Free Heavy-Tailed Lunch for Muon: A Theoretical Justification of Empirical Success

Florian Hübler[1,2,*] **Thomas Pethick**[3] **Suvrit Sra**[2,4]

[1]Department of Computer Science, ETH Zurich, Switzerland
[2]Department of Mathematics, Technical University of Munich, Germany
[3]Harvard University [4]Munich Center for Machine Learning (MCML)

Abstract

Non-Euclidean optimisation methods with matrix-valued updates, such as MUON and SCION, have recently shown strong empirical performance for training Transformer models, yet their theoretical advantages over Euclidean methods remain poorly understood. We address this gap in the heavy-tailed non-convex regime, where stochastic gradients have bounded p-th central moments, $p \in (1, 2]$. We show that certain non-Euclidean methods achieve optimal sample complexity under stronger stationarity measures, while Euclidean methods incur additional dimension-dependent costs. As a consequence, for $m \times n$ matrices, MUON finds an ε-stationary point in nuclear norm within $\mathcal{O}\left(\frac{\Delta_1 L}{\varepsilon^2} + \min\{m, n\}\frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)$ samples, absorbing heavy-tailed noise without extra dimension dependence. We further prove this sample complexity, including its dimension dependence, is optimal for all first-order methods under nuclear-norm stationarity. Experiments on large language models support our theory. Surprisingly, our results suggest that other Schatten geometries beyond the spectral geometry of MUON can perform competitively in certain settings.

1 Introduction

We consider the unconstrained stochastic optimisation problem

$$\min_{x \in \mathcal{B}} \{F(x) \coloneqq \mathbb{E}\left[f(x, \xi)\right]\}, \tag{1}$$

where ξ is distributed according to some unknown distribution and $\mathcal{B}$ is a Banach space such as $\mathbb{R}^d$ or $\mathbb{R}^{m \times n}$. This setting has been the focus of a large body of work in optimisation due to its prevalence in machine learning and data-driven optimisation (Bottou et al., 2018).

While ADAM (Kingma and Ba, 2014) has been the de facto standard optimisation algorithm for training Transformer-based models for the last decade, a recent line of literature has proposed a new class of methods based on steepest descent in general normed spaces that are empirically competitive with ADAM and its variants. The general form of these algorithms is given by the normalised steepest descent (Boyd and Vandenberghe, 2004) update

$$x_{t+1} \leftarrow x_t + \eta \operatorname*{argmin}_{\|d\| \leq 1} \langle g_t, d \rangle, \tag{2}$$

where $\eta > 0$ is the stepsize, g_t some gradient estimator and $\|\cdot\|$ a norm on $\mathcal{B}$ (Pethick et al., 2025b). The most prominent example is MUON (Tuddenham et al., 2022; Jordan et al., 2024b), which uses (1) the spectral norm $\|\cdot\| = \|\cdot\|_{\mathrm{S}_\infty}$, (2) the momentum gradient estimator, and (3) Newton-Schulz

*Corresponding author: florian.huebler@inf.ethz.ch

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

iterations (Kovarik, 1970) to estimate the argmin. This variant achieves impressive results on the modded nanoGPT speedrun (Jordan et al., 2024a) and was used to train recent trillion parameter language models such as Kimi K2 (Kimi Team et al., 2026) and DeepSeek V4 (DeepSeek-AI et al., 2026).

From a theoretical perspective, recent works have established convergence of such algorithms in the non-convex setting. These guarantees are typically of the form

$$\mathbb{E}\left[\|\nabla F(x_T)\|_*\right] \leq \varepsilon \quad \text{after} \quad T = \mathcal{O}\left(\rho^2 \cdot \frac{\Delta_1 L \sigma^2}{\varepsilon^4}\right) \quad \text{samples,} \tag{3}$$

where F is L-smooth and the stochastic gradients have bounded variance σ^2, $\|\cdot\|_*$ is the dual norm and ρ is the norm-equivalence constant between the dual and Euclidean norm, with $\rho^2 = d$ for ℓ_∞ geometry and $\rho^2 = \min\{m, n\}$ for spectral geometry (e.g., Pethick et al., 2025b; Shen et al., 2025). Importantly, (3) guarantees $\mathbb{E}\left[\|\nabla F(x_T)\|_*\right] \leq \varepsilon$, i.e., x_T is an ε-$\|\cdot\|_*$-stationary point. For MUON, the dual is the nuclear norm $\|A\|_{\mathrm{S}_1} = \sum_i \sigma_i(A)$, where $\sigma_i(A)$ denote its singular values. Since $\|A\|_F \leq \|A\|_{\mathrm{S}_1}$, nuclear-norm ε-stationarity implies Frobenius-norm ε-stationarity, while the converse generally incurs a dimension-dependent factor. While informative, these guarantees face two difficulties in capturing the potential benefits of non-Euclidean methods in stochastic settings.

First, the appearance of the norm-equivalence constant ρ can make the bound effectively dimension-dependent, thereby obscuring potential advantages over Euclidean methods. This term arises from the treatment of the gradient noise. Indeed, the analysis must control terms of the form $\mathbb{E}[\|\sum(\nabla F(x_t) - \nabla f(x_t, \xi_t))\|_*]$. Such bounds are straightforward in Hilbert spaces, where the inner-product structure is available, but much harder for general norms. Existing analyses therefore use the norm equivalence $\|\cdot\|_* \leq \rho\,\|\cdot\|_2$ to a Euclidean norm, before controlling the noise in the Euclidean geometry. In particular, for MUON's nuclear norm on $\mathbb{R}^{m\times n}$, we get $\rho = \sqrt{\min\{m, n\}}$.

Second, these results rely on the bounded variance assumption, which can be overly restrictive. Indeed, empirical evidence from image classification (Simsekli et al., 2019; Battash et al., 2024), language modelling (J. Zhang et al., 2020a; Ahn et al., 2024), and reinforcement learning (RL) (Garg et al., 2021) suggests that stochastic gradients can exhibit heavy-tailed behaviour. These observations suggest that heavy-tailed noise provides a more faithful model of stochastic gradients in modern machine learning. Consequently, if one seeks to understand the empirical advantage of non-Euclidean optimisation methods, it is natural to analyse them in this regime. These two limitations naturally raise the following question:

Can non-Euclidean methods such as MUON *admit provable improvements over Euclidean methods in stochastic regimes, particularly under heavy-tailed noise?*

To study this question, we depart from the usual Euclidean noise assumption and instead make the heavy-tailed assumption directly in the native geometry: for some tail index $p \in (1, 2]$ we assume

$$\mathbb{E}\left[\|\nabla f(x, \xi) - \nabla F(x)\|_*^p\right] \leq \sigma^p.$$

This choice is the root of the main technical challenges, but also what enables sharper, geometry-aware guarantees.

Contributions. In this work we show that non-Euclidean methods can achieve strict improvements over their Euclidean counterparts in the heavy-tailed setting.

(i) We show that the optimal heavy-tailed Euclidean rate remains attainable for stronger dual-norm stationarity measures, without a norm-equivalence penalty. For p-uniformly smooth dual norms (see Definition 2.1), normalised steepest descent with momentum reaches an ε-$\|\cdot\|_*$-stationary point using $\mathcal{O}\left(\frac{\Delta_1 L}{\varepsilon^2}\left(1 + \left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)\right)$ samples (Theorem 3.1). A matching lower bound for first-order methods establishes optimality (Theorem F.9), while the upper bound extends to weak (L_0, L_1)-smoothness (Theorem C.3).

(ii) Based on this result, we obtain improved convergence guarantees for MUON, whose dual norm $\|\cdot\|_{\mathrm{S}_1}$ is not uniformly smooth. We show that MUON achieves the leading stochastic sample complexity

$$\mathcal{O}\left(\min\{m, n\} \cdot \frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right) \qquad \text{vs.} \qquad \Omega\left(\min\{m, n\}^{\frac{3p-2}{2(p-1)}} \cdot \frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)$$

required by classical Euclidean NSSD-M for the same stationarity measure (Theorems 3.5 and 3.6). This shows that MUON can absorb heavy-tailed noise essentially without extra dimensional cost, whereas Euclidean NSSD-M can match the same stationarity measure only at the price of a strictly worse, p-dependent dimension factor.

(iii) We provide a first-order sample complexity lower-bound of

$$\Omega\left(\min\{m,n\}\cdot\frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)$$

for $\|\cdot\|_{\mathrm{S}_1}$-stationarity. In particular, MUON is first-order optimal and there cannot exist a dimension-free guarantee for $\|\cdot\|_{\mathrm{S}_1}$-stationarity (Theorems 3.8 and 3.9).

As a side result, we provide tools from functional analysis to control the gradient noise for general norms, sidestepping the usual reduction to Euclidean martingale bounds through norm equivalence. These tools may be of independent interest for the analysis of general-norm algorithms beyond the heavy-tailed setting.

2 Preliminaries

Notation. Throughout, $(\mathcal{B}, \|\cdot\|)$ denotes a Banach space with dual $(\mathcal{B}^*, \|\cdot\|_*)$. We denote the duality pairing by $\langle x^*, x\rangle \coloneqq x^*(x)$ for $x \in \mathcal{B}, x^* \in \mathcal{B}^*$. To keep consistency with the typical notation in $\mathbb{R}^d$ we abuse notation[2] by writing $\nabla F(x)$ for the Fréchet derivative $DF(x) \in \mathcal{B}^*$ and $\langle \nabla F(x), d\rangle$ for the duality pairing between the Fréchet derivative and $d \in \mathcal{B}$. For a matrix A with singular values $s_i(A)$, the Schatten-r norm is $\|A\|_{S_r} = (\sum_i s_i(A)^r)^{1/r}$ for $1 \le r < \infty$, and $\|A\|_{S_\infty} = \max_i s_i(A)$. Its dual is $\|\cdot\|_{S_{r'}}$, where $1/r + 1/r' = 1$. The cases $r = 1, 2, \infty$ are the nuclear, Frobenius and spectral norms, respectively.

Algorithm. We study the stochastic normalised steepest descent method with momentum,

$$\begin{aligned} m_t &\leftarrow \beta_t m_{t-1} + (1-\beta_t)\nabla f(x_t, \xi_t), \\ x_{t+1} &\leftarrow x_t + \eta_t \operatorname*{argmin}_{\|d\|\le 1} \langle m_t, d\rangle, \end{aligned} \tag{NSSD-M}$$

where $\xi_1, \xi_2, \ldots \overset{\text{i.i.d.}}{\sim} \xi$ are independent copies of ξ, $\nabla f(x_t, \xi_t)$ is a stochastic gradient oracle of $\nabla F(x_t)$, $\beta \in [0,1)$ is the momentum parameter and $m_0 \coloneqq 0$. When choosing $(\mathcal{B}, \|\cdot\|) = (\mathbb{R}^{m\times n}, \|\cdot\|_{\mathrm{S}_\infty})$ this corresponds to an idealised version of MUON with an exact lmo, and $(\mathcal{B}, \|\cdot\|) = (\mathbb{R}^d, \|\cdot\|_{\ell_\infty})$ to sign descent with momentum (SIGNUM). The classical Euclidean version of NSSD-M, $x_{t+1} \leftarrow x_t - \eta_t \frac{m_t}{\|m_t\|}$, corresponds to $(\mathbb{R}^d, \|\cdot\|_{\ell_2})$ in the vector, and $(\mathbb{R}^{m\times n}, \|\cdot\|_{\mathrm{F}})$ in the matrix case. We consider the following standard assumptions.

Assumption 1. There exists $\Delta_1 < \infty$ such that $F(x_1) - \inf_{x\in\mathcal{B}} F(x) \le \Delta_1$.

For ease of exposition, we present the results for the case of L-smoothness as defined below. Our results readily also hold for the weaker notion of (L_0, L_1)-smoothness and can be found in Sections C and D.

Assumption 2. There exists $L \ge 0$ such that for all $x, y \in \mathcal{B}$, $\|\nabla F(x) - \nabla F(y)\|_* \le L\|x - y\|$.

Finally, as mentioned in the introduction, we consider the heavy-tailed setting where the stochastic gradient oracle has bounded p-th central moment for some $p \in (1, 2]$.

Assumption 3. There exists $p \in (1,2], \sigma \ge 0$ such that the stochastic gradient oracle is unbiased and has bounded p-th central moment, i.e., for all $x \in \mathcal{B}$,

$$\mathbb{E}[\nabla f(x,\xi)] = \nabla F(x), \qquad \text{and} \qquad \mathbb{E}[\|\nabla f(x,\xi) - \nabla F(x)\|_*^p] \le \sigma^p.$$

[2]The gradient is generally not well defined in Banach spaces, as its definition from $DF(x)$ relies on the Riesz representation theorem only available in Hilbert spaces.

2.1 Moment Bounds for General Norms

We first summarise the necessary technical tools to control the gradient approximation error for general norms, sidestepping the usual reduction to Euclidean martingale bounds through norm equivalence. Therefore we first recall the following key geometric definitions of general norms.

Definition 2.1 (Ball et al., 1994, Equation (2.7))**.** Let $(X, \|\cdot\|)$ be a Banach space.

i) The *p-smoothness constant* of $(X, \|\cdot\|)$ is defined as

$$S_p(X) \coloneqq \inf\{S \geq 1 \mid \forall\, x, y \in X\colon\ \|x+y\|^p + \|x-y\|^p \leq 2\|x\|^p + 2S^p\|y\|^p\},$$

and we call $(X, \|\cdot\|)$ p-uniformly smooth if $S_p(X) < \infty$.

ii) The *q-convexity constant* of $(X, \|\cdot\|)$ is defined as

$$K_q(X) \coloneqq \inf\{K \geq 1 \mid \forall\, x, y \in X\colon 2\|x\|^q + 2/K^q\|y\|^q \leq \|x+y\|^q + \|x-y\|^q\},$$

and we call $(X, \|\cdot\|)$ q-uniformly convex if $K_q(X) < \infty$.

The following strong duality holds between the convexity and smoothness constants.

Lemma 2.2 (Ball et al., 1994, Lemma 5)**.** *Let $(X, \|\cdot\|)$ be a Banach space and $1 < p \leq 2 \leq q < \infty$ conjugate exponents, i.e., $1/p + 1/q = 1$. Then $S_p(X^*) = K_q(X)$.*

The above duality relationship allows us to verify the smoothness assumption on the dual space by checking a convexity property of the primal space, which is often more natural. Next let us get some intuition of the above definition for the special cases of $\mathbb{R}^d$ and $\mathbb{R}^{m\times n}$.

Example 2.3 (Hirzallah and Kittaneh, 2002)**.** Consider the ℓ_r space $(X, \|\cdot\|) = (\mathbb{R}^d, \|\cdot\|_r)$ or the Schatten-r space $(X, \|\cdot\|) = (\mathbb{R}^{m\times n}, \|\cdot\|_{S_r})$. In both cases the following holds.

i) If $1 < r \leq 2$, then X is r-uniformly smooth with $S_r(X) = 1$.

ii) If $2 \leq r < \infty$, then X is r-uniformly convex with $K_r(X) = 1$.

Finally we provide the main technical tool which will allow us to control the gradient approximation error for general norms, sidestepping the usual reduction to Euclidean martingale bounds through norm equivalence.

Theorem 2.4 (Pisier, 2016, Theorem 10.22)**.** *Let $p \in (1, 2]$. Every martingale difference sequence $Y_1, \ldots, Y_n$ in $\mathcal{B}^*$ satisfies*

$$\mathbb{E}\left[\left\|\sum_{t=1}^n Y_t\right\|_*^p\right] \leq 2S_p(\mathcal{B}^*)^p \cdot \sum_{t=1}^n \mathbb{E}\left[\|Y_t\|_*^p\right].$$

For completeness, we provide a proof of Theorem 2.4 in Section A. We note that the above inequality is vacuous if $\mathcal{B}^*$ is not p-uniformly smooth, i.e., if $S_p(\mathcal{B}^*) = \infty$.

3 Main Result

In this section we present our theoretical results. In Section 3.1 we provide convergence guarantees for NSSD-M for uniformly convex norms under heavy-tailed noise. In Section 3.2 we extend these results to the case of MUON. Finally, in Section 3.3 we provide a lower bound showing that the obtained dimension dependence is first-order optimal for nuclear-stationarity guarantees.

3.1 Guarantees for Uniformly Convex Norms

We are now ready to prove the convergence guarantee for uniformly convex norms under heavy-tailed noise. For ease of exposition we will present it under the standard L-smoothness assumption. Extended results for (L_0, L_1)-smoothness and the corresponding proofs can be found in Section C.2.

Theorem 3.1 (Uniformly-Convex Norm Guarantee)**.** *Suppose that Assumptions 1 to 3 hold. Furthermore let $q = p/(p-1)$ and assume that $(\mathcal{B}, \|\cdot\|)$ is q-uniformly convex with q-convexity constant K. Then the iterates generated by* NSSD-M *with parameters*

$$\eta_t \equiv \sqrt{\frac{\Delta_1(1-\beta)}{LT}}, \qquad \beta_t \equiv \beta = 1 - \min\left\{1, \left(\frac{\Delta_1 L}{\sigma^2 K^2 T}\right)^{\frac{p}{3p-2}}\right\}$$

satisfy

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(x_t)\|_*\right] \leq 18\sqrt{\frac{\Delta_1 L}{T}} + 28\left(\frac{\Delta_1 L}{T}(\sigma K)^{\frac{p}{p-1}}\right)^{\frac{p-1}{3p-2}}.$$

This shows that NSSD-M achieves the optimal Euclidean heavy-tailed sample complexity in these possibly infinite-dimensional Banach spaces. Importantly, the stationarity measure $\|\cdot\|_*$ can in general be stronger than in the Euclidean case. This becomes particularly transparent for Schatten and ℓ norms.

Corollary 3.2. *Let $(\mathcal{B}, \|\cdot\|) = (\mathbb{R}^{m\times n}, \|\cdot\|_{S_r})$ or $(\mathbb{R}^d, \|\cdot\|_{\ell_r})$, where $2 \leq r \leq p/(p-1)$. Then, in the setting of Theorem 3.1,* NSSD-M *satisfies*

$$\mathbb{E}\left[\|\nabla F(x_\tau)\|_*\right] \leq \varepsilon \quad \textit{after} \quad T = \mathcal{O}\left(\frac{\Delta_1 L}{\varepsilon^2}\left(1 + \left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)\right) \quad \textit{samples,} \tag{4}$$

where $\tau \sim \mathrm{Unif}([T])$. Here $\|\cdot\|_* = \|\cdot\|_{S_{r'}}$ and $\|\cdot\|_{\ell_{r'}}$, where $r' = r/(r-1)$, for the Schatten and ℓ case respectively.

In particular, when choosing $r = 2$, Corollary 3.2 recovers the leading rates of the existing heavy-tailed results for the Euclidean case (Hübler et al., 2025; Z. Liu and Zhou, 2025). For $2 < r \leq p/(p-1)$, the result is strictly stronger, as $\|\cdot\|_* \geq \|\cdot\|_2 \geq 1/\rho\,\|\cdot\|_*$, where $\rho = \min\{m,n\}^{(r-2)/2r}$ in Schatten-, and $\rho = d^{(r-2)/2r}$ in ℓ-geometry. Retaining only the leading stochastic term, a stationarity-measure comparable result from the Euclidean setting hence has the form

$$\mathbb{E}\left[\|\nabla F(x_\tau)\|_*\right] \leq \varepsilon \quad \text{after} \quad T = \mathcal{O}\left(\rho^{\frac{3p-2}{p-1}} \cdot \frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right) \quad \text{samples.} \tag{5}$$

This multiplicative factor can be considerable, as already for moderate values such as $p = 1.5$ and $r = p/(p-1)$ we have $\rho^{(3p-2)/(p-1)} \approx 10^4$ in the Schatten-, and $\approx 10^7$ in the ℓ-geometry for modern transformers (Chowdhery et al., 2023). The matching lower bound in Theorem F.9 further establishes first-order optimality for the Schatten case.

The main technical novelty in the proof of Theorem 3.1 is based on the careful application of Theorem 2.4 to derive the following deviation bound on our momentum gradient estimator. We remark that mini-batching with a suitable batch size would achieve the same sample complexity.

Lemma 3.3. *Suppose that Assumptions 1 to 3 hold, and let $q = p/(p-1)$. Assume that $(\mathcal{B}, \|\cdot\|)$ is q-uniformly convex with q-convexity constant K. Then the iterates generated by* NSSD-M *with $\eta_t \equiv \eta$ and $\beta_t \equiv \beta$ satisfy*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(x_t) - m_t\|_*\right] \leq \sqrt{\frac{2\Delta_1 L}{T(1-\beta)}} + 2\sigma K(1-\beta)^{\frac{p-1}{p}} + \frac{\eta L}{1-\beta}.$$

The above Lemma allows us to control the convergence of NSSD-M directly in its native geometry, without having to reduce to the Euclidean case through norm equivalence, thereby avoiding the dimension dependence. The extended version of this Lemma and its proof can be found in Lemma C.2.

Remark 3.4. It is well-known that weight-decay is often required in practice to efficiently train large language models (Kimi Team et al., 2026). For normalised steepest descent based methods such as NSSD-M, this regularisation has a natural constrained interpretation as a conditional gradient step for optimisation over the corresponding norm ball (Pethick et al., 2025b). We extend our convergence analysis to this constrained stochastic conditional gradient setting in Section C.3.

3.2 Guarantees for MUON

While the previous results established convergence of NSSD-M for a broad range of norms, they do not cover MUON as the Schatten-∞ norm is not uniformly convex. Instead, we will show in this section that MUON achieves a reduced dimension dependence without any modification to the algorithm. A more general version under (L_0, L_1)-smoothness and the corresponding proofs can be found in Section D.

The main technical idea is to use equivalence of norms only to get from the nuclear norm to the Schatten-p norm, where p is the tail index, instead of going all the way to the Frobenius norm. This allows us to heavily reduce the dimension dependence as the noise becomes more heavy-tailed, while still controlling a strictly stronger stationarity measure than the Frobenius norm. Let us now state the result.

Theorem 3.5 (Convergence of MUON under Heavy-Tailed Noise)**.** *Consider* $(\mathcal{B}, \|\cdot\|) = (\mathbb{R}^{m\times n}, \|\cdot\|_{\mathrm{S}_\infty})$ *and suppose Assumptions 1 to 3 hold. Furthermore let* $\kappa := \min\{m,n\}^{(p-1)/p}$. *Then the iterates generated by* MUON *with parameters*

$$\eta_t \equiv \sqrt{\frac{\Delta_1(1-\beta)}{LT}}, \qquad \beta_t \equiv 1 - \min\left\{1, \left(\frac{\Delta_1 L}{\sigma^2\kappa^2 T}\right)^{\frac{p}{3p-2}}\right\}$$

satisfy

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|_{\mathrm{S}_1}\right] \le 18\sqrt{\frac{\Delta_1 L}{T}} + 28\left(\min\{m,n\}\frac{\Delta_1 L\sigma^{\frac{p}{p-1}}}{T}\right)^{\frac{p-1}{3p-2}}.$$

MUON's sample complexity guarantee for nuclear-norm stationarity is therefore

$$\mathcal{O}\left(\frac{\Delta_1 L}{\varepsilon^2} + \min\{m,n\}\cdot\frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right).$$

This improves the dimension dependence of existing heavy-tailed Euclidean guarantees (Hübler et al., 2025; Z. Liu and Zhou, 2025) when expressed in terms of nuclear-norm stationarity (cf. (5)). Comparing upper bounds alone, however, does not establish a worst-case separation between the methods. We therefore prove the following matching lower bound.

Theorem 3.6 (Lower Bound for Euclidean NSSD-M)**.** *Let* $\min\{m,n\} \ge 2$, $p \in (1,2]$ *and* $0 < \varepsilon \le c\min\{\sigma, \sqrt{\Delta_1 L}\}$, *for a sufficiently small universal* $c > 0$. *For every* $T \ge 1$ *and constant parameters* $\beta \in [0,1)$ *and* $0 < \eta \le \sqrt{\Delta_1(1-\beta)/(LT)}$, *there exists a function and gradient oracle satisfying the assumptions of Theorem 3.5 such that Euclidean* NSSD-M *requires at least*

$$\Omega\left(\min\{m,n\}^{\frac{3p-2}{2(p-1)}}\cdot\frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right) \tag{6}$$

samples to find an iterate x_t, $t \le T$, *with* $\mathbb{E}\left[\|\nabla F(x_t)\|_{\mathrm{S}_1}\right] \le \varepsilon$. *Moreover, the gradient oracle can be chosen so that its nuclear and Frobenius* p*-th noise moments coincide.*

The proof is given in Theorem E.1. Together, Theorems 3.5 and 3.6 establish a strict worst-case separation in dimension dependence between MUON and Euclidean NSSD-M in nuclear norm stationarity. The stepsize restriction is mild, as it includes the optimal choice in our upper bounds. To illustrate, for $p = 1.5$ and matrix dimensions $\min\{m,n\} \approx 10^4$, as in modern transformer based models (Chowdhery et al., 2023), the dimension factors are 10^4 for MUON and 10^{10} for Euclidean NSSD-M, a difference of six orders of magnitude.

We note that Theorem 3.5 makes the noise assumption directly in the nuclear norm rather than in the Frobenius norm. Changing norms in the assumptions may introduce implicit dimension dependence, as discussed by Shen et al. (2025) in the context of smoothness. For the gradient noise $\gamma := \nabla f(x,\xi) - \nabla F(x)$, the inequality $\|\gamma\|_{\mathrm{S}_1} \le \sqrt{\min\{m,n\}}\,\|\gamma\|_{\mathrm{F}}$ gives the worst-case factor by which the noise constant can increase. Since the lower bound in Theorem 3.6 holds with equal noise moments, the separation above does not rely on this conversion. In Section 4, we empirically observe that this conversion is much less severe in practice (Figure 1). Furthermore, the statement of Theorem 3.5 verbatim holds when measuring the p-th noise moment in the Schatten-p norm, giving

a potentially smaller σ in both the parameter choice and the bound. For more details we refer to the discussion following Theorem D.2.

Remark 3.7. For simplicity of presentation, Theorem 3.5 is stated for a single matrix. However, the extended result Theorem D.2 directly applies to neural networks consisting of a collection of weight matrices $\{W_l\}_{l\in[D]}$, by equipping $\mathbb{R}^{m_1\times n_1}\times\ldots\times\mathbb{R}^{m_D\times n_D}$ with the norm $\max_{l\in[D]}\|W_l\|_{S_\infty}$ (Bernstein and Newhouse, 2024). In this case, the dimension dependence becomes $\sum_{l\in[D]}\min\{m_l,n_l\}$. A further extension to full networks including vector-valued weights is discussed in Remark D.5.

3.3 First-Order Lower Bounds for Nuclear-Norm Stationarity

While MUON improves on the dimension dependence of existing Euclidean guarantees, it is natural to ask whether the remaining factor $\min\{m,n\}$ can be removed altogether. We answer this question in the negative by providing a first-order lower bound that matches this linear dependence.

We follow the sample complexity framework of Arjevani et al. (2023), which we recall in Section F. Informally, $\mathfrak{m}_\varepsilon^{\text{rand}}\big(\|\cdot\|_{\mathrm{S}_1},\Delta_1,L,p,\sigma\big)$ gives a lower bound on the number of stochastic gradient samples needed by the best first-order algorithm to reach $\mathbb{E}\left[\|\nabla F(x)\|_{\mathrm{S}_1}\right]\le\varepsilon$ in the worst case over instances. The objectives and gradient oracles are assumed to satisfy Assumptions 1 to 3 on $(\mathbb{R}^{m\times n},\|\cdot\|_{\mathrm{S}_\infty})$. We first show that the linear dependence on $\min\{m,n\}$ is unavoidable for every matrix shape, even for convex objectives. Later we match the full stochastic rate under an additional dimension restriction.

Theorem 3.8. *Let $\Delta_1,L,\sigma>0$, $p\in(1,2]$ and assume $0<\varepsilon<\min\left\{\frac{\sigma}{4},\sqrt{\Delta_1 L}/2\right\}$. Then, for all $m,n\in\mathbb{N}_{\ge2}$, we have*

$$\mathfrak{m}_\varepsilon^{\mathit{rand}}\big(\|\cdot\|_{\mathrm{S}_1},\Delta_1,L,p,\sigma\big)\ge\frac{\min\{m,n\}}{2}\cdot\left(\frac{\sigma}{4\varepsilon}\right)^{\frac{p}{p-1}}.$$

Moreover, the lower bound remains valid when the function class is restricted to convex quadratic objectives.

We construct a convex quadratic whose minimiser encodes $\kappa:=\min\{m,n\}$ independent signs. Each oracle sample reveals one randomly chosen sign with probability $q=(4\varepsilon/\sigma)^{p/(p-1)}$. Reaching ε-stationarity requires learning a constant fraction of these signs, which takes $\Omega(\kappa/q)$ samples. The full construction and proof are given in Section F.1.

Theorem 3.8 establishes the optimal dimension dependence, but does not recover the factor $\Delta_1 L/\varepsilon^2$ in Theorem 3.5. The following result matches the full stochastic rate when the matrix aspect ratio is sufficiently large. This condition arises from our embedding of the high-dimensional Euclidean hard instances of Arjevani et al. (2023).

Theorem 3.9. *Let $c:=1/12000,\Delta_1,L,\sigma>0$, $p\in(1,2]$ and assume $0<\varepsilon\le c\min\left\{\sigma,\sqrt{\Delta_1 L}\right\}$. Then, for all $m,n\in\mathbb{N}_{\ge2}$ with $\kappa:=\min\{m,n\}$ and*

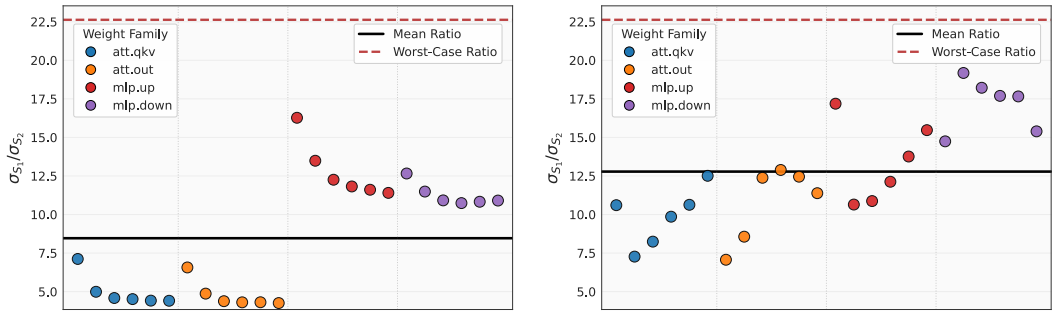
$$\max\{m,n\}\ge64\kappa\frac{\Delta_1 L}{\varepsilon^2}\log\left(2\kappa\left(\frac{\Delta_1 L}{\varepsilon^2}\right)^2\left(\frac{c\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right),$$

we have

$$\mathfrak{m}_\varepsilon^{\mathit{rand}}\big(\|\cdot\|_{\mathrm{S}_1},\Delta_1,L,p,\sigma\big)\ge c^2\kappa\cdot\frac{\Delta_1 L}{\varepsilon^2}\left(\frac{c\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}.$$

In particular, MUON *is first-order optimal for $\|\cdot\|_{\mathrm{S}_1}$-stationarity in this dimension regime.*

We embed κ independently rotated Euclidean hard instances (Arjevani et al., 2023) into a $\kappa d\times\kappa$ matrix, placing each instance in a separate column with disjoint row support. The resulting gradient has orthogonal columns, so its nuclear norm equals the sum of the Euclidean gradient norms. The random rotations hide unexplored directions from the algorithm. Each instance uses dimension d of order $\Delta_1 L/\varepsilon^2$, with an additional logarithmic factor due to these rotations. Since the embedding has aspect ratio d, this gives the dimension restriction. The full proof is given in Section F.2.



(a) Ratios at Initialisation

(b) Ratios at Final Checkpoint

Figure 1: Ratios $\sigma_{S_1,p}/\sigma_{F,p}$ for $p = 1.5$ across all weight matrices trained by MUON. Colours indicate weight families, ordered by layer depth within. While non-trivial, the noise ratios stay considerably below the worst-case ratio $\sqrt{\min\{m,n\}} \approx 22.5$.

4 Experiments

This section aims to empirically validate the theoretical findings of this paper. Since it was empirically observed that heavy-tailed gradient distributions occur in language modelling (J. Zhang et al., 2020b; Ahn et al., 2024), we focus on this task.[3]

Experimental Details. We pre-train the 70M-parameter architecture of the `Pythia` series of models (Biderman et al., 2023) from scratch on the FineWeb-Edu dataset[4] (Penedo et al., 2024) for Chinchilla-optimal 1.4B tokens (Hoffmann et al., 2022). Following previous work, we choose the sequence length to be 2048 and the batch size to be 512 (Biderman et al., 2023). Additionally we use the standard hyper-parameters $\beta = 0.95, \lambda = 0.1$ for `MUON` and use the WSD learning rate schedule (Zhai et al., 2022; Hägele et al., 2024; Hu et al., 2024) with a warmup of 5% and a cooldown period of 20%. We tune learning rates across the grid $[0.005, 0.0075, 0.01, 0.0125, 0.015, 0.03, 0.06]$ and repeat the experiments across the seeds $0, 1, 2$. Further details can be found in Section G.

Norm Change in Noise Assumption. Section 3 analyses NSSD-M directly in its native geometry rather than reducing to the Euclidean case. In particular, we consider the quantity

$$\sigma^p_{\|\cdot\|_*,p} := \mathbb{E}\left[\|\nabla f(x,\xi) - \nabla F(x)\|_*^p\right] \quad \text{rather than} \quad \sigma^p_{F,p} := \mathbb{E}\left[\|\nabla f(x,\xi) - \nabla F(x)\|_F^p\right]$$

which would be consistent with previous work. As noted by Shen et al. (2025), such a norm change in assumptions may implicitly introduce dimension dependencies. In the case of MUON, i.e., $\|\cdot\|_* = \|\cdot\|_{S_1}$, we have the worst-case guarantee $\sigma_{S_1,p} \leq \sqrt{\min\{m,n\}}\sigma_{F,p}$. We therefore empirically calculate the ratio $\sigma_{S_1,p}/\sigma_{F,p}$ for $p = 1.5$ for all weight matrices trained by MUON, as shown in Figure 1. We chose $p = 1.5$ a priori based on earlier empirical studies (Simsekli et al., 2019; Garg et al., 2021). For fixed $\sigma_{F,p}$, a lower ratio gives a sharper guarantee.

We observe that, while non-trivial, the noise ratios are far below the worst-case factor $\sqrt{\min\{m,n\}} \approx 22.5$. Furthermore the ratio is lower at initialisation than at the final checkpoint. We repeat the same experiment while adding the input and output embedding layer, but measure the noise in ℓ-geometry, reporting a $[0,1]$ normalised ratio in Figure 2. We observe that embeddings have a lower normalised ℓ ratio than the other weights. For additional details we refer to Section G and Shen et al. (2025), who conduct similar experiments for the smoothness constant.

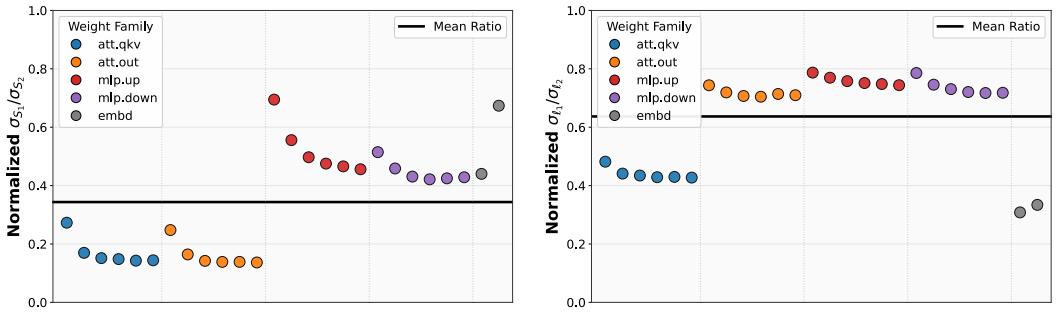
Table 1: Final perplexity for Schatten-r geometries (mean $\pm$ standard deviation over three seeds).

r	Perplexity ↓
1	88.89 ± 0.48
$8/7$	86.50 ± 0.91
$4/3$	67.61 ± 5.98
$3/2$	60.57 ± 4.48
2	40.18 ± 0.09
3	34.51 ± 0.13
4	33.54 ± 0.04
8	$\mathbf{33.02 \pm 0.01}$
∞	$\underline{33.34 \pm 0.02}$

Impact of Norm Choice. To test the theoretical insights from Theorem 3.1, we compare NSSD-M in different Schatten-r geometries, $r \in \{1, 8/7, 4/3, 3/2, 2, 3, 4, 8, \infty\}$. Here $r = 2$ corresponds to the standard Euclidean

[3] Our code is available on GitHub: `https://github.com/fhueb/ht-schatten-experiments`

[4] `https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu`



(a) Ratios in S_1.

(b) Ratios in ℓ_1.

Figure 2: Normalised noise ratios including the embedding and unembedding layer at initialisation. A normalised value of 0 corresponds to $\sigma_1 = \sigma_2$, a value of 1 corresponds to $\sigma_1 = \rho\sigma_2$, where ρ is the worst-case ratio. Interestingly, we find that the ratio is smaller for the embedding and unembedding layers for $\sigma_{\ell_1}/\sigma_{\ell_2}$.

geometry, while $r = \infty$ corresponds to the spectral geometry of MUON. Larger r corresponds to a stronger dual stationarity measure without an additional explicit dimension factor for $2 \leq r \leq p/p-1$. Beyond this point, the stronger stationarity comes with additional dimension dependence (Theorem D.2).

The results qualitatively agree with this trade-off. Performance improves sharply from $r = 1$ to $r = 3$, with smaller improvements up to $r = 8$, which gives the lowest perplexity (Table 1). We also observe that the spectral geometry of MUON, corresponding to $r = \infty$, substantially outperforms the Euclidean geometry. Finally, although intermediate Schatten geometries perform competitively, our experiments use a full SVD for $r \in \{3, 4, 8\}$.

5 Related Work

Steepest Descent in General Norms. While classical SGD can be seen as steepest descent in the Euclidean ℓ_2 norm, early works have proposed to consider general-norm (normalised) steepest descent algorithms (Boyd and Vandenberghe, 2004, Section 9.4). In particular, Kelner et al. (2014) proposed considering ℓ_∞-steepest descent for specific graph tasks, and Carlson et al. (2015a) suggest considering S_∞-steepest descent for Restricted Boltzmann Machines and Deep Learning (Carlson et al., 2015b). Independently, Tuddenham et al. (2022) proposed normalised S_∞-steepest descent with momentum for machine learning tasks under the name ORTHOGONAL-SGDM. This algorithm first orthogonalises each stochastic gradient and then applies momentum to the resulting orthogonalised directions. Neither approach was widely adopted until recently, when Bernstein and Newhouse (2024) proposed using Newton-Schulz iterations (Kovarik, 1970) to efficiently compute the orthogonalisation step. The blog (Jordan et al., 2024b) popularised the normalised S_∞-steepest descent approach under the name MUON, by changing the order of orthogonalisation and momentum in ORTHOGONAL-SGDM. Recently, Pethick et al. (2025b) and J. Liu et al. (2025) suggest MUON with decoupled weight decay as a natural extension from the theoretical and empirical perspective respectively.

Theory of MUON. Motivated by the strong empirical performance of MUON, a rapidly growing line of literature studies its convergence properties. J. Li and Hong (2025) and Shen et al. (2025) analyse MUON under L-smoothness and Euclidean bounded variance, and later relate its behaviour to low-rank and approximately blockwise-diagonal Hessian structure. Going beyond the S_∞ geometry of MUON, Kovalev (2025) and Pethick et al. (2025b) link this family of algorithms to linear minimisation oracle (lmo) based algorithms, and establish convergence for arbitrary norms in finite-dimensional spaces. Guarantees for MUON with decoupled weight decay (Pethick et al., 2025b) and extensions to Nesterov momentum (Chen et al., 2025; Sfyraki and Wang, 2025) draw a close connection to stochastic conditional gradient (Mokhtari et al., 2020) and Lion-$\mathcal{K}$. G. Y. Kim and Oh (2026) analyse MUON with finite Newton–Schulz iterations under bounded Frobenius variance. Their approximation argument also applies to our heavy-tailed setting (Remark D.6). X. Zhang and Gao (2025) study heavy-tailed FedMuon with Frobenius smoothness, noise moments and stationarity.

While the above works make their noise assumptions in the Euclidean space, Kovalev and Borodich (2025) consider the weighted-Euclidean assumption $\mathbb{E}[\|\nabla f(x,\xi) - \nabla F(x)\|^2_{\Sigma^{-1}}] \leq \|\Sigma\|_{S_1}$, where Σ is a self-adjoint positive definite operator and $\|v\|^2_{\Sigma^{-1}} = v^\top \Sigma^{-1} v$.

Heavy-Tailed Noise. A separate line of work suggests that bounded variance may be too restrictive for modern deep learning, and instead suggests studying stochastic gradients with bounded p-th central moments for $p \in (1, 2]$ (Nemirovskij and Yudin, 1983; J. Zhang et al., 2020a; Garg et al., 2021). In Euclidean geometry, a large body of work establishes convergence of gradient-clipping based methods under heavy-tailed noise (e.g., J. Zhang et al., 2020a; Cutkosky and Mehta, 2021; Sadiev et al., 2023). Recent works establish optimal sample complexities for NSSD-M without the need for gradient-clipping under heavy-tailed noise (Hübler et al., 2025; Z. Liu and Zhou, 2025). Yu et al. (2026) analyse sign-based optimisers, including MUON, under weighted smoothness and structured, gradient-dependent heavy-tailed noise. Iiduka (2026) studies heavy-tailed MUON under Frobenius Hölder smoothness. The paper's comparison with SGD counts iterations under different batch-size requirements, so it does not by itself establish a sample-complexity separation. Beyond Euclidean spaces, Cutkosky and Mehta (2021) provide guarantees for 2-uniformly convex Banach spaces, but require an additional gradient clipping step. This extension covers Schatten-r norms for $r \in (1, 2]$, but not the Schatten-∞ geometry required for MUON. Complementary literature examines heavy-tailed *data* distributions (Kunstner and Bach, 2025; J. Kim et al., 2026).

Concurrent Work. While preparing this paper, we became aware of the concurrent preprint (Y. Zhang and Lin, 2026). The authors analyse MUON under heavy-tailed noise and (L_0, L_1)-smoothness, where both assumptions are stated in the Frobenius norm. This results in a $\mathcal{O}\Big(\rho^{\frac{3p-2}{p-1}} \frac{\Delta_1 L}{\varepsilon^2} \big(\frac{\sigma_F}{\varepsilon}\big)^{\frac{p}{p-1}}\Big)$ convergence guarantee, which does not improve over the stationarity adjusted guarantee of Euclidean algorithms (see Equation (5)). In contrast, our work considers general norms beyond S_∞, assumptions stated in the native optimisation geometry and obtains a strictly improved dimension dependence. Additionally we provide lower-bounds showing that our dimension dependence is first-order optimal.

6 Conclusion and Limitations

We provide convergence guarantees for normalised steepest descent with momentum in general Banach spaces under heavy-tailed noise. In particular, for MUON, we obtain a nuclear-stationarity guarantee with substantially improved dimension dependence compared to the corresponding stationarity-adjusted Euclidean guarantees. We further prove that the remaining dimension dependence is first-order optimal and therefore cannot be removed. Our experiments connect these results to empirical observations in language-model training: varying the Schatten-r descent geometry substantially improves perplexity over Euclidean descent for $r > 2$, and weight matrices trained by MUON exhibit favourable Schatten noise ratios.

Our empirical evaluation uses a 70M-parameter model trained on a single dataset, so the observed phenomena should be validated at larger scales and across datasets. Our lower bound matches the full stochastic rate for sufficiently large matrix aspect ratios. Determining the optimal sample complexity for smaller aspect ratios remains an interesting open question. Extending our expectation guarantees to high-probability bounds under heavy-tailed noise is another direction for future work.

Acknowledgments and Disclosure of Funding

FH gratefully acknowledges financial support from the ETH research grant and Swiss National Science Foundation (SNSF) Project Funding No. 200021-207343, the Alexander von Humboldt Foundation, and the European Union's Horizon Europe research and innovation programme under grant agreements No. 101120237 (ELIAS) and No. 101070617 (ELSA). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them. SS gratefully acknowledges generous support from the Alexander von Humboldt Foundation. This research was funded in part by the Swiss National Science Foundation (SNSF) under grant number P500-2_242086. We thank Jack Quinn for pointing out an error in the arXiv version of this paper. We also thank the anonymous reviewers for their helpful comments and suggestions. The authors declare no competing interests.

References

Kovarik, Zdislav (1970). "Some iterative methods for improving orthonormality". In: *SIAM Journal on Numerical Analysis* 7.3, pp. 386–389.

Nemirovskij, Arkadij Semenovič and David Borisovich Yudin (1983). *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons.

Ball, Keith, Eric A. Carlen, and Elliott H. Lieb (Dec. 1994). "Sharp Uniform Convexity and Smoothness Inequalities for Trace Norms". In: *Inventiones mathematicae* 115.1, pp. 463–482. ISSN: 1432-1297. DOI: 10.1007/BF01231769.

Hirzallah, Omar and Fuad Kittaneh (2002). "Non-Commutative Clarkson Inequalities for Unitarily Invariant Norms". In: *Pacific J. Math* 202.2, pp. 363–369.

Boyd, Stephen and Lieven Vandenberghe (2004). *Convex Optimization*. Cambridge: Cambridge University Press.

Bhatia, Rajendra (2013). *Matrix Analysis*. Springer Science & Business Media.

Kelner, Jonathan A. et al. (2014). "An Almost-Linear-Time Algorithm for Approximate Max Flow in Undirected Graphs, and Its Multicommodity Generalizations". In: *Proceedings of the 2014 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 217–226. DOI: 10.1137/1.9781611973402.16.

Kingma, Diederik P. and Jimmy Ba (2014). "Adam: A Method for Stochastic Optimization". In: *arXiv preprint arXiv:1412.6980*.

Mendel, Manor and Assaf Naor (2014). "Nonlinear Spectral Calculus and Super-Expanders". In: *Publications mathématiques de l'IHÉS* 119, pp. 1–95.

Carlson, David, Volkan Cevher, and Lawrence Carin (May 2015a). "Stochastic Spectral Descent for Restricted Boltzmann Machines". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Guy Lebanon and S. V. N. Vishwanathan. Vol. 38. Proceedings of Machine Learning Research. San Diego, California, USA: PMLR, pp. 111–119.

Carlson, David et al. (2015b). "Preconditioned Spectral Descent for Deep Learning". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/f50a6c02a3fc5a3a5d4d9391f05f3efc-Paper.pdf.

Pisier, Gilles (June 2016). *Martingales in Banach Spaces*. Cambridge University Press. ISBN: 978-1-316-48058-8. DOI: 10.1017/cbo9781316480588.

Bottou, Léon, Frank E. Curtis, and Jorge Nocedal (2018). "Optimization Methods for Large-Scale Machine Learning". In: *SIAM Review* 60.2, pp. 223–311. DOI: 10.1137/16M1080173. URL: https://doi.org/10.1137/16M1080173.

Simsekli, Umut, Levent Sagun, and Mert Gurbuzbalaban (June 2019). "A Tail-Index Analysis of Stochastic Gradient Noise in Deep Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 5827–5837.

Balles, Lukas, Fabian Pedregosa, and Nicolas Le Roux (2020). "The Geometry of Sign Gradient Descent". In: *arXiv preprint arXiv:2002.08056*. arXiv: 2002.08056.

Cutkosky, Ashok and Harsh Mehta (July 2020). "Momentum Improves Normalized SGD". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 2260–2268.

Mokhtari, Aryan, Hamed Hassani, and Amin Karbasi (2020). "Stochastic Conditional Gradient Methods: From Convex Minimization to Submodular Maximization". In: *Journal of Machine Learning Research* 21.105, pp. 1–49.

Zhang, Bohang et al. (2020). "Improved Analysis of Clipping Algorithms for Non-Convex Optimization". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 15511–15521.

Zhang, Jingzhao et al. (2020a). "Why are Adaptive Methods Good for Attention Models?" In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 15383–15393.

Zhang, Jingzhao et al. (2020b). "Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity". In: *International Conference on Learning Representations*.

Cutkosky, Ashok and Harsh Mehta (2021). "High-Probability Bounds for Non-Convex Stochastic Optimization with Heavy Tails". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 4883–4895.

Garg, Saurabh et al. (July 2021). "On Proximal Policy Optimization's Heavy-Tailed Gradients". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 3610–3619.

Zhao, Shen-Yi, Yin-Peng Xie, and Wu-Jun Li (Feb. 2021). "On the Convergence and Improvement of Stochastic Normalized Gradient Descent". In: *Science China Information Sciences* 64.3, p. 132103. ISSN: 1869-1919. DOI: 10.1007/s11432-020-3023-7.

Crawshaw, Michael et al. (2022). "Robustness to Unbounded Smoothness of Generalized SignSGD". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., pp. 9955–9968.

Hoffmann, Jordan et al. (2022). "An Empirical Analysis of Compute-Optimal Large Language Model Training". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., pp. 30016–30030.

Tuddenham, Mark, Adam Prügel-Bennett, and Jonathan Hare (2022). "Orthogonalising Gradients to Speed up Neural Network Optimisation". In: *arXiv preprint arXiv:2202.07052*. arXiv: 2202.07052.

Zhai, Xiaohua et al. (2022). "Scaling vision transformers". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12104–12113.

Arjevani, Yossi et al. (May 2023). "Lower Bounds for Non-Convex Stochastic Optimization". In: *Mathematical Programming* 199.1, pp. 165–214. ISSN: 1436-4646. DOI: 10.1007/s10107-022-01822-7.

Biderman, Stella et al. (2023). "Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling". In: *International Conference on Machine Learning*. PMLR, pp. 2397–2430.

Chowdhery, Aakanksha et al. (2023). "PaLM: Scaling Language Modeling with Pathways". In: *Journal of machine learning research* 24.240, pp. 1–113.

Faw, Matthew et al. (July 2023). "Beyond Uniform Smoothness: A Stopped Analysis of Adaptive SGD". In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Ed. by Gergely Neu and Lorenzo Rosasco. Vol. 195. Proceedings of Machine Learning Research. PMLR, pp. 89–160.

Kunstner, Frederik et al. (2023). "Noise Is Not the Main Factor Behind the Gap Between Sgd and Adam on Transformers, But Sign Descent Might Be". In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=a65YK0cqH8g.

Li, Haochuan, Alexander Rakhlin, and Ali Jadbabaie (2023). "Convergence of Adam Under Relaxed Assumptions". In: *Thirty-Seventh Conference on Neural Information Processing Systems*.

Sadiev, Abdurakhmon et al. (July 2023). "High-Probability Bounds for Stochastic Optimization and Variational Inequalities: The Case of Unbounded Variance". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 29563–29648.

Ahn, Kwangjun et al. (2024). "Linear attention is (maybe) all you need (to understand Transformer optimization)". In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=0uI5415ry7.

Battash, Barak, Lior Wolf, and Ofir Lindenbaum (May 2024). "Revisiting the Noise Model of Stochastic Gradient Descent". In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics* (May 2–4, 2024). Ed. by Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li. Vol. 238. Proceedings of Machine Learning Research. PMLR, pp. 4780–4788. URL: https://proceedings.mlr.press/v238/battash24a.html.

Bernstein, Jeremy and Laker Newhouse (2024). "Old Optimizer, New Norm: An Anthology". In: *arXiv preprint arXiv:2409.20325*.

Hägele, Alexander et al. (2024). "Scaling laws and compute-optimal training beyond fixed training durations". In: *Advances in Neural Information Processing Systems* 37, pp. 76232–76264.

Hu, Shengding et al. (2024). "MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies". In: *First Conference on Language Modeling*. URL: https://openreview.net/forum?id=3X2L2TFr0f.

Hübler, Florian et al. (May 2024). "Parameter-Agnostic Optimization under Relaxed Smoothness". In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. Ed. by Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li. Vol. 238. Proceedings of Machine Learning Research. PMLR, pp. 4861–4869. URL: https://proceedings.mlr.press/v238/hubler24a.html.

Jordan, Keller et al. (2024a). *modded-nanogpt: Speedrunning the NanoGPT baseline*. URL: https://github.com/KellerJordan/modded-nanogpt.

Jordan, Keller et al. (2024b). *Muon: An Optimizer for Hidden Layers in Neural Networks*. URL: https://kellerjordan.github.io/posts/muon/.
Penedo, Guilherme et al. (2024). "The fineweb datasets: Decanting the web for the finest text data at scale". In: *Advances in Neural Information Processing Systems* 37, pp. 30811–30849.
Vankov, Daniil et al. (2024). "Optimizing (L_0,L_1)-Smooth Functions by Gradient Methods". In: *arXiv preprint arXiv:2410.10800*. arXiv: 2410.10800.
Chen, Lizhang, Jonathan Li, and Qiang Liu (2025). "Muon Optimizes Under Spectral Norm Constraints". In: *arXiv preprint arXiv:2506.15054*. arXiv: 2506.15054.
Chezhegov, Savelii et al. (2025). "Convergence of Clipped-SGD for Convex (L_0, L_1)-Smooth Optimization with Heavy-Tailed Noise". In: *arXiv preprint arXiv:2505.20817*.
Hübler, Florian, Ilyas Fatkhullin, and Niao He (May 2025). "From Gradient Clipping to Normalization for Heavy Tailed SGD". In: *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*. Ed. by Yingzhen Li et al. Vol. 258. Proceedings of Machine Learning Research. PMLR, pp. 2413–2421. URL: https://proceedings.mlr.press/v258/hubler25a.html.
Kovalev, Dmitry (2025). "Understanding Gradient Orthogonalization for Deep Learning Via Non-Euclidean Trust-Region Optimization". In: *arXiv preprint arXiv:2503.12645*. arXiv: 2503.12645.
Kovalev, Dmitry and Ekaterina Borodich (2025). "Non-Euclidean SGD for Structured Optimization: Unified Analysis and Improved Rates". In: *arXiv preprint arXiv:2511.11466*. arXiv: 2511.11466.
Kunstner, Frederik and Francis Bach (2025). "Scaling Laws for Gradient Descent and Sign Descent for Linear Bigram Models under Zipf's Law". In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=VUbwLjLkws.
Li, Jiaxiang and Mingyi Hong (2025). "A Note on the Convergence of Muon". In: *arXiv preprint arXiv:2502.02900*. arXiv: 2502.02900.
Liu, Jingyuan et al. (2025). "Muon Is Scalable for LLM Training". In: *arXiv preprint arXiv:2502.16982*. arXiv: 2502.16982.
Liu, Zijian and Zhengyuan Zhou (2025). "Nonconvex Stochastic Optimization under Heavy-Tailed Noises: Optimal Convergence without Gradient Clipping". In: *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=NKotdPUc3L.
Pethick, Thomas et al. (2025a). "Generalized Gradient Norm Clipping & Non-Euclidean (L_0, L_1)-Smoothness". In: *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*.
Pethick, Thomas et al. (July 2025b). "Training Deep Learning Models with Norm-Constrained LMOs". In: *Proceedings of the 42nd International Conference on Machine Learning*. Ed. by Aarti Singh et al. Vol. 267. Proceedings of Machine Learning Research. PMLR, pp. 49069–49104.
Riabinin, Artem et al. (2025). "Gluon: Making Muon & Scion Great Again! (Bridging Theory and Practice of Lmo-Based Optimizers for Llms)". In: *arXiv preprint arXiv:2505.13416*. arXiv: 2505.13416.
Sfyraki, Maria-Eleni and Jun-Kun Wang (2025). "Lions and Muons: Optimization Via Stochastic Frank-Wolfe". In: *arXiv preprint arXiv:2506.04192*. arXiv: 2506.04192.
Shen, Wei et al. (2025). "On the Convergence Analysis of Muon". In: *arXiv preprint arXiv:2505.23737*. arXiv: 2505.23737.
Zhang, Xinwen and Hongchang Gao (2025). "On Provable Benefits of Muon in Federated Learning". In: *arXiv preprint arXiv:2510.03866*. arXiv: 2510.03866.
DeepSeek-AI et al. (2026). *DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence*. arXiv: 2606.19348 [cs.CL]. URL: https://arxiv.org/abs/2606.19348.
Iiduka, Hideaki (2026). "Muon Converges Under Heavy-Tailed Noise: Nonconvex Hölder-Smooth Empirical Risk Minimization". In: *arXiv preprint arXiv:2603.15059*. arXiv: 2603.15059.
Kim, Gyu Yeol and Min-hwan Oh (2026). "Convergence of Muon with Newton-Schulz". In: *International Conference on Learning Representations*. Ed. by C. Vondrick et al. Vol. 2026, pp. 45672–45718.
Kim, Juno et al. (2026). "Sharp Capacity Scaling of Spectral Optimizers in Learning Associative Memory". In: *arXiv preprint arXiv:2603.26554*.
Kimi Team et al. (2026). *Kimi K2: Open Agentic Intelligence*.
Yu, Dingzhi et al. (2026). "Sign-Based Optimizers Are Effective Under Heavy-Tailed Noise". In: *arXiv preprint arXiv:2602.07425*. arXiv: 2602.07425.

Zhang, Yubo and Junhong Lin (2026). “On Convergence of Muon for Nonconvex Stochastic Optimization Under Generalized Smoothness”. In: *TechRxiv* 2026.0227. DOI: 10.36227/techrxiv.177220100.05629800/v1.

Contents

1 Introduction **1**

2 Preliminaries **3**

2.1 Moment Bounds for General Norms 4

3 Main Result **4**

3.1 Guarantees for Uniformly Convex Norms 4

3.2 Guarantees for MUON 6

3.3 First-Order Lower Bounds for Nuclear-Norm Stationarity 7

4 Experiments **8**

5 Related Work **9**

6 Conclusion and Limitations **10**

A Banach Space Preliminaries **16**

B (L_0, L_1)-Smoothness **16**

B.1 Related Work on (L_0, L_1)-Smoothness 16

B.2 Properties of (L_0, L_1)-Smoothness in Banach Spaces 16

C Convergence Guarantees in General Norms **17**

C.1 Descent and Momentum Bounds 17

C.2 Convergence of NSSD-M 19

C.3 Extension to Stochastic Conditional Gradient 21

D Convergence Guarantees for MUON **23**

D.1 Banach Space Setting 23

D.2 Application to MUON 25

E Euclidean Lower Bound **27**

F First-Order Lower Bounds **30**

F.1 Convex Objectives 30

F.2 Non-Convex Objectives 33

G Experimental Details **36**

A Banach Space Preliminaries

Lemma A.1 (see Mendel and Naor, 2014, Equation (167)). *Let $(X, \|\cdot\|)$ be a Banach space.*

i) For all $1 < p' \leq p \leq 2$ we have $S_{p'}(X) \leq S_p(X)$.

ii) For all $2 \leq q \leq q' < \infty$ we have $K_{q'}(X) \leq K_q(X)$.

Lemma A.2 (see Bhatia, 2013). *Let $1 \leq p \leq q \leq \infty$. Then the Schatten-$p$ and Schatten-q norms satisfy*

$$\|A\|_{S_q} \leq \|A\|_{S_p} \leq \rho \|A\|_{S_q},$$

where $\rho = \min\{m, n\}^{\frac{1}{p}-\frac{1}{q}}$, for all $A \in \mathbb{R}^{m\times n}$.

For completeness, we reproduce the argument of Pisier (2016, Theorem 10.22) in our notation.

Proof of Theorem 2.4. It suffices to consider $S := S_p(\mathcal{B}^*) < \infty$ and increments with finite p-th moments. Write $M_t := \sum_{s=1}^t Y_s$, $M_0 := 0$, and let $(\mathcal{F}_t)_{t=0}^n$ be the underlying filtration. By conditional Jensen's inequality and the martingale difference property,

$$\mathbb{E}\left[\|M_{t-1} - Y_t\|_*^p \,|\, \mathcal{F}_{t-1}\right] \geq \|M_{t-1} - \mathbb{E}\left[Y_t \,|\, \mathcal{F}_{t-1}\right]\|_*^p = \|M_{t-1}\|_*^p.$$

By Definition 2.1 and the preceding inequality,

$$\begin{aligned}\mathbb{E}\left[\|M_t\|_*^p \,|\, \mathcal{F}_{t-1}\right] &\leq 2\|M_{t-1}\|_*^p + 2S^p\mathbb{E}\left[\|Y_t\|_*^p \,|\, \mathcal{F}_{t-1}\right] - \mathbb{E}\left[\|M_{t-1} - Y_t\|_*^p \,|\, \mathcal{F}_{t-1}\right] \\ &\leq \|M_{t-1}\|_*^p + 2S^p\mathbb{E}\left[\|Y_t\|_*^p \,|\, \mathcal{F}_{t-1}\right].\end{aligned}$$

Taking expectations, summing over $t = 1, \ldots, n$, and using $M_0 = 0$ finishes the proof. □

B (L_0, L_1)-Smoothness

We will derive our results for the weaker smoothness notion of (L_0, L_1)-smoothness. This section contains the related work, and required definitions and properties of this smoothness notion in general Banach spaces.

B.1 Related Work on (L_0, L_1)-Smoothness

Motivated by observations in language modelling, J. Zhang et al. (2020b) propose considering (L_0, L_1)-smoothness, allowing the smoothness to grow affinely with the gradient norm. A long line of works analyse diverse algorithms in this setting (e.g., Crawshaw et al., 2022; Faw et al., 2023; H. Li et al., 2023; Chezhegov et al., 2025). In Euclidean geometry, B. Zhang et al. (2020) and Zhao et al. (2021) prove the convergence of NSSD-M, and Z. Liu and Zhou (2025) extend this result to the heavy-tailed setting. Beyond Euclidean geometry, Balles et al. (2020) provide guarantees for the ℓ_∞-geometry in the deterministic setting, and Pethick et al. (2025a) provide convergence guarantees in arbitrary finite-dimensional spaces in the stochastic setting by adding gradient-clipping, while Riabinin et al. (2025)[5] concurrently prove convergence without the clipping step.

B.2 Properties of (L_0, L_1)-Smoothness in Banach Spaces

Definition B.1 ((L_0, L_1)-smoothness). Let $L_0, L_1 \geq 0$. A function $F\colon \mathcal{B} \to \mathbb{R}$ is said to be (L_0, L_1)-*smooth* if it is differentiable and, for all $x, y \in \mathcal{B}$,

$$\|\nabla F(x) - \nabla F(y)\|_* \leq (L_0 + L_1\|\nabla F(x)\|_*)A(L_1\|x - y\|)\|x - y\|,$$

where $A(c) := \frac{e^c - 1}{c}$, with $A(0) = 1$ by continuity.

[5] We note that this preprint has a typo. While their variance assumption (Assumption 2) is stated in the dual norm, the authors use the inner product structure of the Frobenius norm in the proof of Theorem 2. In particular, a fix requires switching to σ_F and introduces a ρ^2 dependence.

Lemma B.2 (Smoothness Upper-Bound)**.** *Let $F\colon \mathcal{B} \to \mathbb{R}$ be an (L_0, L_1)-smooth function. Then, for all $x, y \in \mathcal{B}$,*

$$F(y) \leq F(x) + \langle \nabla F(x), y - x \rangle + (L_0 + L_1 \|\nabla F(x)\|_*) B(L_1 \|x - y\|) \|x - y\|^2 ,$$

where $B(c) = \frac{e^c - 1 - c}{c^2}$, with $B(0) = 1/2$ by continuity.

Proof. This directly follows with the same proof as (Vankov et al., 2024, Lemma 2.2) by noting that all used inequalities also hold in general Banach spaces. □

C Convergence Guarantees in General Norms

This section contains generalisations of the main results from Section 3.1 and their proofs. We will derive the unconstrained guarantees for the weaker (L_0, L_1)-smoothness notion popularised by J. Zhang et al. (2020b).

Assumption 4 ((L_0, L_1)-smoothness)**.** We assume the objective function F is (L_0, L_1)-smooth in the sense of Definition B.1.

C.1 Descent and Momentum Bounds

We first extend the descent lemma of Hübler et al. (2024, Lemma 12) to Banach spaces.

Lemma C.1 (Descent Lemma)**.** *Suppose Assumptions 1 and 4 hold. Furthermore let $T \in \mathbb{N}_{\geq 1}$ and consider the iterates (x_t) generated by* NSSD-M*. Then*

$$\sum_{t=1}^{T} \left(\eta_t - L_1 \eta_t^2 B_t\right) \|\nabla F(x_t)\|_* \leq \Delta_1 + L_0 \sum_{t=1}^{T} \eta_t^2 B_t + 2 \sum_{t=1}^{T} \eta_t \|\nabla F(x_t) - m_t\|_* ,$$

where $B_t \coloneqq B(L_1 \eta_t)$, $B(c) = \frac{e^c - 1 - c}{c^2}$.

Proof. Using Lemma B.2 we get

$$\begin{aligned} F(x_{t+1}) - F(x_t) &\leq \langle \nabla F(x_t), x_{t+1} - x_t \rangle + (L_0 + L_1 \|\nabla F(x_t)\|_*) B(L_1 \eta_t) \|x_{t+1} - x_t\|^2 \\ &\leq \eta_t \langle \nabla F(x_t), \operatorname{lmo}(m_t) \rangle + (L_0 + L_1 \|\nabla F(x_t)\|_*) B(L_1 \eta_t) \eta_t^2 . \end{aligned}$$

Furthermore

$$\begin{aligned} \langle \nabla F(x_t), \operatorname{lmo}(m_t) \rangle &= \langle \nabla F(x_t) - m_t, \operatorname{lmo}(m_t) \rangle + \langle m_t, \operatorname{lmo}(m_t) \rangle \\ &\leq \|\nabla F(x_t) - m_t\|_* \|\operatorname{lmo}(m_t)\| - \|m_t\|_* \\ &\leq 2 \|\nabla F(x_t) - m_t\|_* - \|\nabla F(x_t)\|_* , \end{aligned}$$

where we used the definition of lmo and the dual norm in the first, and the triangle inequality in the second step. Combining yields

$$\begin{aligned} F(x_{t+1}) - F(x_t) &\leq \eta_t \left(2 \|\nabla F(x_t) - m_t\|_* - \|\nabla F(x_t)\|_*\right) + (L_0 + L_1 \|\nabla F(x_t)\|_*) B(L_1 \eta_t) \eta_t^2 \\ &= \left(L_1 \eta_t^2 B(L_1 \eta_t) - \eta_t\right) \|\nabla F(x_t)\|_* + 2 \eta_t \|\nabla F(x_t) - m_t\|_* + L_0 B(L_1 \eta_t) \eta_t^2 . \end{aligned}$$

Summing up and re-arranging finishes the proof. □

Next we derive the main technical lemma of our analysis, which bounds the deviation of the momentum from the true gradient for general norms, without requiring a conversion to the Euclidean norm.

Lemma C.2 (Momentum Deviation Bound)**.** *Suppose Assumptions 3 and 4 hold and that $(\mathcal{B}^*, \|\cdot\|_*)$ is p'-uniformly smooth with smoothness constant $S_{p'}(\mathcal{B}^*) \leq S$. Then the iterates generated by* NSSD-M *satisfy*

$$\begin{aligned} \mathbb{E}\left[\|\nabla F(x_t) - m_t\|_*\right] \leq a_t \|\nabla F(x_1)\|_* &+ 2\sigma S \left(\sum_{\tau=1}^{t} b_{\tau,t}^r\right)^{1/r} \\ &+ L_0 \sum_{\tau=2}^{t} c_{\tau,t} + L_1 \sum_{\tau=2}^{t} c_{\tau,t} \mathbb{E}\left[\|\nabla F(x_\tau)\|_*\right] , \end{aligned}$$

where $r \coloneqq \min\{p, p'\}$ *and, for easier readability, we define*

$$a_t \coloneqq \prod_{\kappa=1}^{t} \beta_\kappa, \qquad b_{\tau,t} \coloneqq (1-\beta_\tau) \prod_{\kappa=\tau+1}^{t} \beta_\kappa \quad (\tau \geq 1),$$

$$c_{\tau,t} \coloneqq \eta_{\tau-1} A_{\tau-1} \prod_{\kappa=\tau}^{t} \beta_\kappa \quad (\tau \geq 2),$$

where $A_\tau = A(L_1\eta_\tau) = \frac{e^{L_1\eta_\tau}-1}{L_1\eta_\tau}$. *In particular, for constant parameters* $\eta_t \equiv \eta$ *and* $\beta_t \equiv \beta$, *we get*

$$\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(x_t) - m_t\|_*\right] \leq \|\nabla F(x_1)\|_* \frac{\beta(1-\beta^T)}{1-\beta} + 2\sigma S T (1-\beta)^{\frac{r-1}{r}}$$
$$+ L_0 \frac{\eta A T}{1-\beta} + L_1 \frac{\eta A}{1-\beta} \sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(x_t)\|_*\right],$$

where $A \equiv A(L_1\eta)$.

Proof. First note that the standard decomposition of $\mu_t \coloneqq m_t - \nabla F(x_t)$ into initialisation, noise and smoothness terms (Cutkosky and Mehta, 2020),

$$\mu_t = -a_t \nabla F(x_1) + \sum_{\tau=1}^{t} b_{\tau,t}(\nabla f(x_\tau, \xi_\tau) - \nabla F(x_\tau)) + \sum_{\tau=2}^{t} \prod_{\kappa=\tau}^{t} \beta_\kappa (\nabla F(x_{\tau-1}) - \nabla F(x_\tau)),$$

is norm independent and hence also holds in general Banach spaces. Next, we have

$$\|\nabla F(x_\tau) - \nabla F(x_{\tau-1})\|_* \leq (L_0 + L_1 \|\nabla F(x_\tau)\|_*) A(L_1\eta_{\tau-1}) \|x_\tau - x_{\tau-1}\| \leq L_\tau A_{\tau-1}\eta_{\tau-1},$$

where we use the slight abuse of notation $L_\tau \coloneqq L_0 + L_1 \|\nabla F(x_\tau)\|_*$. Hence, by our definition of $b_{\tau,t}$ and $c_{\tau,t}$ in the statement, the triangle inequality yields

$$\begin{aligned}
\mathbb{E}\left[\|\mu_t\|_*\right] &\leq a_t \|\nabla F(x_1)\|_* + \mathbb{E}\left[\left\|\sum_{\tau=1}^{t} b_{\tau,t}(\nabla f(x_\tau, \xi_\tau) - \nabla F(x_\tau))\right\|_*\right] \\
&\quad + \sum_{\tau=2}^{t} \left(\prod_{\kappa=\tau}^{t} \beta_\kappa\right) \mathbb{E}\left[\|\nabla F(x_{\tau-1}) - \nabla F(x_\tau)\|_*\right] \\
&\leq a_t \|\nabla F(x_1)\|_* + \mathbb{E}\left[\left\|\sum_{\tau=1}^{t} b_{\tau,t}(\nabla f(x_\tau, \xi_\tau) - \nabla F(x_\tau))\right\|_*\right] + \sum_{\tau=2}^{t} c_{\tau,t} \mathbb{E}\left[L_\tau\right].
\end{aligned}$$

Now note that $(\nabla f(x_\tau, \xi_\tau) - \nabla F(x_\tau))_{\tau=1}^{t}$ is a martingale difference sequence in $\mathcal{B}^*$, and, by Lemma A.1, $\mathcal{B}^*$ is also r-uniformly smooth with $S_r(\mathcal{B}^*) \leq S_{p'}(\mathcal{B}^*) \leq S$. Hence, we may apply Theorem 2.4 to get

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{\tau=1}^{t} b_{\tau,t}(\nabla f(x_\tau, \xi_\tau) - \nabla F(x_\tau))\right\|_*\right] &\leq \mathbb{E}\left[\left\|\sum_{\tau=1}^{t} b_{\tau,t}(\nabla f(x_\tau, \xi_\tau) - \nabla F(x_\tau))\right\|_*^r\right]^{1/r} \\
&\leq 2S\left(\sum_{\tau=1}^{t} b_{\tau,t}^r \mathbb{E}\left[\|\nabla f(x_\tau, \xi_\tau) - \nabla F(x_\tau)\|_*^r\right]\right)^{1/r} \\
&\leq 2\sigma S\left(\sum_{\tau=1}^{t} b_{\tau,t}^r\right)^{1/r},
\end{aligned}$$

where we used $r \leq p$ and Assumption 3 in the last step. For the constant parameter version note that

$$\sum_{\tau=1}^{t} b_{\tau,t}^r = (1-\beta)^r \sum_{\tau=1}^{t} \beta^{r(t-\tau)} \leq \frac{(1-\beta)^r}{1-\beta^r} \leq (1-\beta)^{r-1}$$

$$\sum_{\tau=2}^{t} c_{\tau,t} = \eta A \sum_{\tau=1}^{t-1} \beta^\tau \leq \frac{\eta A}{1-\beta},$$

where $A \equiv A(L_1\eta)$, and hence

$$
\begin{aligned}
2\sigma S \sum_{t=1}^{T} \left(\sum_{\tau=1}^{t} b_{\tau,t}^{r} \right)^{1/r} &\leq 2\sigma S T (1-\beta)^{\frac{r-1}{r}} \\
L_0 \sum_{t=1}^{T} \sum_{\tau=2}^{t} c_{\tau,t} &\leq L_0 \frac{\eta A T}{1-\beta} \\
L_1 \sum_{t=1}^{T} \sum_{\tau=2}^{t} c_{\tau,t} \mathbb{E}\left[\|\nabla F(x_\tau)\|_* \right] &= L_1 \sum_{\tau=2}^{T} \left(\sum_{t=\tau}^{T} c_{\tau,t} \right) \mathbb{E}\left[\|\nabla F(x_\tau)\|_* \right] \\
&\leq L_1 \sum_{\tau=2}^{T} \frac{\eta A}{1-\beta} \mathbb{E}\left[\|\nabla F(x_\tau)\|_* \right] \\
&\leq \frac{L_1 \eta A}{1-\beta} \sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(x_t)\|_* \right],
\end{aligned}
$$

The initialisation term satisfies $\|\nabla F(x_1)\|_* \sum_{t=1}^{T} a_t = \|\nabla F(x_1)\|_* \beta(1-\beta^T)/(1-\beta)$. Combining the four bounds finishes the proof. □

Proof of Lemma 3.3. By Lemma 2.2, the dual space is p-uniformly smooth with constant K. Applying Lemma C.2 with $L_0 = L$, $L_1 = 0$ and $r = p$, and dividing by T, gives

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(x_t) - m_t\|_* \right] \leq \frac{\|\nabla F(x_1)\|_*}{T} \sum_{t=1}^{T} \beta^t + 2\sigma K (1-\beta)^{\frac{p-1}{p}} + \frac{\eta L}{1-\beta}.
$$

Assumptions 1 and 2 give $\|\nabla F(x_1)\|_* \leq \sqrt{2\Delta_1 L}$, while Cauchy–Schwarz yields $\sum_{t=1}^{T} \beta^t \leq \sqrt{T/(1-\beta)}$. Substituting these bounds proves the claim. □

C.2 Convergence of NSSD-M

Finally we are able to derive the main convergence result for NSSD-M in general Banach spaces.

Theorem C.3. *Suppose Assumptions 1, 3 and 4 hold and that $(\mathcal{B}, \|\cdot\|)$ is q-uniformly convex with constant $K_q(\mathcal{B}) \leq K$. Furthermore let p' be the conjugate exponent of q, i.e., $1/q + 1/p' = 1$, and $r = \min\{p, p'\}$. Then the iterates generated by NSSD-M with*

$$
\begin{aligned}
\eta_t \equiv \eta &= \min\left\{ \frac{1-\beta}{5L_1}, \sqrt{\frac{\Delta_1(1-\beta)}{L_0 T}} \right\}, \\
\beta_t \equiv \beta &= 1 - \min\left\{ 1, \max\left\{ \left(\frac{\Delta_1 L_0}{\sigma^2 K^2 T} \right)^{\frac{r}{3r-2}}, \left(\frac{5L_1\Delta_1}{2\sigma K T} \right)^{\frac{r}{2r-1}} \right\} \right\}
\end{aligned}
$$

satisfy

$$
\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(x_t)\|_* \right] \leq 30 \frac{\Delta_1 L_1}{T} + 18 \sqrt{\frac{\Delta_1 L_0}{T}} + 30 \left(\frac{\Delta_1 L_1 (\sigma K)^{\frac{r}{r-1}}}{T} \right)^{\frac{r-1}{2r-1}} \\
+ 28 \left(\frac{\Delta_1 L_0 (\sigma K)^{\frac{r}{(r-1)}}}{T} \right)^{\frac{r-1}{3r-2}}.
\end{aligned}
$$

In particular, the sample complexity to reach an ε-$\|\cdot\|_*$-stationary point is at most

$$
\mathcal{O}\left(\frac{\Delta_1 L_1}{\varepsilon} + \frac{\Delta_1 L_0}{\varepsilon^2} + \frac{\Delta_1 L_1}{\varepsilon} \left(\frac{\sigma K}{\varepsilon} \right)^{\frac{r}{r-1}} + \frac{\Delta_1 L_0}{\varepsilon^2} \left(\frac{\sigma K}{\varepsilon} \right)^{\frac{r}{r-1}} \right).
$$

Proof. Write $G_1 \coloneqq \|\nabla F(x_1)\|_*$. Following the argument of Vankov et al. (2024, Lemma B.1), using Lemma B.2 and lower boundedness, we obtain

$$G_1 \leq 2\sqrt{\Delta_1 L_0} + 4\Delta_1 L_1. \tag{7}$$

Applying Lemma C.1 yields

$$\begin{aligned}
\eta \sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(x_t)\|_*\right] &\leq \Delta_1 + L_0\eta^2 BT + L_1\eta^2 B \sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(x_t)\|_*\right] \\
&\quad + 2\sum_{t=1}^{T} \eta_t \mathbb{E}\left[\|\nabla F(x_t) - m_t\|_*\right] \\
&\leq \Delta_1 + L_0\eta^2 T\left(B + \frac{2A}{1-\beta}\right) + 2\eta G_1 \min\left\{T, \frac{1}{1-\beta}\right\} \\
&\quad + 4\sigma\eta KT(1-\beta)^{\frac{r-1}{r}} \\
&\quad + L_1\eta^2\left(B + \frac{2A}{1-\beta}\right)\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(x_t)\|_*\right].
\end{aligned}$$

Next we calculate

$$L_1\eta^2\left(B + \frac{2A}{1-\beta}\right) \leq \eta\left(\frac{L_1\eta}{1-\beta}\frac{A(5-\beta)}{2}\right) \leq \frac{5}{9}\eta, \tag{8}$$

where we used $B \leq A/2$ in the first, and our choice of η, which implies $\frac{L_1\eta}{1-\beta} \leq \frac{1}{5}$ and $A \leq 10/9$, in the last inequality. Hence, re-arranging further yields

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(x_t)\|_*\right] &\leq \frac{9\Delta_1}{4\eta T} + \frac{6L_0\eta A}{(1-\beta)} + \frac{9}{2}G_1 \min\left\{1, \frac{1}{T(1-\beta)}\right\} + 10\sigma K(1-\beta)^{\frac{r-1}{r}} \\
&\leq 18\sqrt{\frac{\Delta_1 L_0}{T(1-\beta)}} + \frac{30\Delta_1 L_1}{T(1-\beta)} + 10\sigma K(1-\beta)^{\frac{r-1}{r}},
\end{aligned} \tag{9}$$

where we used that $1/\min\{a,b\} \leq 1/a + 1/b$ and the definition of η in the last step. The initialisation term is absorbed using (7) and $\min\left\{1, z^{-1}\right\} \leq z^{-1/2}$ for $z > 0$. Next we upper bound the remaining 3 terms by our choice of momentum. To simplify exposition, let us denote

$$u \coloneqq \left(\frac{\Delta_1 L_0}{\sigma^2 K^2 T}\right)^{\frac{r}{3r-2}}, \qquad v \coloneqq \left(\frac{5L_1\Delta_1}{2\sigma KT}\right)^{\frac{r}{2r-1}}$$

such that $1 - \beta = \min\{1, \max\{u, v\}\}$. First note that, due to $\frac{1}{\sqrt{1-\beta}} \leq \frac{1}{\min\left\{1,\sqrt{u}\right\}} \leq 1 + \frac{1}{\sqrt{u}}$,

$$18\sqrt{\frac{\Delta_1 L_0}{T(1-\beta)}} \leq 18\sqrt{\frac{\Delta_1 L_0}{T}} + 18\left(\frac{\Delta_1 L_0 (\sigma K)^{\frac{r}{(r-1)}}}{T}\right)^{\frac{r-1}{3r-2}}.$$

For the second term we use the same argument as before, this time using v, to get

$$\frac{30\Delta_1 L_1}{T(1-\beta)} \leq 30\frac{\Delta_1 L_1}{T} + 30\left(\frac{2}{5}\right)^{\frac{r}{2r-1}}\left(\frac{\Delta_1 L_1 (\sigma K)^{\frac{r}{r-1}}}{T}\right)^{\frac{r-1}{2r-1}}.$$

Finally, for the last term we use $(1-\beta)^\alpha \leq u^\alpha + v^\alpha$ to get

$$\begin{aligned}
10\sigma K(1-\beta)^{\frac{r-1}{r}} &\leq 10\sigma K\left(\frac{\Delta_1 L_0}{\sigma^2 K^2 T}\right)^{\frac{r-1}{3r-2}} + 10\sigma K\left(\frac{5L_1\Delta_1}{2\sigma KT}\right)^{\frac{r-1}{2r-1}} \\
&= 10\left(\frac{\Delta_1 L_0 (\sigma K)^{\frac{r}{r-1}}}{T}\right)^{\frac{r-1}{3r-2}} + 10\left(\frac{5\Delta_1 L_1 (\sigma K)^{\frac{r}{r-1}}}{2T}\right)^{\frac{r-1}{2r-1}}.
\end{aligned}$$

Plugging these bounds into (9) and using $22\left(\frac{5}{2}\right)^{\frac{r-1}{2r-1}} \leq 30$ finishes the proof. □

Theorem 3.1 follows from Theorem C.3 by taking $L_0 = L$, $L_1 = 0$ and $p' = p$. For Schatten-r and ℓ_r spaces with $2 \le r \le p/(p-1)$, Example 2.3 and lemma A.1 give $K_{p/(p-1)} = 1$. Hence Corollary 3.2 follows by taking $K = 1$ and choosing τ uniformly from $[T]$.

Remark C.4. The above result can readily be extended to the parameter-agnostic setting by using the same arguments as in (Hübler et al., 2024) to handle the non-constant parameters and error accumulation before $\eta_t \lesssim 1/L_1$.

C.3 Extension to Stochastic Conditional Gradient

Let $\mathcal{D} \subset \mathcal{B}$ be convex and bounded, with diameter

$$D_{\mathcal{D}} \coloneqq \sup_{x,y \in \mathcal{D}} \|x - y\| .$$

Consider

$$\min_{x \in \mathcal{D}} F(x),$$

and assume that the linear minimisation oracle over $\mathcal{D}$ is exact, i.e., each subproblem $\inf_{u \in \mathcal{D}} \langle \phi, u \rangle$ admits a minimiser. We consider the stochastic conditional gradient (SCG) iteration

$$\begin{aligned} m_t &\leftarrow \beta_t m_{t-1} + (1-\beta_t)\nabla f(x_t, \xi_t), \\ v_t &\in \operatorname*{argmin}_{u \in \mathcal{D}} \langle m_t, u \rangle , \\ x_{t+1} &\leftarrow (1-\eta_t)x_t + \eta_t v_t, \end{aligned} \tag{SCG}$$

where $m_0 \coloneqq 0$, $x_1 \in \mathcal{D}$, $\eta_t \in [0,1]$, and $\beta_t \in [0,1)$. We measure stationarity by the Frank-Wolfe gap

$$g_{\mathcal{D}}(x) \coloneqq \sup_{u \in \mathcal{D}} \langle \nabla F(x), x - u \rangle .$$

For the analysis, define $\widehat{m}_t \coloneqq m_t / (1 - \prod_{\tau=1}^{t} \beta_\tau)$. This positive rescaling leaves the linear minimisation oracle unchanged.

Lemma C.5 (Constrained descent bound)**.** *Suppose Assumption 2 holds. Let $\mu_t \coloneqq \widehat{m}_t - \nabla F(x_t)$. Then, for every $t \ge 1$,*

$$\eta_t \mathbb{E}\left[g_{\mathcal{D}}(x_t)\right] \le \mathbb{E}\left[F(x_t) - F(x_{t+1})\right] + \eta_t D_{\mathcal{D}} \mathbb{E}\left[\|\mu_t\|_*\right] + \tfrac{L}{2}\eta_t^2 D_{\mathcal{D}}^2.$$

Proof. For any $u \in \mathcal{D}$, smoothness and $x_{t+1} = x_t + \eta_t(v_t - x_t)$ give

$$\begin{aligned} F(x_{t+1}) &\le F(x_t) + \eta_t \langle \nabla F(x_t), v_t - x_t \rangle + \tfrac{L}{2}\eta_t^2 \|v_t - x_t\|^2 \\ &\le F(x_t) + \eta_t \langle \nabla F(x_t), u - x_t \rangle + \eta_t \langle \mu_t, u - v_t \rangle + \tfrac{L}{2}\eta_t^2 \|v_t - x_t\|^2 , \end{aligned}$$

where we used $\langle m_t, v_t \rangle \le \langle m_t, u \rangle$. Since $u, v_t, x_t \in \mathcal{D}$,

$$\eta_t \langle \nabla F(x_t), x_t - u \rangle \le F(x_t) - F(x_{t+1}) + \eta_t D_{\mathcal{D}} \|\mu_t\|_* + \tfrac{L}{2}\eta_t^2 D_{\mathcal{D}}^2.$$

Taking the supremum over u and then expectations proves the claim. □

The proof relies on an error control lemma analogous to Lemma C.2. Here we bound the error of $\widehat{m}_t$, taking the normalisation and the domain diameter into account.

Lemma C.6 (Momentum error for SCG)**.** *Suppose Assumptions 2 and 3 hold, and that $(\mathcal{B}, \|\cdot\|)$ is q-uniformly convex with q-convexity constant K. Let p' be the conjugate exponent of q and set $r \coloneqq \min\{p, p'\}$. For constant parameters $\eta_t \equiv \eta$, and $\beta_t \equiv \beta \in [0,1)$,*

$$\textstyle\sum_{t=1}^{T} \mathbb{E}\left[\|\widehat{m}_t - \nabla F(x_t)\|_*\right] \le 4\sigma K T^{1/r} + 2\sigma K T (1-\beta)^{\frac{r-1}{r}} + \frac{\eta L D_{\mathcal{D}} T}{1-\beta}.$$

Proof. Writing $\delta_t \coloneqq \nabla f(x_t, \xi_t) - \nabla F(x_t)$, unrolling the momentum recursion gives

$$\mu_t \coloneqq \widehat{m}_t - \nabla F(x_t) = \sum_{\tau=1}^{t} b_{\tau,t}\delta_\tau + \sum_{\tau=2}^{t} c_{\tau,t}(\nabla F(x_{\tau-1}) - \nabla F(x_\tau)),$$

where $b_{\tau,t} = (1-\beta)\beta^{t-\tau}/(1-\beta^t)$ and $c_{\tau,t} = \beta^{t-\tau+1}(1-\beta^{\tau-1})/(1-\beta^t)$. By Assumption 2,

$$\|\nabla F(x_{\tau-1}) - \nabla F(x_\tau)\|_* \le L\|x_{\tau-1} - x_\tau\|.$$

The SCG update implies $x_t - x_{t-1} = \eta(v_{t-1} - x_{t-1})$ with $x_{t-1}, v_{t-1} \in \mathcal{D}$, hence

$$\|x_t - x_{t-1}\| \le \eta D_{\mathcal{D}}.$$

Therefore

$$\mathbb{E}\left[\|\mu_t\|_*\right] \le \mathbb{E}\left[\left\|\sum_{\tau=1}^t b_{\tau,t}\delta_\tau\right\|_*\right] + \eta L D_{\mathcal{D}} \sum_{\tau=2}^t c_{\tau,t}.$$

Since $(\delta_\tau)_{\tau=1}^t$ is a martingale difference sequence in $\mathcal{B}^*$, the same application of Theorem 2.4 and lemma A.1 as in Lemma C.2 yields

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^t b_{\tau,t}\delta_\tau\right\|_*\right] \le 2\sigma K\left(\sum_{\tau=1}^t b_{\tau,t}^r\right)^{1/r}.$$

Since $\sum_{\tau=1}^t b_{\tau,t} = 1$ and

$$b_{\tau,t} \le b_{t,t} = 1 - \beta + \frac{(1-\beta)\beta^t}{1-\beta^t} \le 1 - \beta + \frac{1}{t},$$

the coefficients satisfy

$$\sum_{t=1}^T \left(\sum_{\tau=1}^t b_{\tau,t}^r\right)^{1/r} \le \sum_{t=1}^T \left(1-\beta+\frac{1}{t}\right)^{\frac{r-1}{r}} \le T(1-\beta)^{\frac{r-1}{r}} + 2T^{1/r},$$

$$\sum_{\tau=2}^t c_{\tau,t} \le \sum_{\tau=1}^{t-1} \beta^\tau \le \frac{1}{1-\beta},$$

where we used $\sum_{t=1}^T t^{-(r-1)/r} \le rT^{1/r} \le 2T^{1/r}$ in the first line. Combining these bounds proves the claim. □

We are now ready to state the convergence theorem for SCG.

Theorem C.7 (Frank-Wolfe gap bound for SCG)**.** *Suppose Assumptions 2 and 3 hold, and that $(\mathcal{B}, \|\cdot\|)$ is q-uniformly convex with q-convexity constant K. Let p' be the conjugate exponent of q, $r \coloneqq \min\{p, p'\}$, and*

$$\Delta_{\mathcal{D}} \coloneqq F(x_1) - \inf_{x\in\mathcal{D}} F(x).$$

Consider SCG with

$$\eta_t \equiv \eta \coloneqq \min\left\{1, \sqrt{\tfrac{\Delta_{\mathcal{D}}(1-\beta)}{LD_{\mathcal{D}}^2 T}}\right\},$$

$$\beta_t \equiv \beta \coloneqq 1 - \min\left\{1, \max\left\{\left(\tfrac{\Delta_{\mathcal{D}} L}{\sigma^2 K^2 T}\right)^{\frac{r}{3r-2}}, T^{-1}\right\}\right\}.$$

Then

$$\tfrac{1}{T}\textstyle\sum_{t=1}^T \mathbb{E}\left[g_{\mathcal{D}}(x_t)\right] \le \tfrac{\Delta_{\mathcal{D}}}{T} + 3D_{\mathcal{D}}\sqrt{\tfrac{\Delta_{\mathcal{D}} L}{T}} + 4D_{\mathcal{D}}\left(\tfrac{\Delta_{\mathcal{D}} L(\sigma K)^{\frac{r}{r-1}}}{T}\right)^{\frac{r-1}{3r-2}} + 6D_{\mathcal{D}}\tfrac{\sigma K}{T^{\frac{r-1}{r}}}.$$

Proof. The case $\Delta_{\mathcal{D}} = 0$ is immediate, so assume $\Delta_{\mathcal{D}} > 0$. Summing Lemma C.5 and telescoping gives

$$\eta\textstyle\sum_{t=1}^T \mathbb{E}\left[g_{\mathcal{D}}(x_t)\right] \le \Delta_{\mathcal{D}} + \eta D_{\mathcal{D}} \sum_{t=1}^T \mathbb{E}\left[\|\mu_t\|_*\right] + \frac{L}{2}\eta^2 D_{\mathcal{D}}^2 T.$$

Plugging in Lemma C.6 yields

$$\tfrac{1}{T}\textstyle\sum_{t=1}^T \mathbb{E}\left[g_{\mathcal{D}}(x_t)\right] \le \tfrac{\Delta_{\mathcal{D}}}{\eta T} + 4D_{\mathcal{D}}\tfrac{\sigma K}{T^{\frac{r-1}{r}}} + 2D_{\mathcal{D}}\sigma K(1-\beta)^{\frac{r-1}{r}} + \tfrac{\eta L D_{\mathcal{D}}^2}{1-\beta} + \tfrac{L\eta D_{\mathcal{D}}^2}{2}.$$

Let $\bar{\eta} := \sqrt{\frac{\Delta_{\mathcal{D}}(1-\beta)}{LD_{\mathcal{D}}^2 T}}$, so that $\eta = \min\{1, \bar{\eta}\}$. We have that $\frac{1}{\eta} \leq 1 + \frac{1}{\bar{\eta}}$ and $\eta \leq \bar{\eta}$. Consequently,

$$\frac{\Delta_{\mathcal{D}}}{\eta T} \leq \frac{\Delta_{\mathcal{D}}}{T} + D_{\mathcal{D}}\sqrt{\frac{\Delta_{\mathcal{D}} L}{T(1-\beta)}}, \quad \frac{\eta L D_{\mathcal{D}}^2}{1-\beta} \leq D_{\mathcal{D}}\sqrt{\frac{\Delta_{\mathcal{D}} L}{T(1-\beta)}}, \quad \frac{L\eta D_{\mathcal{D}}^2}{2} \leq \frac{D_{\mathcal{D}}}{2}\sqrt{\frac{\Delta_{\mathcal{D}} L(1-\beta)}{T}} \leq \frac{D_{\mathcal{D}}}{2}\sqrt{\frac{\Delta_{\mathcal{D}} L}{T}}.$$

Next, let

$$U := \left(\frac{\Delta_{\mathcal{D}} L (\sigma K)^{\frac{r}{r-1}}}{T}\right)^{\frac{r-1}{3r-2}}, \qquad W := \frac{\sigma K}{T^{\frac{r-1}{r}}}, \qquad u := \left(\frac{\Delta_{\mathcal{D}} L}{\sigma^2 K^2 T}\right)^{\frac{r}{3r-2}}, \qquad w := T^{-1},$$

so that $1 - \beta = \min\{1, \max\{u, w\}\}$. Hence

$$\frac{1}{\sqrt{1-\beta}} \leq \frac{1}{\min\{1, \sqrt{u}\}} \leq 1 + \frac{1}{\sqrt{u}},$$

which implies

$$2D_{\mathcal{D}}\sqrt{\frac{\Delta_{\mathcal{D}} L}{T(1-\beta)}} + \frac{D_{\mathcal{D}}}{2}\sqrt{\frac{\Delta_{\mathcal{D}} L}{T}} \leq 3D_{\mathcal{D}}\sqrt{\frac{\Delta_{\mathcal{D}} L}{T}} + 2D_{\mathcal{D}} U.$$

Using $(1-\beta)^\alpha \leq u^\alpha + w^\alpha$ for $\alpha = \frac{r-1}{r}$,

$$2D_{\mathcal{D}}\sigma K(1-\beta)^{\frac{r-1}{r}} \leq 2D_{\mathcal{D}} U + 2D_{\mathcal{D}} W.$$

Combining the bounds gives

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[g_{\mathcal{D}}(x_t)\right] \leq \frac{\Delta_{\mathcal{D}}}{T} + 3D_{\mathcal{D}}\sqrt{\frac{\Delta_{\mathcal{D}} L}{T}} + 4D_{\mathcal{D}} U + 6D_{\mathcal{D}} W,$$

which is exactly the stated claim. □

Remark C.8. Compared with the unconstrained bound, the additional term $\Delta_{\mathcal{D}}/T$ comes from capping the stepsize at one to keep the iterates in $\mathcal{D}$. Moreover, the constrained suboptimality $\Delta_{\mathcal{D}}$ does not control the initial gradient norm $\|\nabla F(x_1)\|_*$. We therefore analyse the rescaled estimator $\widehat{m}_t$, which removes the initialisation-bias term without changing the iterates. Its initial noise-averaging error contributes the additional term $D_{\mathcal{D}} W = D_{\mathcal{D}}\sigma K / T^{(r-1)/r}$.

D Convergence Guarantees for MUON

In Theorem C.3 we provide guarantees for arbitrary p'-uniformly smooth dual norms $\|\cdot\|_*$, but do not achieve optimal rates whenever $p' < p$. In this section we will use an orthogonal approach, where we instead assume that $\|\cdot\|_* \leq \rho \|\cdot\|_p$ for some p-uniformly smooth norm $\|\cdot\|_p$. Such a constant always exists in the finite dimensional case, and will allow us to achieve optimal rates. We first derive a general result for NSSD-M in Section D.1, before specialising it to MUON in Section D.2.

D.1 Banach Space Setting

Therefore we first formulate Lemma C.2 for this approach.

Lemma D.1. *Suppose Assumptions 3 and 4 hold, and that $\|\cdot\|_p \leq \|\cdot\|_* \leq \rho \|\cdot\|_p$, where $(\mathcal{B}^*, \|\cdot\|_p)$ is p-uniformly smooth with smoothness constant $S_p(\mathcal{B}^*, \|\cdot\|_p) \leq S$. Then the iterates generated by* NSSD-M *with constant parameters $\eta_t \equiv \eta$ and $\beta_t \equiv \beta$ satisfy*

$$\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(x_t) - m_t\|_*\right] \leq \|\nabla F(x_1)\|_* \frac{\beta(1-\beta^T)}{1-\beta} + 2\rho\sigma S T (1-\beta)^{\frac{p-1}{p}} + L_0 \frac{\eta A T}{1-\beta} + L_1 \frac{\eta A}{1-\beta}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(x_t)\|_*\right],$$

where $A \equiv A(L_1\eta) = \frac{e^{L_1\eta}-1}{L_1\eta}$.

We note that such a ρ always exists in finite dimensions, but may not exist in general Banach spaces.

Proof. We proceed exactly as in the proof of Lemma C.2, but replace

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^{t} b_{\tau,t}(\nabla f(x_\tau,\xi_\tau)-\nabla F(x_\tau))\right\|_*\right] \le \mathbb{E}\left[\left\|\sum_{\tau=1}^{t} b_{\tau,t}(\nabla f(x_\tau,\xi_\tau)-\nabla F(x_\tau))\right\|_*^r\right]^{1/r}$$

with

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{\tau=1}^{t} b_{\tau,t}(\nabla f(x_\tau,\xi_\tau)-\nabla F(x_\tau))\right\|_*\right] &\le \rho\mathbb{E}\left[\left\|\sum_{\tau=1}^{t} b_{\tau,t}(\nabla f(x_\tau,\xi_\tau)-\nabla F(x_\tau))\right\|_p\right] \\ &\le \rho\mathbb{E}\left[\left\|\sum_{\tau=1}^{t} b_{\tau,t}(\nabla f(x_\tau,\xi_\tau)-\nabla F(x_\tau))\right\|_p^p\right]^{1/p}.\end{aligned}$$

The proof closes the same way as before by using the p-uniform smoothness of $\|\cdot\|_p$. □

Now we are ready to prove the norm-constant version of the convergence result for NSSD-M under heavy-tailed noise and (L_0, L_1)-smoothness.

Theorem D.2. *Suppose Assumptions 1, 3 and 4 hold. Further assume that $\|\cdot\|_p \le \|\cdot\|_* \le \rho\|\cdot\|_p$, where $(\mathcal{B}^*, \|\cdot\|_p)$ is p-uniformly smooth with smoothness constant $S_p(\mathcal{B}^*, \|\cdot\|_p) \le S$. Then the iterates generated by NSSD-M with*

$$\eta_t \equiv \eta = \min\left\{\frac{1-\beta}{5L_1}, \sqrt{\frac{\Delta_1(1-\beta)}{L_0 T}}\right\},$$
$$\beta_t \equiv \beta = 1 - \min\left\{1, \max\left\{\left(\frac{\Delta_1 L_0}{\sigma^2\rho^2 S^2 T}\right)^{\frac{p}{3p-2}}, \left(\frac{5L_1\Delta_1}{2\sigma\rho S T}\right)^{\frac{p}{2p-1}}\right\}\right\}$$

satisfy

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|_*\right] \le 18\sqrt{\frac{\Delta_1 L_0}{T}} + 30\frac{\Delta_1 L_1}{T} + 28\left(\frac{\Delta_1 L_0(\rho\sigma S)^{\frac{p}{p-1}}}{T}\right)^{\frac{p-1}{3p-2}} \\ + 30\left(\frac{\Delta_1 L_1(\rho\sigma S)^{\frac{p}{p-1}}}{T}\right)^{\frac{p-1}{2p-1}}.\end{aligned}$$

In particular, the sample complexity to reach an ε-$\|\cdot\|_*$-stationary point is at most

$$\mathcal{O}\left(\frac{\Delta_1 L_1}{\varepsilon} + \frac{\Delta_1 L_0}{\varepsilon^2} + \frac{\Delta_1 L_1}{\varepsilon}\left(\frac{\rho\sigma S}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\Delta_1 L_0}{\varepsilon^2}\left(\frac{\rho\sigma S}{\varepsilon}\right)^{\frac{p}{p-1}}\right).$$

We note that Assumption 3 could be replaced with the S_p version $\mathbb{E}\left[\|\nabla f(x,\xi)-\nabla F(x)\|_p^p\right] \le \sigma^p$ for all $x \in \mathcal{B}$, which would allow for smaller σ values. However, we choose to state the result with the stronger Assumption 3 for better comparability with the previous results for NSSD-M.

Proof. For notational conciseness denote $\kappa := \rho\sigma S$. We proceed as in the proof of Theorem C.3, but replacing Lemma C.2 by Lemma D.1, to get

$$\begin{aligned}\eta\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|_*\right] \le \Delta_1 + L_0\eta^2 T\left(B + \frac{2A}{1-\beta}\right) + 2\eta\|\nabla F(x_1)\|_* \min\left\{T, \frac{1}{1-\beta}\right\} \\ + 4\kappa\eta T(1-\beta)^{\frac{p-1}{p}} \\ + L_1\eta^2\left(B + \frac{2A}{1-\beta}\right)\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|_*\right].\end{aligned}$$

By Equation (8) $L_1\eta^2\left(B + \frac{2A}{1-\beta}\right) \leq \frac{5}{9}\eta$ and re-arranging thus yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla F(x_t)\right\|_*\right] \leq \frac{9\Delta_1}{4\eta T} + \frac{6L_0\eta A}{(1-\beta)} + \frac{9}{2}\left\|\nabla F(x_1)\right\|_* \min\left\{1, \frac{1}{T(1-\beta)}\right\} + 10\kappa(1-\beta)^{\frac{p-1}{p}}.$$

Using $\eta = \min\left\{\frac{1-\beta}{5L_1}, \sqrt{\frac{\Delta_1(1-\beta)}{L_0 T}}\right\}$ and absorbing the initialisation error using (7) as above, we further obtain

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla F(x_t)\right\|_*\right] \leq 18\sqrt{\frac{\Delta_1 L_0}{T(1-\beta)}} + \frac{30\Delta_1 L_1}{T(1-\beta)} + 10\kappa(1-\beta)^{\frac{p-1}{p}}.$$

Now denote

$$U \coloneqq \left(\frac{\Delta_1 L_0 \kappa^{\frac{p}{p-1}}}{T}\right)^{\frac{p-1}{3p-2}}, \qquad V \coloneqq \left(\frac{\Delta_1 L_1 \kappa^{\frac{p}{p-1}}}{T}\right)^{\frac{p-1}{2p-1}}$$

With the same arguments as in the proof of Theorem C.3 we derive the upper bounds

$$\begin{aligned}
18\sqrt{\frac{\Delta_1 L_0}{T(1-\beta)}} &\leq 18\sqrt{\frac{\Delta_1 L_0}{T}} + 18U \\
\frac{30\Delta_1 L_1}{T(1-\beta)} &\leq 30\frac{\Delta_1 L_1}{T} + 30\left(\frac{2}{5}\right)^{\frac{p}{2p-1}} V \\
10\kappa(1-\beta)^{\frac{p-1}{p}} &\leq 10U + 10\left(\frac{5}{2}\right)^{\frac{p-1}{2p-1}} V
\end{aligned}$$

on the remaining terms. Combining these bounds yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla F(x_t)\right\|_*\right] \leq 18\sqrt{\frac{\Delta_1 L_0}{T}} + 30\frac{\Delta_1 L_1}{T} + 28U + 30V,$$

which finishes the proof. □

D.2 Application to MUON

Next we extend the previous result to MUON, which is an instance of NSSD-M with $\|\cdot\| = \|\cdot\|_{\mathrm{S}_\infty}$, i.e.,

$$\begin{aligned}
m_t &\leftarrow \beta_t m_{t-1} + (1-\beta_t)\nabla f(x_t, \xi_t) \\
x_{t+1} &\leftarrow x_t + \eta_t \operatorname*{argmin}_{\|d\|_{\mathrm{S}_\infty} \leq 1} \langle m_t, d\rangle,
\end{aligned} \tag{MUON}$$

where one choice of minimiser is $-UV^\top$ for a singular value decomposition $m_t = U\Sigma V^\top$. In order to analyse MUON, we need to upper-bound the nuclear norm $\|\cdot\|_{\mathrm{S}_1} \leq \rho\|\cdot\|_p$ with a p-uniformly smooth norm $\|\cdot\|_p$, to use the machinery from Section D.1. In particular, we want to choose a p-uniformly smooth norm that has the smallest product ρS of norm equivalence and smoothness constants. We motivate our choice of p-uniformly smooth norm with the following lemma, which shows that the Schatten-p norm is the best possible choice in the above sense.

Lemma D.3. *Let $\|\cdot\|$ be a p-uniformly smooth norm on $\mathbb{R}^{m\times n}$ with smoothness constant $S_p(\|\cdot\|) \leq S$. Furthermore assume the normalisation $\left\|uv^\top\right\| = 1$ for all $u \in \mathbb{R}^m, v \in \mathbb{R}^n$ with $\|u\|_2 = \|v\|_2 = 1$. Then we have*

$$\sup_{A\in\mathbb{R}^{m\times n}\setminus\{0\}} \frac{\|A\|_{\mathrm{S}_1}}{\|A\|} \geq \max\left\{1, \frac{\min\{m,n\}^{\frac{p-1}{p}}}{S}\right\}.$$

In particular, the Schatten-p norm achieves this lower-bound exactly with $S = 1$.

Proof. Denote $d \coloneqq \min\{m,n\}$ and define

$$C \coloneqq \sup_{A \in \mathbb{R}^{m\times n}\setminus\{0\}} \frac{\|A\|_{\mathrm{S}_1}}{\|A\|}.$$

Furthermore let $E_i \coloneqq e_i f_i^\top \in \mathbb{R}^{m\times n}$, where $(e_i)_{i=1}^m, (f_i)_{i=1}^n$ are the standard bases of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively, for $i = 1,\dots,d$, and $\varepsilon_1,\dots,\varepsilon_d$ be independent Rademacher random variables. Define

$$M_k \coloneqq \sum_{i=1}^k \varepsilon_i E_i.$$

Then, by definition of the smoothness constant $S_p(\mathbb{R}^{m\times n})$ (see Definition 2.1), we have

$$\mathbb{E}\left[\|M_k\|^p \,|\, \varepsilon_1,\dots,\varepsilon_{k-1}\right] = \frac{\|M_{k-1}+E_k\|^p + \|M_{k-1}-E_k\|^p}{2} \le \|M_{k-1}\|^p + S^p \|E_k\|^p.$$

Hence, by induction and $\|E_k\| = 1$, we get $\mathbb{E}\left[\|M_d\|^p\right] \le dS^p$. On the other hand note that

$$C \ge \frac{\|M_d\|_{\mathrm{S}_1}}{\|M_d\|} = \frac{d}{\|M_d\|}$$

and hence

$$\frac{d^p}{C^p} \le \mathbb{E}\left[\|M_d\|^p\right] \le dS^p.$$

Rearranging yields the claim. □

Finally we can state the (L_0, L_1)-smooth version of Theorem 3.5.

Corollary D.4. *Suppose Assumptions 1, 3 and 4 hold and let* $\kappa \coloneqq \min\{m,n\}^{\frac{(p-1)}{p}}$. *Then the iterates generated by* MUON *with*

$$\eta_t \equiv \eta = \min\left\{\frac{1-\beta}{5L_1}, \sqrt{\frac{\Delta_1(1-\beta)}{L_0 T}}\right\},$$
$$\beta_t \equiv \beta = 1 - \min\left\{1, \max\left\{\left(\frac{\Delta_1 L_0}{\kappa^2\sigma^2 T}\right)^{\frac{p}{3p-2}}, \left(\frac{5L_1\Delta_1}{2\kappa\sigma T}\right)^{\frac{p}{2p-1}}\right\}\right\}$$

satisfy

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\|\nabla F(x_t)\|_{\mathrm{S}_1}\right] \le 18\sqrt{\frac{\Delta_1 L_0}{T}} + 30\frac{\Delta_1 L_1}{T} + 28\left(\frac{\min\{m,n\}\Delta_1 L_0 \sigma^{\frac{p}{p-1}}}{T}\right)^{\frac{p-1}{3p-2}} + 30\left(\frac{\min\{m,n\}\Delta_1 L_1 \sigma^{\frac{p}{p-1}}}{T}\right)^{\frac{p-1}{2p-1}}.$$

In particular, the sample complexity to reach an ε-$\|\cdot\|_{\mathrm{S}_1}$*-stationary point is at most*

$$\mathcal{O}\left(\frac{\Delta_1 L_1}{\varepsilon} + \frac{\Delta_1 L_0}{\varepsilon^2} + \iota\frac{\Delta_1 L_1}{\varepsilon}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}} + \iota\frac{\Delta_1 L_0}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right),$$

where $\iota \coloneqq \min\{m,n\}$.

Proof. By Lemma A.2 we have $\|A\|_{\mathrm{S}_1} \le \kappa \|A\|_{S_p}$ for all $A \in \mathbb{R}^{m\times n}$, and the Schatten-$p$ norm is p-uniformly smooth with smoothness constant $S_p(\mathbb{R}^{m\times n}) \le 1$ (Example 2.3). Hence we may apply Theorem D.2 with $(\mathcal{B}, \|\cdot\|) = (\mathbb{R}^{m\times n}, \|\cdot\|_{\mathrm{S}_\infty})$, $\|\cdot\|_p = \|\cdot\|_{S_p}$, $\rho \leftarrow \kappa$ and $S = 1$. □

Taking $L_0 = L$ and $L_1 = 0$ in Corollary D.4 gives Theorem 3.5.

Remark D.5. Extending Remark 3.7, we note that Theorem D.2 also extends to full networks including vector valued weights by considering (Pethick et al., 2025b) the Banach space

$$\mathcal{B} = \mathbb{R}^{m_1 \times n_1} \times \ldots \times \mathbb{R}^{m_D \times n_D} \times \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_K},$$
$$\|\cdot\| = \max\left\{ \max_{l \in [D]} \|W_l\|_{\mathrm{S}_\infty}, \max_{l \in [K]} \|v_l\|_{\ell_\infty} \right\}.$$

In this case, vector-valued weights are trained with SIGNUM, which is widely considered to be an idealised version of ADAM (Kunstner et al., 2023). The corresponding dimension dependence is given by

$$\sum_{l \in [D]} \min\{m_l, n_l\} + \sum_{l \in [K]} d_l.$$

Remark D.6 (Newton–Schulz Approximation)**.** The approximation argument of G. Y. Kim and Oh (2026) extends to Theorem 3.5. Let P_t be the partial polar factor of m_t, with $P_t = 0$ if $m_t = 0$, and use the update $x_{t+1} = x_t - \eta_t O_t$, where $\|O_t\|_{\mathrm{S}_\infty} \leq 1$ and $\|O_t - P_t\|_{\mathrm{S}_\infty} \leq \delta$ for a fixed $\delta \in (0, 1)$ at every iteration. Then

$$\langle \nabla F(x_t), -O_t \rangle \leq -(1-\delta) \|\nabla F(x_t)\|_{\mathrm{S}_1} + 2 \|\nabla F(x_t) - m_t\|_{\mathrm{S}_1}.$$

Since $\|x_{t+1} - x_t\|_{\mathrm{S}_\infty} \leq \eta_t$, the momentum bound in Lemma D.1 remains unchanged. Hence, with the same parameters, the bound in Theorem 3.5 is multiplied by at most $(1-\delta)^{-1}$. If δ is independent of the dimension and ε, the sample complexity retains its exponents. The Newton–Schulz iteration analysed by G. Y. Kim and Oh (2026) achieves this accuracy while staying in the spectral unit ball, but the required number of inner iterations depends on the singular values of the scaled momentum. Our assumptions therefore do not guarantee these exponents for a fixed number of inner iterations.

E Euclidean Lower Bound

In this section we prove Theorem 3.6. Therefore we consider the Euclidean version of NSSD-M, i.e.,

$$\begin{aligned} m_t &= \beta m_{t-1} + (1-\beta)\nabla f(x_t, \xi_t), \\ x_{t+1} &= x_t - \eta m_t / \|m_t\|_F, \end{aligned} \tag{10}$$

with $\beta \in [0, 1)$ and $0 < \eta \leq \sqrt{\Delta_1(1-\beta)/(LT)}$. Without loss of generality, we fix $x_1 = 0$.

Theorem E.1 (Lower Bound for Euclidean NSSD-M)**.** *Let $m, n \geq 2$, $p \in (1, 2]$ and $\Delta_1, L, \sigma > 0$. Suppose that $0 < \varepsilon \leq c \min\{\sigma, \sqrt{\Delta_1 L}\}$. For every integer $T \geq 1$ and constant parameters η, β as above, there exist an objective F satisfying Assumptions 1 and 2 in the spectral norm and a gradient oracle satisfying Assumption 3 in the nuclear norm such that the iterates of (10) satisfy*

$$\min_{1 \leq t \leq T} \mathbb{E}\left[\|\nabla F(x_t)\|_{S_1}\right] \geq 2\varepsilon \quad \textit{whenever} \quad T \leq c_p \min\{m, n\}^{\frac{3p-2}{2(p-1)}} \frac{\Delta_1 L}{\varepsilon^2} \left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}.$$

Here $c > 0$ and $c_p > 0$ are sufficiently small constants, with c universal and c_p depending only on p. Moreover, the noise has rank at most one, so its Frobenius and nuclear p-th moments coincide.

To prove Theorem E.1, we construct an objective and gradient oracle and distinguish two cases depending on β. In the first case, the momentum does not sufficiently reduce the noise, so the expected gradient norm remains large for any stepsize. In the second case, we bound the distance travelled by the iterates and show that they remain in a region of constant nonzero gradient up to time T. The table below summarises these cases under the stated bound on T, with $k = \min\{m, n\}$, $\lambda = (\varepsilon/(\sqrt{k}\sigma))^{p/(p-1)}$ and A_p given in (13).

Momentum β	Stepsize	Gradient lower bound
$\beta \leq 1 - A_p\lambda$	Any $\eta > 0$	$\mathbb{E}\left[\|\nabla F(x_t)\|_{S_1}\right] \geq 2\varepsilon$ for all $t \geq 1$ (Lemma E.2).
$\beta > 1 - A_p\lambda$	$\eta \leq \sqrt{\Delta_1(1-\beta)/(LT)}$	$\|\nabla F(x_t)\|_{S_1} = 4\varepsilon$ for all $t \leq T$ with probability one (Lemma E.3).

The construction has a region of constant nonzero gradient. We choose its gradient G to have nuclear norm 4ε and Frobenius norm of order $\varepsilon/\sqrt{k}$. To this end, we set

$$G = a\,\mathrm{Diag}\left(1, \frac{1}{\sqrt{k-1}}, \ldots, \frac{1}{\sqrt{k-1}}\right), \qquad a = \frac{4\varepsilon}{1+\sqrt{k-1}}. \tag{11}$$

This gives $\|G\|_{S_1} = 4\varepsilon$ and $\|G\|_F = \sqrt{2}a \le 4\sqrt{2}\varepsilon/\sqrt{k}$.

Lemma E.2. *Suppose $\varepsilon \le \min\{\sigma/16, \sqrt{\Delta_1 L}/4\}$. Then there exist an objective F and a gradient oracle satisfying the assumptions of Theorem 3.5, with initial gap Δ_1 and rank-one noise, such that for every $\eta > 0$ and $\beta \in [0,1)$ with $1-\beta \ge 1024\delta$, the iterates of (10) satisfy*

$$\mathbb{E}\left[\|\nabla F(x_t)\|_{S_1}\right] \ge 2\varepsilon \qquad \textit{for all } t \ge 1,$$

where $\delta = (2^{2p+1}a^p/\sigma^p)^{1/(p-1)}$.

Proof. We adapt the stochastic lower-bound argument of Hübler et al. (2025, Appendix E.2) from minibatching to momentum and the nuclear norm. Write $\alpha = 1-\beta$. Let ψ be the one-sided Huber function, with a flat, quadratic and linear part, defined by $\psi(0) = 0$ and $\psi'(z) = \min\{4\varepsilon, \max\{0, Lz\}\}$. We choose C so that the linear part of F has gradient G, setting

$$C = \frac{G}{4\varepsilon}, \qquad R = \frac{\Delta_1 - 8\varepsilon^2/L}{4\varepsilon}, \qquad F(x) = \psi(\langle C, x\rangle_F + R + 4\varepsilon/L).$$

Since $\|C\|_{S_1} = 1$, F is L-smooth in the spectral norm. Moreover, $\inf F = 0$ and $F(0) = \Delta_1$. We can write its gradient as $\nabla F(x) = \vartheta(x)G$, where $\vartheta(x) \in [0,1]$ and $\vartheta(x) = 1$ whenever $\langle C, x\rangle_F \ge -R$. For the stochastic oracle, we take $f(x, I) = F(x) + \langle Z(I), x\rangle_F$, where $Z(I) = 4a(I/\delta - 1)E_{11}$, with E_{11} the first matrix unit and $I \sim \mathrm{Bernoulli}(\delta)$. Our restriction on ε gives $a \le \sigma/8$ and hence $\delta \le 1/2$. The noise has rank one and satisfies

$$\mathbb{E}[Z] = 0, \qquad \mathbb{E}\left[\|Z\|_{S_1}^p\right] = \mathbb{E}\left[\|Z\|_F^p\right] \le 2^{2p+1}a^p\delta^{1-p} = \sigma^p.$$

Writing $s_t := 1-\beta^t$ and unrolling the momentum recursion gives $m_t = s_t\left[u_t G + 4a(z_t/\delta - 1)E_{11}\right]$, where

$$z_t := \frac{\alpha}{s_t}\sum_{j=1}^{t}\beta^{t-j}I_j, \qquad u_t := \frac{\alpha}{s_t}\sum_{j=1}^{t}\beta^{t-j}\vartheta(x_j) \in [0,1].$$

To lower-bound the expected gradient norm, it suffices to show that $\langle C, x_t\rangle_F \ge -R$ with probability at least $1/2$ for each t, since the gradient then has nuclear norm 4ε. As $C = G/(4\varepsilon)$, each update satisfies

$$\langle C, x_{t+1} - x_t\rangle_F = -\frac{\eta}{4\varepsilon}\frac{\langle G, m_t\rangle_F}{\|m_t\|_F}.$$

We therefore seek an upper bound on $\langle G, m_t\rangle_F/\|m_t\|_F$. Since $u_t \le 1$, we can replace it by 1 in this upper bound. Indeed, with the noise term rE_{11} fixed,

$$\frac{d}{du}\frac{\langle G, uG + rE_{11}\rangle_F}{\|uG + rE_{11}\|_F} = \frac{r^2a^2}{\|uG + rE_{11}\|_F^3} \ge 0$$

for $u \ge 0$. Using $\|G\|_F = \sqrt{2}a$, we obtain $\langle G, m_t\rangle_F/\|m_t\|_F \le \|G\|_F h(z_t)$, where

$$h(z) := \frac{4z/\delta - 2}{\sqrt{2\big((4z/\delta - 3)^2 + 1\big)}}.$$

Summing the updates gives

$$\langle C, x_{n+1}\rangle_F \ge -\frac{\eta\|G\|_F}{4\varepsilon}\sum_{t=1}^{n}h(z_t). \tag{12}$$

For $0 \le z \le \delta/4$, we have $h(z) \le -1/(2\sqrt{5})$, while $h(z) \le 1$ for all $z \ge 0$. Thus, writing $c := 1/(2\sqrt{5})$, we get

$$h(z_t) \le -c + (1+c)b_t, \qquad b_t := \min\{4z_t/\delta, 1\}.$$

We can therefore bound the probability of a positive sum in (12) by Markov's inequality. For each $n \geq 1$,

$$\Pr\left\{\sum_{t=1}^{n} h(z_t) > 0\right\} \leq \Pr\left\{\sum_{t=1}^{n} b_t > \frac{cn}{1+c}\right\} \leq \frac{1+c}{cn} \sum_{t=1}^{n} \mathbb{E}\left[b_t\right].$$

It remains to bound $\mathbb{E}\left[b_t\right]$. If $t < 1/\alpha$, then $b_t = 0$ whenever $I_1 = \cdots = I_t = 0$, so

$$\mathbb{E}\left[b_t\right] \leq \Pr\{I_j = 1 \text{ for some } j \leq t\} \leq t\delta \leq \frac{\delta}{\alpha}.$$

If $t \geq 1/\alpha$, then $s_t = 1 - \beta^t \geq 1 - e^{-1} > 1/2$. The definition of z_t and subadditivity of $u \mapsto \min\{u, 1\}$ give

$$\mathbb{E}\left[b_t\right] \leq \delta \sum_{j \geq 0} \min\left\{\frac{8\alpha\beta^j}{\delta}, 1\right\} \leq \frac{\delta}{\alpha}\left(\log\frac{8\alpha}{\delta} + 2\right),$$

where the last step follows by splitting the sum at $\lceil \alpha^{-1} \log(8\alpha/\delta) \rceil$ and using $\beta^j \leq e^{-\alpha j}$. Combining both cases and using $\alpha/\delta \geq 1024$ gives

$$\frac{1+c}{c}\mathbb{E}\left[b_t\right] \leq (1 + 2\sqrt{5})\frac{\delta}{\alpha}\left(\log\frac{8\alpha}{\delta} + 2\right) \leq \frac{1}{16}.$$

Hence, for each $n \geq 1$, the sum in (12) is nonpositive with probability at least $15/16$. On this event, $\langle C, x_{n+1}\rangle_F \geq 0$ and therefore $\nabla F(x_{n+1}) = G$. Together with $\nabla F(x_1) = G$, this yields $\mathbb{E}\left[\|\nabla F(x_t)\|_{S_1}\right] \geq 15\varepsilon/4 > 2\varepsilon$ for every $t \geq 1$. □

Since $a \leq 4\varepsilon/\sqrt{k}$, the condition $1 - \beta \geq 1024\delta$ in Lemma E.2 follows from

$$1 - \beta \geq A_p\lambda, \qquad A_p = 1024\, 2^{(p+1)/(p-1)} 8^{p/(p-1)}. \tag{13}$$

Lemma E.3. *Consider the objective and gradient oracle constructed in the proof of Lemma E.2, and suppose $\varepsilon \leq \min\{\sigma/16, \sqrt{\Delta_1 L}/4\}$. Let $T \geq 1$ and $\beta \in [0, 1)$ with $\beta > 1 - A_p\lambda$. If*

$$0 < \eta \leq \sqrt{\frac{\Delta_1(1-\beta)}{LT}} \qquad \text{and} \qquad T \leq \frac{k\Delta_1 L}{256 A_p \varepsilon^2 \lambda},$$

then, for every realisation of the noise, the iterates of (10) satisfy

$$\|\nabla F(x_t)\|_{S_1} = 4\varepsilon \qquad \text{for all } 1 \leq t \leq T.$$

Proof. By construction, it suffices to show that $\langle C, x_t\rangle \geq -R$ for all $t \leq T$, since the gradient is G there. As $\|C\|_F = \|G\|_F/(4\varepsilon) \leq \sqrt{2/k}$, each normalised update satisfies

$$\left|\langle C, x_{t+1} - x_t\rangle\right| = \eta \frac{\left|\langle C, m_t\rangle\right|}{\|m_t\|_F} \leq \eta \|C\|_F.$$

Summing from $x_1 = 0$ and using the bounds on η, β and T gives, for $t \leq T$,

$$\begin{aligned}\left|\langle C, x_t\rangle\right| &\leq (t-1)\eta \|C\|_F \leq T\eta\sqrt{\frac{2}{k}} \\ &\leq \sqrt{\frac{2\Delta_1(1-\beta)T}{Lk}} \leq \frac{\Delta_1}{8\varepsilon} \leq R,\end{aligned}$$

where the last inequality uses $R = (\Delta_1 - 8\varepsilon^2/L)/(4\varepsilon) \geq \Delta_1/(8\varepsilon)$. Thus every iterate stays in the linear region, where $\|\nabla F(x_t)\|_{S_1} = \|G\|_{S_1} = 4\varepsilon$. □

Proof of Theorem E.1. Take $c = 1/16$ and $c_p = 1/(256A_p)$. For $\beta \leq 1 - A_p\lambda$, the claim follows from Lemma E.2 and (13). The remaining case follows from Lemma E.3. □

F First-Order Lower Bounds

In this section we prove Theorems 3.8 and 3.9. We first recall the sample complexity framework of Arjevani et al. (2023). Let $\mathcal{A}_{\text{rand}}$ denote the class of possibly randomised first-order algorithms starting at zero and using a stochastic gradient oracle to produce iterates $x_1, x_2, \dots$. For $m, n \in \mathbb{N}_{\geq 2}$, $\Delta_1, L, \sigma \geq 0$ and $p \in (1, 2]$, define the function and oracle classes

$$\mathcal{F}_{\Delta_1,L}(m,n) \coloneqq \left\{ F \colon \mathbb{R}^{m\times n} \to \mathbb{R} \;\middle|\; \begin{array}{c} F(0) - \inf_x F(x) \leq \Delta_1, \\ \|\nabla F(x) - \nabla F(y)\|_{\mathrm{S}_1} \leq L \|x-y\|_{\mathrm{S}_\infty} \end{array} \right\},$$

$$\mathcal{O}_{p,\sigma}(m,n) \coloneqq \left\{ \nabla f \colon \mathbb{R}^{m\times n} \times \Xi \to \mathbb{R}^{m\times n} \;\middle|\; \begin{array}{c} \mathbb{E}\left[\nabla f(x,\xi)\right] = \nabla F(x), \\ \mathbb{E}\left[\|\nabla f(x,\xi) - \nabla F(x)\|_{\mathrm{S}_1}^p\right] \leq \sigma^p \end{array} \right\}.$$

The distributional complexity is

$$\mathfrak{m}_\varepsilon^{\text{rand}}\left(\|\cdot\|_{\mathrm{S}_1}, \Delta_1, L, p, \sigma\right) \coloneqq \sup_O \sup_{P_F} \inf_{\mathcal{A}\in\mathcal{A}_{\text{rand}}} \inf\left\{T \in \mathbb{N} \;\middle|\; \mathbb{E}\left[\|\nabla F(x_T)\|_{\mathrm{S}_1}\right] \leq \varepsilon\right\},$$

where O assigns each objective $F \in \mathcal{F}_{\Delta_1,L}(m,n)$ an admissible oracle $O_F \in \mathcal{O}_{p,\sigma}(m,n)$, and P_F ranges over probability distributions on $\mathcal{F}_{\Delta_1,L}(m,n)$. The expectation is over the random choice of F, the algorithm's random seed and the oracle samples. Lower-bounds on this distributional complexity imply lower-bounds on the worst-case complexity (Nemirovskij and Yudin, 1983).

F.1 Convex Objectives

We now prove Theorem 3.8. Let $m, n \in \mathbb{N}_{\geq 2}$ and set

$$\kappa \coloneqq \min\{m, n\} \geq 2.$$

Further let $(e_i)_{i=1}^m$ and $(f_j)_{j=1}^n$ be the standard bases of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively, and

$$E_i \coloneqq e_i f_i^\top \in \mathbb{R}^{m\times n}.$$

For a sign vector $s \in \{\pm 1\}^\kappa$, define

$$S_s \coloneqq \sum_{i=1}^{\kappa} s_i E_i.$$

Fix $\delta > 0$ to be specified later and $q \coloneqq \left(\frac{2\delta}{\sigma}\right)^{\frac{p}{p-1}}$. Then the hard instance is given by

$$F_s(x) \coloneqq \frac{L}{2\kappa}\|x\|_{\mathrm{F}}^2 - \frac{\delta}{\kappa}\langle S_s, x\rangle_F, \qquad \nabla f_s(x,\xi) \coloneqq \frac{L}{\kappa}x - \frac{\delta}{q}B s_I E_I,$$

where $B \sim \mathrm{Bern}(q)$ and $I \sim \mathrm{Unif}([\kappa])$ independently.

Now we first verify that F_s and ∇f_s satisfy the required properties.

Lemma F.1. *Assume $\delta \leq \sigma/2$. Then the function F_s is L-smooth with $F_s(0) - \inf_x F_s(x) \leq \frac{\delta^2}{2L}$ and the gradient oracle ∇f_s is unbiased and satisfies $\mathbb{E}\left[\|\nabla f_s(x,\xi) - \nabla F_s(x)\|_{\mathrm{S}_1}^p\right] \leq \sigma^p$.*

Proof. The smoothness is immediate by noting that $\|\nabla F_s(x) - \nabla F_s(y)\|_{\mathrm{S}_1} = \frac{L}{\kappa}\|x-y\|_{\mathrm{S}_1} \leq L\|x-y\|_{\mathrm{S}_\infty}$ for all $x, y \in \mathbb{R}^{m\times n}$. To prove the bounded initialisation gap, we complete the square to get

$$F_s(x) = \frac{L}{2\kappa}\left\|x - \frac{\delta}{L}S_s\right\|_{\mathrm{F}}^2 - \frac{\delta^2}{2L\kappa}\|S_s\|_{\mathrm{F}}^2,$$

and hence $F_s(x) \geq -\frac{\delta^2}{2L}$. Unbiasedness again directly follows from the independence of B and I and the definition of q. Finally, we have

$$\mathbb{E}\left[\|\nabla f_s(x,\xi) - \nabla F_s(x)\|_{\mathrm{S}_1}^p\right] = \mathbb{E}\left[\left\|\frac{\delta}{\kappa}S_s - \frac{\delta}{q}B s_I E_I\right\|_{\mathrm{S}_1}^p\right] \leq 2^{p-1}\left(\delta^p + \mathbb{E}\left[\left\|\frac{\delta}{q}B s_I E_I\right\|_{\mathrm{S}_1}^p\right]\right)$$

and

$$\mathbb{E}\left[\left\|\frac{\delta}{q}B s_I E_I\right\|_{\mathrm{S}_1}^p\right] = q\left(\frac{\delta}{q}\right)^p \mathbb{E}\left[\|s_I E_I\|_{\mathrm{S}_1}^p\right] = q\left(\frac{\delta}{q}\right)^p = \frac{\delta^p}{q^{p-1}} = \frac{\sigma^p}{2^p}.$$

Hence the claim follows from our assumption $\delta \leq \frac{\sigma}{2}$. □

Additionally we need the following technical lemmata.

Lemma F.2. *For every $A \in \mathbb{R}^{m\times n}$, $\|A\|_{\mathrm{S}_1} \geq \sum_{i=1}^{\kappa} |A_{ii}|$.*

Proof. Let $\omega_i \coloneqq \operatorname{sgn}(A_{ii})$ for $i = 1, \dots, \kappa$ and $U \coloneqq \sum_{i=1}^{\kappa} \omega_i E_i$. Then $\|U\|_{\mathrm{S}_\infty} \leq 1$ and, by duality of the nuclear and operator norm, $\sum_{i=1}^{\kappa} |A_{ii}| = \langle U, A\rangle_F \leq \|U\|_{\mathrm{S}_\infty} \|A\|_{\mathrm{S}_1} \leq \|A\|_{\mathrm{S}_1}$. □

For a sign vector $s \in \{\pm 1\}^\kappa$ we denote the version which flips the sign at index i as $s^{(i)} = (s_1, \dots, s_{i-1}, -s_i, s_{i+1}, \dots, s_\kappa)$. The following lemma says that, as long as $(B_t, I_t) = (1, i)$ did not yet happen, the iterates of an algorithm run on F_s and $F_{s^{(i)}}$ are identical.

Lemma F.3. *Let $\mathcal{A}$ be an optimisation algorithm, $i \in [\kappa]$ and $s \in \{\pm 1\}^\kappa$. Let $B_1, \dots, B_T$ and $I_1, \dots, I_T$ be independent and run $\mathcal{A}$ on F_s and $F_{s^{(i)}}$ with the same random seed and oracle samples to obtain iterates $x_1, g_1, \dots, x_T, g_T$ and $x'_1, g'_1, \dots, x'_T, g'_T$ respectively. Furthermore let*

$$A_i \coloneqq \bigcap_{t=1}^{T} \{(I_t, B_t) \neq (i, 1)\}.$$

Then we have $\mathbb{1}_{A_i}(x_t, g_t) = \mathbb{1}_{A_i}(x'_t, g'_t)$ for all $t \in [T]$.

Proof. This follows by induction. For $t = 1$ the claim is trivial as $x_1 = x'_1$. Now assume the claim holds for some $t \in [T-1]$. Then, on the event A_i, we have $\nabla f_s(x_t, \xi_t) = \nabla f_{s^{(i)}}(x'_t, \xi_t)$ and hence $x_{t+1} = x'_{t+1}$. □

Finally we need a technical lemma which roughly states that, for $s \sim \mathrm{Unif}(\{\pm 1\}^\kappa)$, as long as the algorithm has not seen an index i, s_i is still uniformly distributed conditioned on the history.

Lemma F.4. *Let $\mathcal{A}$ be an optimisation algorithm with internal randomness r. Let $s \sim \mathrm{Unif}(\{\pm 1\}^\kappa)$ and denote the sequence generated by $\mathcal{A}$ on F_s as $(x_1, g_1), \dots, (x_T, g_T)$. Furthermore define the observed randomness $\mathcal{H}_T \coloneqq \sigma(\{r, g_1, \dots, g_T\})$ and*

$$A_i \coloneqq \bigcap_{t=1}^{T} \{(I_t, B_t) \neq (i, 1)\}.$$

Finally assume that $r, s, B_1, I_1, \dots, B_T, I_T$ are mutually independent. Then

$$\mathbb{1}_{A_i} \mathbb{E}[s_i \mid \mathcal{H}_T] = 0 \qquad \text{a.s.}$$

In particular, for any $\mathcal{H}_T$-measurable ν and Borel-measurable function $\psi\colon \mathbb{R} \times \{\pm 1\} \to \mathbb{R}$, we have

$$\mathbb{1}_{A_i} \mathbb{E}[\psi(\nu, s_i) \mid \mathcal{H}_T] = \mathbb{1}_{A_i} \frac{\psi(\nu, 1) + \psi(\nu, -1)}{2}.$$

Proof. Fix $i \in [\kappa]$, and define the sigma-algebra containing all primitive randomness except s_i as

$$\mathcal{K}_i \coloneqq \sigma(r, I_1, B_1, \dots, I_T, B_T, s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_\kappa).$$

Furthermore let $M \coloneqq (r, g_1, \dots, g_T)$ be the trajectory on F_s, and $M^{(\pm 1)}$ the counterfactual trajectories on $F_{s^{(\pm 1)}}$, where $s^{(\pm 1)} = (s_1, \dots, s_{i-1}, \pm 1, s_{i+1}, \dots, s_\kappa)$.

Note that $A_i \in \mathcal{H}_T$, since $g_t - (L/\kappa) x_t$ reveals whether $(B_t, I_t) = (1, i)$. Now let Z be any bounded $\mathcal{H}_T$-measurable random variable. Since $\mathcal{H}_T = \sigma(M)$, the Doob-Dynkin Lemma supplies a bounded Borel-measurable function h such that $Z = h(M)$. Hence

$$\mathbb{E}[\mathbb{1}_{A_i} Z s_i] = \mathbb{E}[\mathbb{E}[\mathbb{1}_{A_i} Z s_i \mid \mathcal{K}_i]] = \mathbb{E}\left[\mathbb{1}_{A_i} \frac{h(M^{(1)}) - h(M^{(-1)})}{2}\right],$$

where we used $A_i \in \mathcal{K}_i$, $M = M^{(s_i)}$ and that s_i is independent of $\mathcal{K}_i$. By Lemma F.3 we have $\mathbb{1}_{A_i} M^{(1)} = \mathbb{1}_{A_i} M^{(-1)}$ and hence $\mathbb{E}[\mathbb{1}_{A_i} Z s_i] = 0$. The definition of conditional expectation hence implies $\mathbb{E}[\mathbb{1}_{A_i} s_i \mid \mathcal{H}_T] = 0$, which finishes the first part.

For the second statement note that $\psi(\nu, x) = \alpha(\nu) + \beta(\nu)x$, where

$$\alpha(\nu) = \frac{\psi(\nu, 1) + \psi(\nu, -1)}{2}, \qquad \beta(\nu) = \frac{\psi(\nu, 1) - \psi(\nu, -1)}{2}$$

and thus

$$\mathbb{1}_{A_i}\mathbb{E}\left[\psi(\nu, s_i) \,|\, \mathcal{H}_T\right] = \mathbb{1}_{A_i}\alpha(\nu) + \beta(\nu)\mathbb{E}\left[\mathbb{1}_{A_i} s_i \,|\, \mathcal{H}_T\right] = \mathbb{1}_{A_i}\alpha(\nu),$$

where we used the previous result in the last step. This finishes the proof. □

We now choose s uniformly over $\{\pm 1\}^\kappa$ and use the uncertainty in the undiscovered signs to lower-bound $\mathbb{E}\left[\|\nabla F_s(x_T)\|_{\mathrm{S}_1}\right]$.

Proposition F.5. *Let $s \sim \mathrm{Unif}\left(\{\pm 1\}^\kappa\right)$. Then, for every possibly randomised algorithm $\mathcal{A}$ and every $T \in \mathbb{N}_{\geq 1}$, we have*

$$\mathbb{E}\left[\|\nabla F_s(x_T)\|_{\mathrm{S}_1}\right] \geq \delta\Big(1 - \frac{q}{\kappa}\Big)^T.$$

Proof. Denote the sequence generated by $\mathcal{A}$ on F_s as $(x_1, g_1), \ldots, (x_T, g_T)$ and define $\mathcal{H}_T$ as in the previous lemma. For $i \in [\kappa]$ denote $\nu_i \coloneqq \langle E_i, x_T\rangle$ and note that ν_i is $\mathcal{H}_T$-measurable. Moreover, we have

$$\langle E_i, \nabla F_s(x_T)\rangle = \frac{L\nu_i - \delta s_i}{\kappa}$$

and hence, by Lemma F.2,

$$\|\nabla F_s(x_T)\|_{\mathrm{S}_1} \geq \sum_{i=1}^{\kappa} |\langle E_i, \nabla F_s(x_T)\rangle| = \sum_{i=1}^{\kappa} \frac{|L\nu_i - \delta s_i|}{\kappa}.$$

Taking expectation we obtain

$$\mathbb{E}\left[\|\nabla F_s(x_T)\|_{\mathrm{S}_1}\right] \geq \sum_{i=1}^{\kappa} \mathbb{E}\left[\frac{|L\nu_i - \delta s_i|}{\kappa}\right] \geq \sum_{i=1}^{\kappa} \mathbb{E}\left[\mathbb{1}_{A_i}\frac{|L\nu_i - \delta s_i|}{\kappa}\right]$$

and, by Lemma F.4,

$$\mathbb{E}\left[\mathbb{1}_{A_i}\frac{|L\nu_i - \delta s_i|}{\kappa}\right] = \mathbb{E}\left[\mathbb{1}_{A_i}\mathbb{E}\left[\frac{|L\nu_i - \delta s_i|}{\kappa}\,\middle|\,\mathcal{H}_T\right]\right] \geq \mathbb{E}\left[\mathbb{1}_{A_i}\frac{|L\nu_i - \delta| + |L\nu_i + \delta|}{2\kappa}\right].$$

Finally, note that $\frac{|L\nu_i - \delta| + |L\nu_i + \delta|}{2\kappa} \geq \frac{\delta}{\kappa}$, and $\mathbb{E}\left[\mathbb{1}_{A_i}\right] = \Pr(A_i) = (1 - \frac{q}{\kappa})^T$ to close the proof. □

Finally we provide the proof of Theorem 3.8.

Proof of Theorem 3.8. Let $\delta \coloneqq 2\varepsilon$ which satisfies $\delta \leq \frac{\sigma}{2}$ by our assumption on ε. By Lemma F.1, the function F_s satisfies

$$F_s(0) - \inf_{x \in \mathbb{R}^{m\times n}} F_s(x) \leq \frac{\delta^2}{2L} = \frac{2\varepsilon^2}{L} \leq \Delta_1$$

by our assumption on ε. Hence $F_s \in \mathcal{F}_{\Delta_1, L}$ and $\nabla f_s \in \mathcal{O}_{p,\sigma}$. Furthermore, for $s \sim \mathrm{Unif}\left(\{\pm 1\}^\kappa\right)$, Proposition F.5 implies

$$\mathbb{E}\left[\|\nabla F_s(x_T)\|_{\mathrm{S}_1}\right] \geq \delta\Big(1 - \frac{q}{\kappa}\Big)^T$$

and hence, for the output to satisfy $\mathbb{E}\left[\|\nabla F_s(x_T)\|_{\mathrm{S}_1}\right] \leq \varepsilon$, we need $2\varepsilon(1 - q/\kappa)^T \leq \varepsilon$. This is equivalent to

$$\Big(1 - \frac{q}{\kappa}\Big)^T \leq \frac{1}{2} \Leftrightarrow T\log\Big(1 - \frac{q}{\kappa}\Big) \leq \log\Big(\frac{1}{2}\Big) \Leftrightarrow T \geq \frac{\log(2)}{-\log\left(1 - \frac{q}{\kappa}\right)}.$$

Since $\kappa \geq 2$ we have $q/\kappa \leq 1/2$ and thus we can use $-\log(1 - x) \leq 2x\log(2)$, for $x \in [0, 1/2]$, to get

$$-\log\Big(1 - \frac{q}{\kappa}\Big) \leq 2\log(2)\frac{q}{\kappa}.$$

Combining yields

$$T \geq \frac{\log(2)\kappa}{2\log(2)q} = \frac{\kappa}{2} \cdot \Big(\frac{\sigma}{4\varepsilon}\Big)^{\frac{p}{p-1}},$$

finishing the proof. □

F.2 Non-Convex Objectives

We now prove Theorem 3.9 and the Schatten-norm extension in Theorem F.9. We first recall the Euclidean vector construction of Arjevani et al. (2023), which we then embed into matrix space to obtain the Schatten lower bounds.

We introduce the Euclidean $\mathbb{R}^d$-version of the function class as

$$\mathcal{F}^{(2)}_{\Delta_1,L}(d) \coloneqq \left\{ F\colon \mathbb{R}^d \to \mathbb{R} \;\middle|\; \|\nabla F(x) - \nabla F(y)\|_2 \le L\|x-y\|_2, F(0) - \inf_x F(x) \le \Delta_1 \right\}$$

and the Euclidean $\mathbb{R}^d$-version of the oracle class as

$$\mathcal{O}^{(2)}_{p,\sigma}(d) \coloneqq \left\{ \nabla f\colon \mathbb{R}^d \times \Xi \to \mathbb{R}^d \;\middle|\; \begin{array}{c} \mathbb{E}\left[\nabla f(x,\xi)\right] = \nabla F(x), \\ \mathbb{E}\left[\|\nabla f(x,\xi) - \nabla F(x)\|_2^p\right] \le \sigma^p, \forall x \in \mathbb{R}^d \end{array} \right\}.$$

Lemma F.6 (Randomised Euclidean hard instance)**.** *Let $c = 1/12000$, $\Delta_1, L > 0$, $\theta \in (0,1]$ and $0 < \varepsilon \le c\sqrt{\Delta_1 L}$. Then there exist a dimension d and a distribution over objectives $F \in \mathcal{F}^{(2)}_{\Delta_1,L}(d)$ with stochastic gradient oracles ∇f such that, for every possibly randomised first-order algorithm, any output $\hat{x}$ based on at most $c^2\Delta_1 L/(\varepsilon^2\theta)$ independent oracle samples satisfies*

$$\mathbb{E}\left[\|\nabla F(\hat{x})\|_2\right] \ge \frac{3}{2}\varepsilon,$$

where the expectation is over all randomness. Moreover, for every realisation the oracle has the form

$$\nabla f(x,\xi) = \nabla F(x) + \left(\frac{\xi}{\theta} - 1\right)h(x), \qquad \|h(x)\|_2 \le 2000\varepsilon, \qquad \xi \sim \mathrm{Bernoulli}(\theta),$$

and the dimension satisfies

$$d \le 64\frac{\Delta_1 L}{\varepsilon^2}\log\left(\frac{2}{\theta}\left(\frac{\Delta_1 L}{\varepsilon^2}\right)^2\right).$$

Proof. This follows directly from Arjevani et al. (2023, Proposition 1 and the proof of Theorem 3), leaving the Bernoulli parameter free and using failure probability $1/4$ in their Lemma 6. The noise representation and bound follow from their equation (60) and Lemma 15. □

We now embed independent Euclidean instances in columns with disjoint row supports. The number of copies will determine the dimension dependence of the lower bound.

Let $d, \kappa \in \mathbb{N}_{\ge 1}$. We place each of the κ Euclidean instances in one column with its own block of d rows, giving a matrix with $m \coloneqq \kappa d$ rows and $n \coloneqq \kappa$ columns. For $i \in [\kappa]$ and $v \in \mathbb{R}^d$, let $E_i(v) \in \mathbb{R}^{m\times n}$ be the matrix containing v in rows $(i-1)d+1, \dots, id$ of column i and zeros elsewhere. For $A \in \mathbb{R}^{m\times n}$, let $P_i(A) \in \mathbb{R}^d$ extract these entries. Then the following properties hold.

Lemma F.7. *Let $v_1, \dots, v_\kappa \in \mathbb{R}^d$, and $A \in \mathbb{R}^{m\times n}$. Then, for every $s \in [1,\infty]$,*

$$\left\|\sum_{i=1}^{\kappa} E_i(v_i)\right\|_{S_s} = \left(\sum_{i=1}^{\kappa}\|v_i\|_2^s\right)^{1/s}, \qquad \left(\sum_{i=1}^{\kappa}\|P_i(A)\|_2^s\right)^{1/s} \le \|A\|_{S_s}.$$

For $s = \infty$, the expressions involving sums are interpreted as $\max_i \|v_i\|_2$ and $\max_i \|P_i(A)\|_2$, respectively.

Proof. The columns of $\sum_{i=1}^{\kappa} E_i(v_i)$ have disjoint row support, so

$$\left(\sum_{i=1}^{\kappa} E_i(v_i)\right)^\top \left(\sum_{i=1}^{\kappa} E_i(v_i)\right) = \mathrm{Diag}\left(\|v_1\|_2^2, \dots, \|v_\kappa\|_2^2\right)$$

and the singular values are hence $\|v_1\|_2, \dots, \|v_\kappa\|_2$, which proves the first identity. For the second claim, let s' be conjugate to s and take $v_1, \dots, v_\kappa \in \mathbb{R}^d$ with $(\sum_i \|v_i\|_2^{s'})^{1/s'} \le 1$, using the maximum when $s' = \infty$. Then

$$\sum_{i=1}^{\kappa} \langle P_i(A), v_i \rangle = \left\langle A, \sum_{i=1}^{\kappa} E_i(v_i) \right\rangle \le \|A\|_{S_s} \left\| \sum_{i=1}^{\kappa} E_i(v_i) \right\|_{S_{s'}} \le \|A\|_{S_s},$$

where we used Schatten-norm duality and the first identity. By duality of the Euclidean and ℓ_s norms, the supremum of the left-hand side over these vectors is $(\sum_i \|P_i(A)\|_2^s)^{1/s}$, which proves the claim. □

This allows us to embed the hard instance from $\mathbb{R}^d$ into $\mathbb{R}^{m\times n}$ as follows. Draw κ independent pairs $(F_i^{(2)}, \nabla f_i^{(2)})$ from the distribution provided by Lemma F.6, with common parameters $\Delta_1^{(2)}, L^{(2)}, \varepsilon^{(2)}, \theta$. These instances remain fixed throughout the algorithm. Then we define

$$F(x) \coloneqq \sum_{i=1}^{\kappa} F_i^{(2)}(P_i(x)), \qquad \nabla f(x, \xi) \coloneqq \sum_{i=1}^{\kappa} E_i\Big(\nabla f_i^{(2)}(P_i(x), \xi_i)\Big), \tag{14}$$

where $\xi_i \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(\theta)$ are independent across blocks and oracle calls. We next verify that the matrix objective and oracle satisfy Assumptions 1 to 3. The Bernoulli parameter θ is chosen to satisfy the noise bound.

Proposition F.8. *Let $p \in (1, 2]$, $r \in (p/(p-1), \infty]$ with conjugate exponent r', and*

$$\Delta_1^{(2)} = \frac{\Delta_1}{\kappa}, \qquad L^{(2)} \coloneqq \frac{L}{\kappa^{1-2/r}}, \qquad \varepsilon^{(2)} \coloneqq \frac{\varepsilon}{\kappa^{1-1/r}}, \qquad \sigma^{(2)} \coloneqq \frac{\sigma}{\kappa^{1/p}}.$$

Let $F, \nabla f$ be defined as above. Furthermore let $c \coloneqq 1/12000$ and assume $0 < \varepsilon \le c \min\{\sigma, \sqrt{\Delta_1 L}\}$. Choose $\theta \coloneqq (\varepsilon^{(2)}/(c\sigma^{(2)}))^{p/(p-1)}$ and suppose $\kappa\theta \le 2$. Then $F, \nabla f$ satisfy Assumptions 1 to 3 in $(\mathbb{R}^{m\times n}, \|\cdot\|_{S_r})$ with initial point zero for every realisation of the Euclidean instances.

Proof. Our assumptions give $\varepsilon^{(2)} \le c\sqrt{\Delta_1^{(2)} L^{(2)}}$ and $\theta = \kappa^{-1+p/(r(p-1))}(\varepsilon/(c\sigma))^{p/(p-1)} \le 1$, so Lemma F.6 applies. For any realisation of the Euclidean instances, the initial gap satisfies

$$F(0) - \inf_x F(x) = \sum_{i=1}^{\kappa} \left(F_i^{(2)}(0) - \inf_x F_i^{(2)}(x)\right) \le \kappa \Delta_1^{(2)} = \Delta_1,$$

and, for $x, y \in \mathbb{R}^{m\times n}$,

$$\begin{aligned}
\|\nabla F(x) - \nabla F(y)\|_{S_{r'}} &= \left(\sum_{i=1}^{\kappa} \left\| \nabla F_i^{(2)}(P_i(x)) - \nabla F_i^{(2)}(P_i(y)) \right\|_2^{r'} \right)^{1/r'} \\
&\le L^{(2)} \left(\sum_{i=1}^{\kappa} \|P_i(x - y)\|_2^{r'} \right)^{1/r'} \\
&\le L^{(2)} \kappa^{1-2/r} \|x - y\|_{S_r} = L \|x - y\|_{S_r},
\end{aligned}$$

where we used Lemma F.7 and Hölder's inequality. This shows Assumptions 1 and 2. For the oracle properties, note that

$$\mathbb{E}\left[\nabla f(x, \xi)\right] = \sum_{i=1}^{\kappa} E_i\Big(\mathbb{E}\left[\nabla f_i^{(2)}(P_i(x), \xi_i)\right]\Big) = \sum_{i=1}^{\kappa} E_i\Big(\nabla F_i^{(2)}(P_i(x))\Big) = \nabla F(x).$$

To prove the bounded p-th central moment, we make use of the explicit structure of $\nabla f(x, \xi)$ to get

$$\begin{aligned}
\|\nabla f(x, \xi) - \nabla F(x)\|_{S_{r'}} &= \left(\sum_{i=1}^{\kappa} |\xi_i/\theta - 1|^{r'} \|h_i(P_i(x))\|_2^{r'} \right)^{1/r'} \\
&\le 2000\varepsilon^{(2)} \left(\kappa^{1/r'} + \frac{(\sum_{i=1}^{\kappa} \xi_i)^{1/r'}}{\theta} \right),
\end{aligned}$$

using Lemma F.7 and $\|h_i(P_i(x))\|_2 \le 2000\varepsilon^{(2)}$. Now let $N := \sum_{i=1}^{\kappa} \xi_i \sim \mathrm{Binomial}(\kappa, \theta)$, and use $p/r' \le 2$ and $\kappa\theta \le 2$ to obtain

$$\mathbb{E}\left[N^{p/r'}\right] \le \mathbb{E}\left[N^2\right] = \kappa\theta(1-\theta) + (\kappa\theta)^2 \le 3\kappa\theta.$$

Hence

$$\begin{aligned}\mathbb{E}\left[\|\nabla f(x,\xi) - \nabla F(x)\|_{S_{r'}}^p\right] &\le 2^{p-1}2000^p\left(\varepsilon^p + 3(\varepsilon^{(2)})^p\kappa\theta^{1-p}\right)\\ &= 2^{p-1}2000^p(\varepsilon^p + 3c^p\sigma^p)\\ &\le 4\cdot 2^{p-1}(2000c)^p\sigma^p \le \sigma^p,\end{aligned}$$

and hence the claim. □

We now prove Theorem 3.9 as a special case of the following Schatten lower bound.

Theorem F.9 (First-Order Schatten Lower Bound)**.** *Let $c = 1/12000$, $m, n \in \mathbb{N}_{\ge 1}$, $p \in (1, 2]$ and $r \in [1, \infty]$ with conjugate exponent r'. Suppose that $\Delta_1, L, \sigma > 0$, $0 < \varepsilon \le c\min\left\{\sigma, \sqrt{\Delta_1 L}\right\}$ and*

$$\max\{m, n\} \ge 64\frac{\Delta_1 L}{\varepsilon^2}\log\left(2\left(\frac{\Delta_1 L}{\varepsilon^2}\right)^2\left(\frac{c\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right).$$

Then there exists a distribution over objectives F and gradient oracles ∇f satisfying Assumptions 1 to 3 in $(\mathbb{R}^{m\times n}, \|\cdot\|_{S_r})$ with initial point zero, such that every output $\hat{x}$ of a possibly randomised first-order algorithm using T independent oracle samples satisfies

$$\mathbb{E}\left[\|\nabla F(\hat{x})\|_{S_{r'}}\right] \ge \frac{3}{2}\varepsilon \quad \textit{whenever} \quad T \le c^2\frac{\Delta_1 L}{\varepsilon^2}\left(\frac{c\sigma}{\varepsilon}\right)^{\frac{p}{p-1}},$$

where the expectation is over all randomness. Moreover, if $r > p/(p-1)$ and m, n satisfy the dimension condition in Theorem 3.9, then the same conclusion holds whenever

$$T \le c^2\frac{\Delta_1 L}{\varepsilon^2}\min\left\{\min\{m,n\}^{1-\frac{p}{r(p-1)}}\left(\frac{c\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}, \left(\frac{c\sigma}{\varepsilon}\right)^r\right\}.$$

For $r = \infty$, the second term in the minimum is omitted.

Proof. For both bounds, we use the Euclidean construction from Lemma F.6 and its matrix embedding (14). We may assume $m \ge n$ by transposing the construction.

For the bound valid for all $r \in [1, \infty]$, we use a single Euclidean instance and first verify its p-th noise bound. Choose the probability of revealing the hidden gradient component as $\theta := (\varepsilon/(c\sigma))^{p/(p-1)} \in (0, 1]$ and let $(F^{(2)}, \nabla f^{(2)})$ have the distribution provided by Lemma F.6 with parameters $\Delta_1, L, \varepsilon, \theta$. For each realisation, its noise satisfies

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla f^{(2)}(x,\xi) - \nabla F^{(2)}(x)\right\|_2^p\right] &\le (2000\varepsilon)^p\mathbb{E}\left[|\xi/\theta - 1|^p\right]\\ &\le 2(2000\varepsilon)^p\theta^{1-p} = 2(2000c)^p\sigma^p \le \sigma^p.\end{aligned}$$

We now embed this instance in the first column by (14) with $\kappa = 1$, padding unused rows and columns with zeros. The dimension assumption ensures $m \ge d$. We next verify Assumptions 1 to 3 for the resulting matrix instance. By Lemma F.7, all Schatten norms of $E_1(v)$ equal $\|v\|_2$, and $\|P_1(x)\|_2 \le \|x\|_{S_r}$. The initial gap and noise moment are therefore unchanged, unbiasedness follows by linearity of E_1, and smoothness follows from

$$\|\nabla F(x) - \nabla F(y)\|_{S_{r'}} = \left\|\nabla F^{(2)}(P_1(x)) - \nabla F^{(2)}(P_1(y))\right\|_2 \le L\|x - y\|_{S_r}.$$

Any algorithm reaching ε-stationarity for this matrix instance would also reach it for the Euclidean instance with the same number of oracle calls. Indeed, the matrix oracle at x returns E_1 applied to the Euclidean oracle response at $P_1(x)$. Since $\|\nabla F(\hat{x})\|_{S_{r'}} = \left\|\nabla F^{(2)}(P_1(\hat{x}))\right\|_2$, the bound follows from Lemma F.6.

For $r > p/(p-1)$, we obtain the stronger bound by combining κ independent Euclidean instances. We therefore let

$$\kappa := \min\left\{n, \left\lceil \left(\frac{c\sigma}{\varepsilon}\right)^r \right\rceil\right\}$$

when $r < \infty$, and $\kappa := n$ when $r = \infty$. For each Euclidean instance, we choose the parameters $\Delta_1^{(2)}, L^{(2)}, \varepsilon^{(2)}, \sigma^{(2)}$ and θ as in Proposition F.8, i.e.,

$$\frac{\Delta_1^{(2)} L^{(2)}}{(\varepsilon^{(2)})^2} = \frac{\Delta_1 L}{\varepsilon^2}, \qquad \theta = \kappa^{-1+\frac{p}{r(p-1)}} \left(\frac{\varepsilon}{c\sigma}\right)^{\frac{p}{p-1}}.$$

Note that $\kappa\theta \le 2^{p/(r(p-1))} \le 2$ for finite r, and $\kappa\theta \le 1$ for $r = \infty$. Draw κ independent pairs $(F_i^{(2)}, \nabla f_i^{(2)})$ from the distribution in Lemma F.6 with these parameters, and define $F, \nabla f$ by (14). Thus, by Proposition F.8, the function and oracle satisfy the required assumptions for every realisation. The dimension assumption ensures that the blocks fit in $\mathbb{R}^{m\times n}$. Indeed, $\theta^{-1} \le n(c\sigma/\varepsilon)^{p/(p-1)}$ and Lemma F.6 give

$$\kappa d \le 64n\frac{\Delta_1 L}{\varepsilon^2} \log\left(2n\left(\frac{\Delta_1 L}{\varepsilon^2}\right)^2 \left(\frac{c\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right) \le m.$$

As $\kappa \le n$, we can therefore pad unused rows and columns with zeros.

Now let $\mathcal{A}$ be any possibly randomised first-order algorithm and denote its queries on $F, \nabla f$ by $(x_t)_{t=1}^T$. Fix $i \in [\kappa]$. We show that the projected sequence $(P_i(x_t))_{t=1}^T$ can be generated by a randomised Euclidean algorithm using only the oracle of instance i. For this, sample the other instances and their oracle samples $\xi_{j,t}$ as internal randomness, independently of instance i. At each query x_t, obtain g_t from the oracle of instance i at $P_i(x_t)$ and give $\mathcal{A}$ the response

$$E_i(g_t) + \sum_{j\neq i} E_j\left(\nabla f_j^{(2)}(P_j(x_t), \xi_{j,t})\right).$$

The terms with $j \neq i$ are computed from the sampled instances. By induction, the instances, queries and output have the same joint distribution as when $\mathcal{A}$ is run on the matrix oracle. Returning $P_i(\hat{x})$ therefore gives a Euclidean algorithm using T oracle calls. In particular, Lemma F.6 yields

$$\mathbb{E}\left[\left\|\nabla F_i^{(2)}(P_i(\hat{x}))\right\|_2\right] \ge \frac{3}{2}\varepsilon^{(2)} \quad \text{whenever} \quad T \le \frac{c^2\Delta_1^{(2)}L^{(2)}}{(\varepsilon^{(2)})^2\theta} = \frac{c^2\Delta_1 L}{\varepsilon^2\theta}.$$

Combining these bounds using Lemma F.7 and Hölder's inequality gives

$$\mathbb{E}\left[\|\nabla F(\hat{x})\|_{S_{r'}}\right] \ge \kappa^{-1/r}\sum_{i=1}^{\kappa}\mathbb{E}\left[\left\|\nabla F_i^{(2)}(P_i(\hat{x}))\right\|_2\right] \ge \frac{3}{2}\kappa^{1-1/r}\varepsilon^{(2)} = \frac{3}{2}\varepsilon.$$

Finally, for finite r, the choice of κ gives

$$\frac{1}{\theta} = \kappa^{1-\frac{p}{r(p-1)}}\left(\frac{c\sigma}{\varepsilon}\right)^{\frac{p}{p-1}} \ge \min\left\{n^{1-\frac{p}{r(p-1)}}\left(\frac{c\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}, \left(\frac{c\sigma}{\varepsilon}\right)^r\right\}.$$

Thus the stated restriction on T implies $T \le c^2\Delta_1 L/(\varepsilon^2\theta)$, as required. For $r = \infty$, the same conclusion follows from $\theta^{-1} = n(c\sigma/\varepsilon)^{p/(p-1)}$. □

Taking $r = \infty$ in Theorem F.9 proves Theorem 3.9.

G Experimental Details

In this section we provide further details and extensions of our experiments in Section 4.

Further Details. We follow the usual convention of training all 2-d parameters but the input- and output-embeddings with MUON, while all other parameters are trained with ADAMW. For the Schatten-r sweep, we vary the descent geometry on these matrices.

Table 2: Mean layerwise noise ratio $\sigma_{S_1,p}/\sigma_{F,p}$ for different moment orders p.

p	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Initialisation	8.320	8.274	8.216	8.142	8.048	7.931	7.785	7.609	7.399	7.155
Final	13.162	13.162	13.162	13.162	13.161	13.161	13.161	13.161	13.161	13.160

Computational Resources. All experiments were conducted on an internal cluster, using 2 H100 GPUs. Overall the experiments required 606 GPU hours, out of which 311 were spent on preliminary experiments, 283 on the Schatten-r experiments, and an additional 12 GPU hours to calculate the gradient noise.

Licenses. The FineWeb-Edu dataset (Penedo et al., 2024) is distributed under the Open Data Commons Attribution License (ODC-By) v1.0. Our use complies with this license and adheres to CommonCrawl's Terms of Use.

Details on Figure 1 and 2. We consider the checkpoint at initialisation and at the end of training for the optimal learning rate $\eta = 0.015$, and seed 0. For both checkpoints, we first calculate the true gradient $\nabla F(x_i), i \in \{1, T\}$ by going over the whole 1.4B token training dataset $\{\xi_1, \ldots, \xi_T\}$. In a second pass we then calculate, for each weight matrix $x^{(j)}$

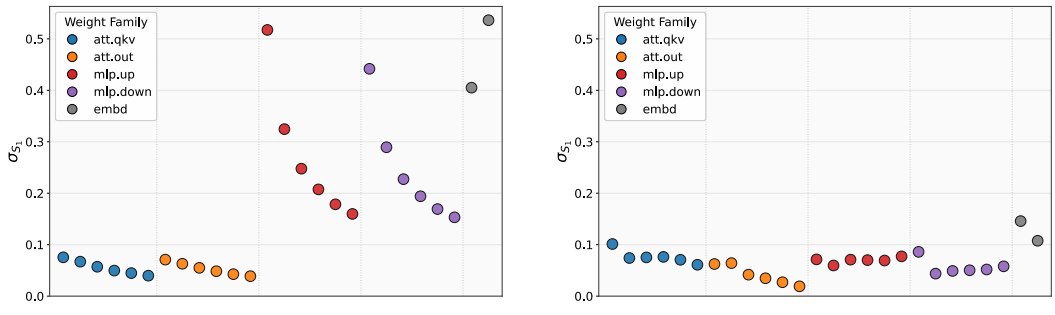
$$\sigma^p_{S_1,p} = \frac{1}{T}\sum_{\tau=1}^{T} \|\partial_{x^{(j)}} F(x_i) - \partial_{x^{(j)}} f(x_i, \xi_\tau)\|^p_{\mathrm{S}_1}\,,$$

$$\sigma^p_{F,p} = \frac{1}{T}\sum_{\tau=1}^{T} \|\partial_{x^{(j)}} F(x_i) - \partial_{x^{(j)}} f(x_i, \xi_\tau)\|^p_{\mathrm{F}}\,,$$

where $\xi_1, \ldots, \xi_T$ are the mini-batches of size 512. We then calculate the ratio $\sigma_{S_1,p}/\sigma_{F,p}$ for all weight matrices trained by MUON and report them in Figure 1. We additionally visualise the plain $\sigma_{S_1,p}$ and $\sigma_{\ell_1,p}$ in Figure 3. For Figure 2, the ℓ norms act on the vector of matrix entries. We normalise each ratio $R = \sigma_{1,p}/\sigma_{2,p}$ as $(R-1)/(\rho-1)$, where $\rho = \sqrt{mn}$ for the entrywise norms and $\rho = \sqrt{\min\{m,n\}}$ for the Schatten norms of an $m \times n$ weight matrix.

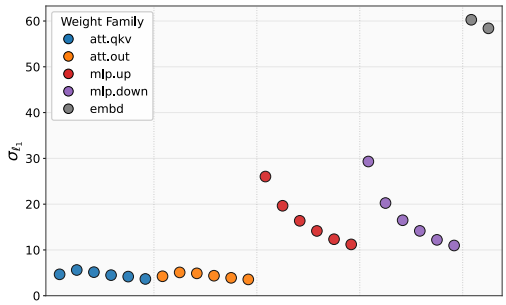
Sensitivity to p. We repeat the noise-ratio measurements for $p \in \{1.1, 1.2, \ldots, 2.0\}$ and report the arithmetic mean of the layerwise ratios $\sigma_{S_1,p}/\sigma_{F,p}$ over all weight matrices trained by MUON in Table 2. Across all tested values of p, the mean ratio remains below the worst-case factor of approximately 22.5 and is larger at the final checkpoint than at initialisation.

Details on Table 1. To prevent confounding from the choice of learning rates, we additionally provide the full grid of results for the learning rate and Schatten-r sweep in Figure 4. Importantly, none of the optimal learning rates lie on the end points of the grid.

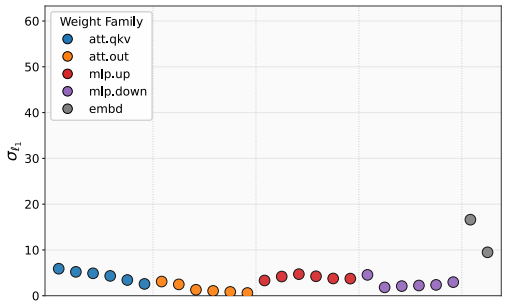


(a) σ_{S_1} at Init

(b) σ_{S_1} at Final



(c) σ_{ℓ_1} at Init



(d) σ_{ℓ_1} at Final

Figure 3: Plain noise values.

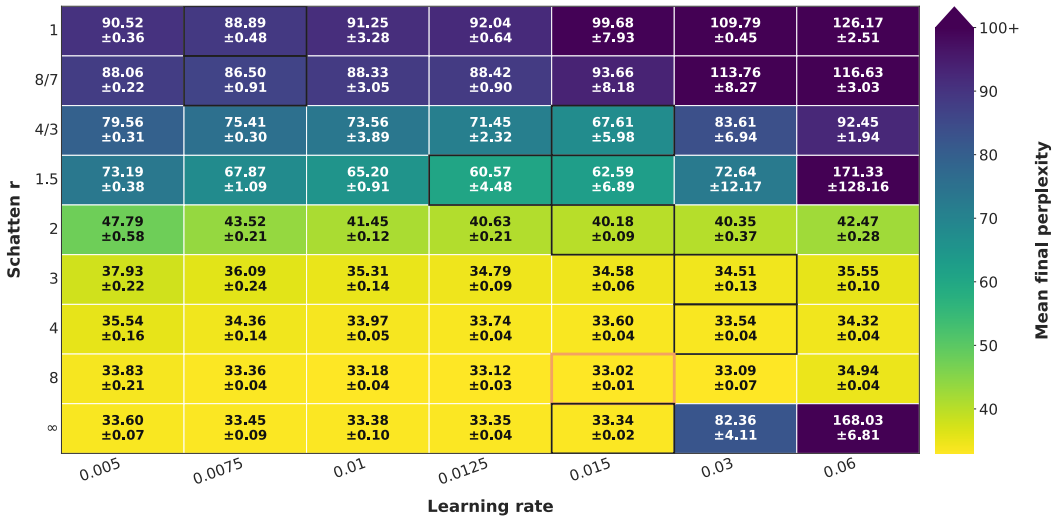


Figure 4: Full result grid for the Schatten-r sweep. The value represents the mean final perplexity $\pm$ its standard deviation over the seeds $0, 1, 2$. Optimal learning rates per geometry are marked with black borders, the overall best result with an orange border.

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 provide proofs or empirical evidence for all claims.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Limitations are discussed in Section 6.

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: Assumptions are included in the theory statements. The corresponding proofs can be found in the Appendix.

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 public code is linked in Section 4. All hyper-parameters including seeds are specified in Section 4.

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: The public code and dataset are linked in Section 4.

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 details can be found in Section 4.

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We provide error bars in Section 4.

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: See Section G.

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: This work is of theoretic nature.

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 work studies generic optimisation algorithms.

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: No data or models are released.

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: See Section G.

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: No new assets are released.

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: This work does not involve crowdsourcing or research with human subjects.

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: No human subjects took part in this study.

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 developments in this research did not involve LLMs as any important, original, or non-standard component.