our guarantees for sign-based methods, existing examples of slow convergence for adaptive methods [34], and the pretraining experiments. These results do not prove a general separation between Lion or Muon and AdamW. A direct comparison requires an AdamW theory that handles adaptive second-moment normalization, bias correction, coordinate rescaling, and weight decay under heavy-tailed noise. This remains difficult even under standard smoothness without the generalized growth terms, and it is explictly posed by Yu et al. [33] as an open problem.

Experimental scope. Our experiments use GPT-2 124M trained on C4. They support the noise model and geometric predictions in this setting, but do not cover larger models, other architectures, post-training, or broader datasets. In non-language tasks with lighter gradient tails, the advantage of sign-based methods may diminish. Testing these regimes would clarify how broadly the findings apply.

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 heavy-tailed 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 demonstrated in the paper rely on specific assumptions, which have been clearly stated in the main text. We also explicitly discuss other limitations in Appendix G.

   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 F.

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: Our code is available at `https://github.com/Dingzhen230/Heavy-tailed-Noise-in-LLMs`.

   Guidelines: Please follow the instructions in `README.md` to reproduce our results.

   - 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 F.

   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: Table 6 reports final validation-loss means and standard deviations across random seeds for AdamW, Lion, and Muon. Table 8 separately reports standard deviations across 100 random perturbations; the full training curves are individual runs, not seed averages.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We have provided the relevant information.

   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.