WHEN TO USE WHAT SCHATTEN-p NORM IN DEEP LEARNING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Schatten-∞ based optimizers such as Muon have shown promising empirical performance, but there remains seemingly conflicting observations regarding whether they are beneficial. We resolve this conflict by showing that the conclusion is regime dependent. Even when the natural problem geometry is the Schatten-∞ norm, smaller Schatten-p geometries can be optimal, specifically in the low-dimensional regime, which we show includes Chinchilla scaling. This conclusion follows from a new noise-robust acceleration result for the SODA framework for $p > 2$. The same analysis explains why Muon-like methods do not require warmup, why they naturally favor large batches, and yields a batch size scaling rule for arbitrary p.

1 INTRODUCTION

Schatten-∞ based optimizers such as Muon (Jordan et al., 2024b) have seen substantial adoption in the machine learning community. However, there remain seemingly conflicting observations around when, and whether, Muon is beneficial.

One prevailing consensus is that the batch size needs to be taken large to see the benefits of the Schatten-∞ norm. This was the case in the original spectral methods (Carlson et al., 2016; 2015), later in the CIFAR-10 speedrun setting where Muon was developed (Jordan, 2024), and further documented in Pethick et al. (2025); Shah et al. (2025). It also seems that Muon is particularly beneficial for short training runs, such as speedruns on modded-nanogpt (Jordan et al., 2024a).

However, for large-scale training runs, where the token budget is also much larger, the verdict is less clear as shown in large-scale benchmark studies such as Wen et al. (2025); Semenov et al. (2025). Even SGD has been shown to outperform common training methods like Adam, specifically in the extremely low batch size regime (Srećković et al., 2025; Marek et al., 2026).

Additionally, HTMuon (Pang et al., 2026) and Soft-Muon (Abrahamsen, 2026), which instead use finite Schatten-p norms, have respectively led to a speedup on Llama pre-training and to the leading optimizer-track result on modded-nanogpt at the time of writing. Similarly, Shumaylov et al. (2026) dispute the need for exclusively using the Schatten-∞ norm.

Essentially, the mechanism behind Muon's success, and the regime in which it should be expected to be beneficial, is not fully understood. This raises the following more general question:

What Schatten-p norm should the optimizer use, and when?

To answer this question, we build on the SODA framework of Pethick et al. (2026), which combines optimistic dual averaging (Rakhlin & Sridharan, 2013) with the schedule-free optimization framework of Defazio et al. (2024). This lets us capture modern optimizer components precisely, including dualization, Nesterov momentum and weight decay.

From this perspective, we identify the possible source of the confusion. The choice of p is simultaneously tied to both the heavy-tailedness of the stochastic gradient and the geometry of the problem, *and* it affects the degree to which the method can accelerate. We find that *even when the objective is smooth in the Schatten-∞ geometry, it can be beneficial to use an optimizer with a smaller Schatten-p norm in the low-dimensional, or "overtraining," regime.* Thus, conclusions will be regime dependent.

Throughout, we use "p-norm" to refer to either a vector ℓ_p norm or a Schatten-p norm, depending on the ambient space. The first challenge we face is that existing analysis of SODA does not allow us to

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

Table 1: Instances of SODA obtained by various choices of p-norm. For Schatten norms, write $Y = U\,\mathrm{diag}(s)V^\top$, and for mixed norms let Y_i denote the ith row. Optimism means $\bar{\alpha}_k \neq 0$ with Nesterov momentum corresponding to $\bar{\alpha}_k = \alpha_k$.

Method	**Reference**	**Optimism**	**Norm**	**Dual map** ($\partial h^*/\nabla h^*$)
Lion	(Chen et al., 2023)	✓	ℓ_∞	$\mathrm{sign}(u)$
	–		ℓ_p	$\mathrm{sign}(u) \odot \lvert u\rvert^{q-1}$
	–		mixed $\ell_{\infty,p}$	$\lVert Y_i\rVert_1^{q-1}\,\mathrm{sign}(Y_i)$
Muon	(Jordan et al., 2024b)	✓	Schatten-∞	$UV^\top$
HTMuon[1]	(Pang et al., 2026)	✓	Schatten-p	$U\,\mathrm{diag}(s^{q-1})V^\top$

[1] The same finite-Schatten norm is used in Soft-Muon and Freon (Abrahamsen, 2026; Shumaylov et al., 2026).

capture p-norm geometries for $p > 2$. A major contribution is thus to extend the analysis of SODA to these p-norm geometries by considering p-uniformly convex reference functions.

Contributions. Concretely, we make the following contributions:

(i) We prove that SODA achieves accelerated rates for optimization over p-norm geometries while remaining robust to stochastic noise. In particular, we obtain the *dimension-free* rate

$$O\left(\frac{L_p R_p^2}{n^{1+2/p}} + \frac{\sigma_q R_p}{n^{1/p}}\right),$$

revealing a connection between the choice of geometry p and the heaviness of the noise tails through the dual exponent $q = p/(p-1)$. This appears to be the first noise-robust acceleration result for arbitrary p-norm geometries. The result extends to the Hölder-smooth case and interestingly yields a p-dependent stepsize schedule.

(ii) We derive stochastic gradient oracle complexity bounds under possibly lighter-tailed noise by using large batch sizes. In the extreme case of bounded variance, the result shows that after N oracle calls, Euclidean methods recover the classical $O(N^{-1/2})$ stochastic error rate, whereas $p \to \infty$ gives $O(N^{-1/3})$. In the process, we develop a norm-dependent batch size scaling rule that increases with p as $B_p \asymp N^{2(p-1)/(3p-2)}$.

(iii) We interpret these results and show that a ∞-norm problem geometry does *not* always imply that the ∞-geometry is optimal for the algorithm. Lower values of p can be preferable both in deterministic settings, due to stronger acceleration, and under heavy-tailed noise, due to better alignment with the noise assumption. This advantage appears in the low-dimensional regime, where the effective dimension d is smaller than the oracle budget N, i.e., $d \lesssim N$. This regime dependency suggests a switching strategy of p during training, which we discuss.

(iv) We connect the low-dimensional regime with modern scaling rules, illustrating that common scaling rules in fact place training well within the low-dimensional regime by scaling the token budget N as d^3 where d is the effective dimension of the model matrices. This might explain why Muon is not necessarily favored under typical scaling recipes.

2 Method

Let $\mathcal{X}$ be a finite-dimensional normed space with dual $\mathcal{X}^*$. We rely on the Fenchel conjugate to map gradient information back to the primal space. For a proper closed convex reference function $h : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$, the Fenchel conjugate over $\mathcal{X}$ is defined as,

$$h^*(m) = \sup_{x \in \mathcal{X}} \{\langle m, x\rangle - h(x)\}, \qquad m \in \mathcal{X}^*.$$

We analyze SODA (Pethick et al., 2026), which generalizes Optimistic Dual Averaging (ODA) by introducing a primal extrapolation sequence (y^k) from Tseng (2008); Lan (2012); Defazio et al. (2024):

$$\begin{aligned} m^{k+1} &= (1-\alpha_k)m^k + \alpha_k \nabla f(y^k, \xi_k), \\ \bar{m}^{k+1} &= (1-\bar{\alpha}_k)m^{k+1} + \bar{\alpha}_k \nabla f(y^k, \xi_k), \\ z^{k+1} &\in \partial h^*(-\gamma_k \bar{m}^{k+1}) = \arg\min_{x \in \mathcal{X}} \gamma_k \langle \bar{m}^{k+1}, x \rangle + h(x), \\ x^{k+1} &= (1-\lambda_k)x^k + \lambda_k z^{k+1}, \\ y^{k+1} &= (1-\bar{\lambda}_k)x^{k+1} + \bar{\lambda}_k z^{k+1}. \end{aligned} \tag{SODA}$$

Here $\alpha_k, \bar{\alpha}_k, \lambda_k, \bar{\lambda}_k \in [0,1]$ and $\gamma_k > 0$. We initialize $m^0 = 0$ and $x^0 = y^0 = z^0 \in \partial h^*(0)$. For $\bar{\lambda}_k = 0$ we obtain an optimistic version of the Double Averaging (Nesterov & Shikhman, 2015).

What is particularly attractive about this algorithmic template in the context of deep learning is that it allows us to capture commonly used techniques such as Nesterov momentum ($\bar{m}^{k+1}$), weight decay (parameterized by λ_k when $\bar{\lambda}_k = 0$) and the use of different geometries (∂h^*), all in a way where their impact can be understood theoretically.

2.1 Norm choices

Previous analyses of SODA were based on strongly convex reference functions, which naturally cover p-norm geometries for $p \in (1,2]$. We instead use uniform convexity, relying on the fact that $h(x) = \frac{1}{p}\|x\|_p^p$ is p-uniformly convex (see Definition 1). At the endpoint $p = \infty$, this regularizer becomes a hard ∞-norm constraint, thus recovering the Schatten-∞ linear minimization oracle (lmo) used in e.g., Muon (Jordan et al., 2024b) and Scion (Pethick et al., 2025).

Vector ℓ_p geometry For vector variables $x \in \mathbb{R}^d$, choose the fixed reference function centered at the initial point z^0,

$$h(x) = \tfrac{1}{p}\left\|x - z^0\right\|_{\ell_p}^p, \qquad p > 2, \qquad q = \tfrac{p}{p-1}.$$

Then $h^*(u) = \langle u, z^0 \rangle + \frac{1}{q}\|u\|_{\ell_q}^q$, and the step in SODA has the following closed form solution:

$$\nabla h^*(-\bar{m}) = z^0 - \mathrm{sign}(\bar{m}) \odot |\bar{m}|^{q-1},$$

where the sign, absolute value, power, and product are taken componentwise. When $z^0 = 0$, this dual map interpolates between the identity map, $\nabla h^*(g) = g$, at the Euclidean endpoint $p = 2$ and the sign map at $p = \infty$. This is different from, for example, the linear minimization oracle in Frank-Wolfe, which interpolates between *normalized* gradients and the sign map.

Schatten-p geometry For matrix variables $X \in \mathbb{R}^{d_1 \times d_2}$, choose instead

$$h(X) = \tfrac{1}{p}\left\|X - Z^0\right\|_{S_p}^p = \tfrac{1}{p}\textstyle\sum_i s_i(X - Z^0)^p, \qquad p > 2, \qquad q = \tfrac{p}{p-1},$$

where $\|\cdot\|_{S_p}$ is the Schatten-p norm and s_i is the ith singular value. Then $h^*(Y) = \langle Y, Z^0 \rangle + \frac{1}{q}\|Y\|_{S_q}^q$. If $\bar{M} = U\,\mathrm{diag}(s)V^\top$ is a singular value decomposition (SVD), then

$$\nabla h^*(-\bar{M}) = Z^0 - U\,\mathrm{diag}(s^{q-1})V^\top.$$

This is the dualization used in HTMuon (Pang et al., 2026), Freon (Shumaylov et al., 2026) and Soft-Muon (Abrahamsen, 2026), with the only difference being that Z^0 can be taken different from zero. The SVD can be computed efficiently using QDWH (Nakatsukasa et al., 2010; Nakatsukasa & Freund, 2016) used in Freon or (warmstarted) power iteration as used in PowerSGD and Dion (Vogels et al., 2019; Ahn et al., 2025).

Spectral geometry At the endpoint $p = \infty$ and $Z^0 = 0$, the mapping becomes the linear minimization oracle over the Schatten-∞/spectral norm ball, $\partial h^*(-\bar{M}) \ni -UV^\top$, for which SODA exactly recovers Muon (Jordan et al., 2024b) with Nesterov momentum and weight decay by taking $\bar{\lambda}_k = 0$.

3 Analysis

We derive convergence guarantees for SODA under uniform convexity in order to generalize to p-norm for $p > 2$. The proof is based on an online regret argument, where we let g^k denote an arbitrary gradient-feedback sequence. In the stochastic optimization setting of SODA, we take $g^k := \nabla f(y^k, \xi_k)$.

Definition 1 (p-uniform convexity)**.** *For $p \geq 2$, a proper closed convex function h is p-uniformly convex with constant $\mu > 0$ with respect to $\|\cdot\|$ if*

$$D_h(u, v; s) := h(u) - h(v) - \langle s, u - v\rangle \geq \tfrac{\mu}{p}\|u - v\|^p \qquad \text{for all } u, v \in \mathcal{X},\ s \in \partial h(v).$$

The conjugate exponent is denoted by $q = p/(p-1)$. The case $p = 2$ is the usual strong convexity.

Assumption 1 (Convex)**.** *The objective f is convex.*

Assumption 2 (L-smooth)**.** *The function f is L-smooth with respect to $\|\cdot\|$.*

Remark 1 (Meaning of $\|\cdot\|_p$). The general statements are formulated for an arbitrary norm $\|\cdot\|$. When the corollaries specialize to $\|\cdot\|_p$, this notation covers both the vector norm $\|\cdot\|_{\ell_p}$ and the Schatten norm $\|\cdot\|_{S_p}$, with dual norm $\|\cdot\|_q$ interpreted in the same ambient space. We write L_p for the corresponding smoothness constant. The specialization is valid because, for finite $p \geq 2$, $h(u) = \frac{1}{p}\left\|u - z^0\right\|_p^p$ is p-uniformly convex with respect to either geometry with admissible constant $\mu = 2^{2-p}$. See Section A for other norms.

Assumption 3 (Unbiased)**.** *Let $\mathcal{F}_k$ be the natural filtration. Then $\mathbb{E}[g^k \mid \mathcal{F}_{k-1}] = \nabla f(y^k)$.*

We rely on the following heavy-tailed type noise assumption which arises naturally in the analysis since p-uniform convexity of h implies that h^* is q-uniformly smooth (Vural et al., 2022, Prop. 1).

Assumption 4 (Gradient variation)**.** *For some $q \in (1, 2]$, the gradients satisfy, $\tau_q \geq 1$, and $\sigma_q \geq 0$,*

$$\mathbb{E}\left[\left\|g^k - g^{k-1}\right\|_*^q\right]^{1/q} \leq \tau_q \mathbb{E}\left[\left\|\nabla f(y^k) - \nabla f(y^{k-1})\right\|_*^q\right]^{1/q} + \sigma_q, \qquad \forall k \geq 0.$$

We are now ready to state our first main convergence result.

Corollary 3.1 (Non-accelerated parameterization)**.** *Let $p \geq 2$, $x^\star \in \arg\min_{x \in \operatorname{dom} h} f(x)$, $h(x) = \frac{1}{p}\left\|x - z^0\right\|_p^p$, and $R_p := \left\|x^\star - z^0\right\|_p$. Consider SODA with*

$$\alpha_k = \tfrac{1}{k+1}, \qquad \bar{\alpha}_k = \lambda_k = \tfrac{1}{k+2}, \qquad \bar{\lambda}_k \leq \tfrac{\lambda_k}{6}, \qquad \gamma_k = \eta(k+2), \quad \forall k = 0, \ldots, n-1.$$

Suppose Assumptions 1 to 4 hold with $q = p/(p-1)$. Choose

$$\eta \asymp_{p,\tau_q} \min\left\{\tfrac{R_p^{p-2}}{L_p}, \tfrac{R_p^{p-1}}{\sigma_q\, n^{(p-1)/p}}\right\}.$$

Then, for every $n \geq 1$,

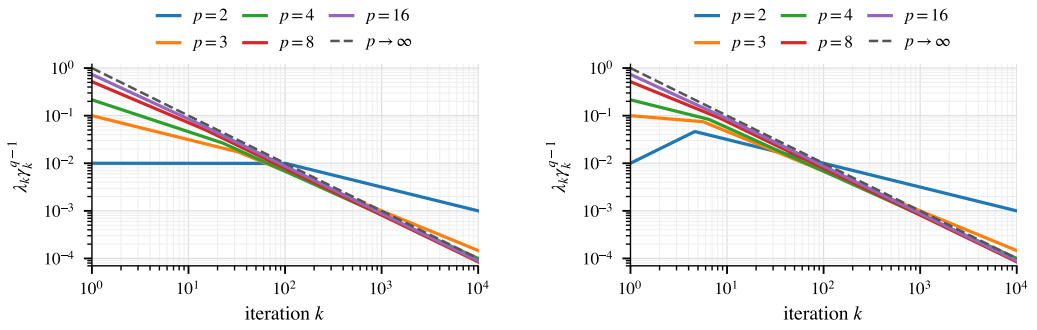
$$\mathbb{E}[f(x^{n-1}) - f(x^\star)] = O_{p,\tau_q}\left(\tfrac{L_p R_p^2}{n} + \tfrac{\sigma_q R_p}{n^{1/p}}\right).$$

For $p > 2$, this follows from the more general Theorem C.1, which allows Hölder smoothness, by taking $\nu = 1$ and $L_\nu = L$, while $p = 2$ is the strongly convex SODA guarantee of Pethick et al. (2026, Cor. 4.6). Corollary C.3 in the appendix shows that the ∞-geometry can be approximated by $p = O(\log(d))$ with only a logarithmic loss. For simplicity of presentation, in the remaining discussion involving $p \to \infty$, we refer to this result for $p = O(\log(d))$ and suppress the logarithmic dependence in the resulting rates and complexities.

Effective stepsize Let us interpret the stepsize choices in Corollary 3.1. For the finite-p regularizer $h(x) = \frac{1}{p}\left\|x - z^0\right\|_p^p$, the update for z^{k+1} in SODA contains the stepsize γ_k^{q-1}. To see how this stepsize changes with the geometry, consider the horizon-free version, where the horizon n in η is replaced by the current iteration k:[1]

$$\eta_k \asymp_{p,\tau_q} \min\left\{C_{\text{det}}, C_{\text{stoch}} k^{-(p-1)/p}\right\}, \quad \text{with} \quad C_{\text{det}} := \tfrac{R_p^{p-2}}{L_p}, \quad C_{\text{stoch}} := \tfrac{R_p^{p-1}}{\sigma_q}.$$

[1] We leave explicit treatment of such anytime parameterization for future work.



(a) Non-accelerated stepsize (Corollary 3.1). (b) Accelerated stepsize (Corollary 3.2).

Figure 1: Effective stepsize $\lambda_k \gamma_k^{q-1}$ for several geometries, using $C_{\text{det}} = 10^{-2}$ and $C_{\text{stoch}} = 10^{-1}$. Interestingly, we find a warmup-like phase for the accelerated parameterization, which is present in the Euclidean case but not for $p \to \infty$, consistent with practice (although the functional form differs).

Since $\gamma_k = \eta_k(k+2)$, this gives

$$\gamma_k^{q-1} \asymp_{p,\tau_q} \min\left\{C_{\text{det}}^{1/(p-1)} k^{1/(p-1)}, C_{\text{stoch}}^{1/(p-1)} k^{1/(p(p-1))}\right\}.$$

For the centered regularizer $h(x) = \frac{1}{p}\left\|x - z^0\right\|_p^p$, the averaged iterate update in SODA reduces to

$$x^{k+1} = (1-\lambda_k)x^k + \lambda_k z^0 - \lambda_k \gamma_k^{q-1} \nabla\left(\tfrac{1}{q}\|\cdot\|_q^q\right)(\bar{m}^{k+1}).$$

From this we see clearly that λ_k acts as weight decay (Hanson & Pratt, 1988) while $\lambda_k \gamma_k^{q-1}$ acts as the effective stepsize of the update, which we plot in Figure 1. Specifically, at the endpoints we see the effect of p on the effective stepsize

$$p = 2: \qquad \lambda_k \gamma_k^{q-1} \asymp_{\tau_2} \min\left\{\tfrac{1}{L_2}, \tfrac{R_2}{\sigma_2 k^{1/2}}\right\}, \qquad p \to \infty: \qquad \lambda_k \gamma_k^{q-1} \asymp k^{-1}.$$

For a more aggressive hyperparameter choice we can accelerate the smooth term (depending on p).

Corollary 3.2 (Accelerated parameterization)**.** *Let $p \geq 2$, $x^\star \in \arg\min_{x \in \operatorname{dom} h} f(x)$, $h(x) = \frac{1}{p}\left\|x - z^0\right\|_p^p$, and $R_p := \left\|x^\star - z^0\right\|_p$. Consider SODA with*

$$\alpha_k = \tfrac{2}{k+2}, \qquad \bar{\alpha}_k = \lambda_k = \tfrac{2}{k+3}, \qquad \bar{\lambda}_k \leq \tfrac{\lambda_k}{6}, \qquad \gamma_k = \eta \tfrac{(k+2)(k+3)}{2}, \quad \forall k = 0, \ldots, n-1.$$

Suppose Assumptions 1 to 4 hold with $q = p/(p-1)$. Choose

$$\eta \asymp_{p,\tau_q} \min\left\{\frac{R_p^{p-2}}{L_p\, n^{(p-2)/p}}, \frac{R_p^{p-1}}{\sigma_q\, n^{(2p-1)/p}}\right\}.$$

Then for every $n \geq 1$, we have

$$\mathbb{E}[f(x^{n-1}) - f(x^\star)] = O_{p,\tau_q}\left(\frac{L_p R_p^2}{n^{1+2/p}} + \frac{\sigma_q R_p}{n^{1/p}}\right).$$

For $p > 2$, this follows from the more general Theorem C.2, which allows Hölder smoothness, by taking $\nu = 1$ and $L_\nu = L$, while $p = 2$ is the accelerated strongly convex SODA guarantee of Pethick et al. (2026, Cor. B.6). For fixed finite $p \geq 2$, the deterministic term is tight up to constants depending on p (see Guzmán & Nemirovski (2018, Cor. 1)). For $p = \infty$, Guzmán & Nemirovski (2018, Cor. 1) only gives tightness up to a logarithmic factor in the dimension, thus matching Corollary C.3.

We can extract a remarkable number of insights from these bounds as we will see next.

3.1 BATCH SIZE UNDER LESS HEAVY-TAILED NOISE

The noise assumption in the convergence theorems is inherently tied to the algorithm norm, since choosing the primal geometry $\|\cdot\|_p$ fixes the dual norm $\|\cdot\|_q$, where $q = p/(p-1)$. This results in an overly pessimistic oracle complexity when a stronger noise assumption is available.

For mini-batching, we therefore strengthen the oracle assumption by requiring an r-moment with $r \geq q$. Larger r means less heavy-tailed noise, and the extra moment is what allows the effective noise scale to decrease with the batch size. We write

$$p_r := \tfrac{r}{r-1}, \qquad \tfrac{1}{p_r} = \tfrac{r-1}{r}.$$

Assumption 5 (Mean-zero r-moment stochastic gradients)**.** *For some $r \in (1, 2]$, for every $x \in \mathcal{X}$,*

$$\mathbb{E}_\xi[\nabla f(x,\xi)] = \nabla f(x), \qquad \mathbb{E}_\xi[\|\nabla f(x,\xi) - \nabla f(x)\|_r^r] \leq \sigma_r^r.$$

This implies the gradient-variation condition in Assumption 4 for $q = r$.

Here we keep the stochastic oracle structure explicit in order to track how batching changes the scale. For a mini-batch average

$$\bar{g}_B(x) := \tfrac{1}{B}\sum_{i=1}^B \nabla f(x,\xi_i),$$

the von Bahr–Esseen inequality (von Bahr & Esseen, 1965; Pinelis & Sakhanenko, 1986) gives

$$\mathbb{E}[\|\bar{g}_B(x) - \nabla f(x)\|_r^r \mid x]^{1/r} \lesssim_r \sigma_r B^{-1/p_r}.$$

Let d denote the ambient dimension, or the effective rank in the Schatten case. If $q \leq r$, then

$$\|u\|_q \leq d^{1/q-1/r}\,\|u\|_r\,. \tag{1}$$

Thus the only role of batching in the rate below is to replace σ_q by $d^{1/q-1/r}\sigma_r B^{-1/p_r}$, while keeping $\tau_q = 1$. We are now ready to state our complexity result.

Corollary 3.3 (Oracle complexity with less heavy-tailed noise)**.** *Let $p \geq 2$, set $q = p/(p-1)$, and suppose $p \geq p_r$. Suppose Assumption 5 holds. Run SODA with gradient feedback $g^k = \bar{g}_B(y^k)$. Then Corollary 3.1 applies with its noise parameter σ_q replaced by $d^{1/q-1/r}\sigma_r B^{-1/p_r}$, giving*

$$\mathbb{E}[f(x^{n-1}) - f(x^\star)] \lesssim \tfrac{A_p}{n} + \tfrac{S_{r,p}}{B^{1/p_r} n^{1/p}},$$

where $R_p := \left\|x^\star - z^0\right\|_p$, $A_p = L_p R_p^2$, and $S_{r,p} := \sigma_r d^{1/q-1/r} R_p$. Then accuracy ϵ is reached by taking

$$n \asymp \tfrac{A_p}{\epsilon}, \qquad B \asymp \left(\tfrac{S_{r,p}}{\epsilon}\right)^{p_r}\left(\tfrac{\epsilon}{A_p}\right)^{p_r/p},$$

and hence with total stochastic gradient budget

$$N = nB = O\left(\tfrac{A_p}{\epsilon} + \left(\tfrac{S_{r,p}}{\epsilon}\right)^{p_r}\left(\tfrac{A_p}{\epsilon}\right)^{1-p_r/p}\right).$$

There are a couple of insights we can extract from this complexity result.

Batch size rule for a fixed geometry Suppose the total stochastic gradient budget $N = nB$ is fixed instead of the accuracy ϵ. Substituting $B = N/n$ in the rate of Corollary 3.3 gives

$$\epsilon_p(n,N) \lesssim \tfrac{A_p}{n} + \tfrac{S_{r,p}}{N^{1/p_r}} n^{1/p_r - 1/p}, \qquad p \geq p_r.$$

At the smallest admissible geometry $p = p_r$, the stochastic term is independent of the iteration count after the substitution $B = N/n$. Thus any batch size choice with

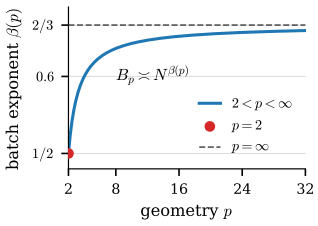
$$\tfrac{A_{p_r}}{n} \lesssim \tfrac{S_{r,p_r}}{N^{1/p_r}} \qquad \Longleftrightarrow \qquad 1 \leq B \lesssim \tfrac{S_{r,p_r}}{A_{p_r}} N^{1-1/p_r},$$

attains the same oracle complexity. For any other p the optimal batchsize is unique. Specifically, for $p > p_r$, balancing the smooth and stochastic terms gives the following scaling rule:

$$n_p \asymp \left(\tfrac{A_p}{S_{r,p}}\right)^{1/(1+1/p_r-1/p)} N^{\frac{1/p_r}{1+1/p_r-1/p}}, \qquad B_p = \tfrac{N}{n_p} \asymp \left(\tfrac{S_{r,p}}{A_p}\right)^{1/(1+1/p_r-1/p)} N^{\frac{1-1/p}{1+1/p_r-1/p}}. \tag{2}$$

Table 2: Batch size scaling rules for a given geometry under bounded variance. Here $N = nB$ is the total number of stochastic gradient oracle calls, which corresponds to the total token budget in LLMs. The batch size increases with p and recovers the BST rule for $p \to \infty$ (Islamov et al., 2026).

Geometry	Iterations	Batch size	Rate
$p = 2$	$n_2 \asymp N^{1/2}$	$B_2 \asymp N^{1/2}$	$\epsilon_N \asymp N^{-1/2}$
$2 < p < \infty$	$n_p \asymp N^{p/(3p-2)}$	$B_p \asymp N^{2(p-1)/(3p-2)}$	$\epsilon_N \asymp N^{-p/(3p-2)}$
$p \to \infty$	$n_\infty \asymp N^{1/3}$	$B_\infty \asymp N^{2/3}$	$\epsilon_N \asymp N^{-1/3}$



Bounded variance Let us get a feeling for this scaling by developing the special case of bounded variance, $r = 2$ and hence $p_r = 2$. The corresponding batch size scaling rules based on (2) are summarized in Table 2. Interestingly, for $p \to \infty$ we recover the BST batch size scaling rule of Islamov et al. (2026), which identically uses $B_\infty \asymp N^{2/3}$.

Specifically, for a fixed oracle budget N, the two endpoint stochastic error rates are

$$p = 2: \quad \epsilon_N = O(N^{-1/2}), \qquad p \to \infty: \quad \epsilon_N = O(d^{1/3} N^{-1/3}).$$

From the above it might appear as if the Euclidean geometry is always better than the ∞-geometry, both in terms of the dependence on the oracle budget N and the dimension d. However, this comparison hides that the smoothness and radius terms depend on the geometry. In the next section we show that, when the problem geometry is the ∞-norm, different regimes emerge depending on d and N.

4 Regimes

4.1 Choosing p under ∞-norm problem geometry

We now ask what the choice of the geometry for the algorithm should be in deep learning based on the derived complexities. Let us have some faith and assume that the natural geometry of neural networks is the Schatten-∞ norm[2], as e.g., argued in the modular norm work (Large et al., 2024; Bernstein & Newhouse, 2025). Maybe surprisingly, we will see that the optimal choice for the algorithm is not necessarily the matching Schatten-∞ geometry.

Assume that f is L_∞-smooth with respect to $\|\cdot\|_\infty$, and that the mini-batch oracle satisfies Assumption 5 with a fixed $r \in (1, 2]$. The oracle fixes the threshold $p_r = r/(r-1)$ for which the p-geometry method is admissible when $p \geq p_r$.

The ∞-smoothness assumption is inherited by every finite-p analysis geometry, since

$$\|\nabla f(x) - \nabla f(y)\|_{p/(p-1)} \leq \|\nabla f(x) - \nabla f(y)\|_1 \leq L_\infty \|x - y\|_\infty \leq L_\infty \|x - y\|_p .$$

Also $R_p := \|x^\star - z^0\|_p \leq d^{1/p} R_\infty$, which shows the dimension price of using a smaller p.

Deterministic setting and acceleration A tradeoff can already be seen in the deterministic setting. When the noise is negligible, we take $B = 1$ for which the oracle budget becomes equivalent to the number of iterations, $N = n$. The accelerated rate (Corollary 3.2) then gives

$$\epsilon_p^{\text{det}}(N) \lesssim \frac{L_\infty R_\infty^2 d^{2/p}}{N^{1+2/p}} = \frac{L_\infty R_\infty^2}{N} \left(\frac{d}{N}\right)^{2/p}.$$

Clearly, when $N \gtrsim d$, picking p small improves the rate through $(\frac{d}{N})^{2/p}$. On the other hand, when $N \lesssim d$, the same factor penalizes small p and large $p \to \infty$ is preferred.

Heavy-tailed noise and batching It turns out that a similar regime separation carries over to the stochastic case (*c.f.* Section D.3). We summarize the regime dependent optimal choice of p in Table 3.

[2]By *natural geometry* we specificaly mean that the product $L_p R_p^2$ is the smallest for the given p.

Table 3: Optimal choice of geometry under smoothness in $p = \infty$. Matching the problem geometry is favorable in the high-dimensional regime ($N \lesssim d$). However, in the low-dimensional regime ($N \gtrsim d$) choosing a smaller p is favorable both deterministically and in the presence of heavy-tailed noise. Here $A := L_\infty R_\infty^2$ and $S := \sigma_r R_\infty$.

Setting	**Regime**	**Preferred geometry**
Deterministic ($\sigma_q = 0$)	$N \gtrsim d$	$p = 2$
Deterministic ($\sigma_q = 0$)	$N \lesssim d$	$p \to \infty$
Heavy-tailed noise ($\sigma_r > 0$)	$N \gtrsim d \max\left\{1, \left(\frac{S}{A}\right)^{p_r}\right\}$	$p = p_r$
Heavy-tailed noise ($\sigma_r > 0$)	$N \lesssim d \max\left\{1, \left(\frac{S}{A}\right)^{p_r}\right\}$	$p \to \infty$

Limitations To compare the different methods we have used worst-case norm conversions both for the radius, from R_{p_r} to R_∞, and for the noise, from the algorithmic dual norm $\|\cdot\|_q$ to the assumed r-moment norm (1), both of which introduce a dimensionality dependency. Sharper problem-dependent relations can improve these constants and move the boundary.

4.1.1 Anytime switching under heavy-tailed noise

The careful reader might have noticed that the above regime discussion requires knowing the budget N in order to pick the optimal geometry p. What if we did not know N? The regime bound suggests switching from p close to ∞ to $p = p_r$ once $N_{\text{sw}} \asymp d \max\left\{1, \left(\frac{S}{A}\right)^{p_r}\right\}$.

When the final oracle budget is not known in advance, this gives the rule

$$p(N) = \begin{cases} p_d, & N < N_{\text{sw}}, \\ p_r, & N \geq N_{\text{sw}}, \end{cases} \qquad p_d = \max\{3, \lceil \log(ed) \rceil\}.$$

This is the optimization analogue of the adaptive-regularization principle in anytime online learning on ℓ_p balls (Johnson et al., 2025). Concretely, use the following restart procedure:

1. Run SODA with $p = p_d$ using the parameterization from Corollary C.3 until N_{sw} oracle calls, choosing the batch size by optimizing the complexity bound at the switching budget.
2. Restart with $p = p_r$ using the parameterization from Corollary D.1, keeping the same batch size.

Remarkably, Soft-Muon (Abrahamsen, 2026) seems to have arrived at such a geometry-schedule empirically by softening the ∞-geometry later in training.[3] In contrast, the proposal above is a *discrete* switching strategy motivated by the regime discussion. We leave further practical exploration of this idea to future work.

4.2 Implications for scaling laws

We now ask what can be extracted from the regime discussion above in the context of deep learning. A typical scaling-law prescription, such as Chinchilla (Hoffmann et al., 2022), is that the model size should scale proportionally to the number of training tokens/stochastic gradient oracle calls,

$$\texttt{model_size} \propto N.$$

For a matrix parameter in a depth-D network, however, the model size is

$$\texttt{model_size} \propto \texttt{fan_in} * \texttt{fan_out} * D.$$

If width and depth are scaled together, then the rank of each matrix grows only like

$$\texttt{rank} \propto (\texttt{model_size})^{1/3} \propto N^{1/3}.$$

For Schatten-p geometries, the rank is the effective dimension d in the regime discussion. The Chinchilla scaling law puts us well within the low-dimensional regime $N \gtrsim d$, where it is optimal to pick the smallest admissible geometry, Schatten-p_r, based on the heavy-tailed noise assumption (see Table 3).

[3] See the speedrun submission.

This may explain why the improvement for Schatten-∞ based methods such as Muon over baselines has been observed to reduce with scale (Wen et al., 2025; Semenov et al., 2025) and the recent interest in finite Schatten-p norm optimizers (Abrahamsen, 2026; Pang et al., 2026; Shumaylov et al., 2026). As promising future work we leave open the question of how to scale more appropriately for Schatten-∞ based optimizers.

5 Related work

The regime discussion is inspired by Johnson et al. (2025), who study FTRL in online learning and identify when the optimal geometry separates from the Euclidean choice $p = 2$. In comparison, we study an optimistic accelerated algorithm in the offline stochastic setting under heavy-tailed noise with moment r, and ask when the smallest admissible geometry p_r separates from the near-∞ geometry.

The deterministic accelerated rate is related to non-Euclidean acceleration and optimistic online-to-batch conversion. d'Aspremont et al. (2018) study accelerated methods under p-uniform convexity, showing the same p-dependent deterministic rate of $L_p R_p^2/n^{1+2/p}$ on ℓ_p balls, matching the lower bounds of Guzmán & Nemirovski (2018). Another route to acceleration was pioneered in the online-learning literature through online-to-offline and adaptive universality results (Levy, 2017; Levy et al., 2018), the constrained UnixGrad framework (Kavis et al., 2019), and the optimistic online-to-batch constructions of Cutkosky (2019); Joulani et al. (2020). This perspective was later used in the schedule-free framework of Defazio et al. (2024). Our analysis follows this latter route, but uses Optimistic Dual Averaging with a p-uniformly convex reference function, which changes the Euclidean accelerated term LR^2/n^2 into $L_p R_p^2/n^{1+2/p}$ and the noise term to be based on a heavy-tailed assumption.

Uniform convexity was used to analyze Online Mirror Descent in the context of online learning in Sridharan & Tewari (2010); Srebro et al. (2011). In stochastic convex optimization with infinite variance, Vural et al. (2022) prove optimal rates for stochastic mirror descent with uniformly convex reference functions under heavy-tailed stochastic gradients with bounded r-th moment, $r \in (1, 2]$. In Euclidean geometry, much of the theory handles such noise by modifying the gradients, e.g, Cutkosky & Mehta (2021) combine clipping, momentum, and normalized gradient descent and Gorbunov et al. (2020) use accelerated gradient clipping. More recent work studies when clipping can be avoided (Hübler et al., 2025; Liu & Zhou, 2025).

Muon can be viewed and analyzed as a conditional gradient method (Pethick et al., 2025; Mokhtari et al., 2020). Specifically with weight decay, Muon reduces to stochastic Frank-Wolfe for which Kovalev (2025) additionally provided analysis under star-convexity. This constrained endpoint is recovered by our template in the limit $p \to \infty$ with primal extrapolation ($\bar{\lambda}_k = 0$) and no optimism ($\bar{\alpha}_k = 0$). Muon can also be seen as an instance of the Lion-$\mathcal{K}$ framework, which allows capturing what we refer to as optimism (Chen et al., 2025; Sfyraki & Wang, 2025). Our analysis differs by focusing on finite p, where we show that optimism helps exploit smoothness and accelerate convergence. For Schatten-∞ geometries we approximate with $p = O(\log d)$, matching practice where $UV^\top$ is only computed approximately.

6 Conclusion

We have learned that the choice of p-norm to use in the algorithm depends on the training regime. For high-dimensional problems ($N \lesssim d$) the optimizer should match the smoothness assumption, while for low-dimensional problems ($N \gtrsim d$) the optimizer should instead match the noise assumption in the stochastic case and use the Euclidean geometry in the deterministic case to accelerate.

What is striking about our results is how closely they match empirical observations. Our batch size analysis provides an explanation for why Schatten-∞ based optimizers like Muon typically require larger batch sizes and gives a batch size scaling rule for any Schatten-p norm. We have also seen an explanation for why stepsize warmup is not needed for the Schatten-∞ norm, quantified when a smaller Schatten-p norm is beneficial and provided theoretical backing for existing empirical schedules of p.

AI USE STATEMENT

In this work, we used generative AI tools for extracting the batchsize conclusions from the complexity results, generating the stepsize and batchsize plots and generalizing the result to Hölder smoothness. It was also extensively used for rearranging and improving the presentation by streamlining notation. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Nilin Abrahamsen. Contra-Muon and Soft-Muon. Blog post, 2026. URL https://nilin.github.io/contra-muon-and-soft-muon/. Accessed: 2026-05-27.

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

Keith Ball, Eric A. Carlen, and Elliott H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones Mathematicae*, 115(1):463–482, 1994. doi: 10.1007/BF01231769.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. In *International Conference on Machine Learning*, 2025.

David Carlson, Ya-Ping Hsieh, Edo Collins, Lawrence Carin, and Volkan Cevher. Stochastic spectral descent for discrete graphical models. *IEEE Journal of Selected Topics in Signal Processing*, 2016.

David E Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned spectral descent for deep learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pp. 2971–2979, 2015.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025. URL https://arxiv.org/abs/2506.15054.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V Le. Symbolic discovery of optimization algorithms. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ne6zeqLFCZ.

Ashok Cutkosky. Anytime online-to-batch, optimism and acceleration. In *International conference on machine learning*, pp. 1446–1454. PMLR, 2019.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4883–4895, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/26901debb30ea03f0aa833c9de6b81e9-Abstract.html.

Alexandre d'Aspremont, Cristóbal Andrés Guzmán Paredes, and Martin Jaggi. Optimal affine-invariant smooth minimization algorithms. *SIAM Journal on Optimization*, 28(3):2384–2405, 2018. doi: 10.1137/17M1116842.

Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled. In *Advances in Neural Information Processing Systems*, volume 37, pp. 9974–10007, 2024. doi: 10.52202/079017-0320. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/136b9a13861308c8948cd308ccd02658-Paper-Conference.pdf.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15042–15053, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/abd1c782880cc59759f4112fda0b8f98-Abstract.html.

Cristóbal Guzmán and Arkadi Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *arXiv preprint arXiv:1307.5001*, 2018. URL https://arxiv.org/abs/1307.5001.

Stephen Hanson and Lorien Pratt. Comparing biases for minimal network construction with back-propagation. *Advances in neural information processing systems*, 1, 1988.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, 2022.

Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed SGD. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2413–2421. PMLR, 2025. URL https://proceedings.mlr.press/v258/hubler25a.html.

Rustem Islamov, Roman Machacek, Aurelien Lucchi, Antonio Silveti-Falls, Eduard Gorbunov, and Volkan Cevher. On the role of batch size in stochastic conditional gradient methods. *arXiv preprint arXiv:2603.21191*, 2026.

Emmeran Johnson, David Martínez-Rubio, Ciara Pike-Burke, and Patrick Rebeschini. On the necessity of adaptive regularisation: Optimal anytime online learning on ℓ_p-balls. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=ONc9vWkwCp. Spotlight.

Keller Jordan. 94% on CIFAR-10 in 3.29 seconds on a single GPU. *arXiv preprint arXiv:2404.00498*, 2024.

Keller Jordan, Jeremy Bernstein, Brendan Rappazzo, @fernbear.bsky.social, Boza Vlado, You Jiacheng, Franz Cesista, Braden Koszarsky, and @Grad62304977. modded-nanogpt: Speedrunning the nanogpt baseline, 2024a. URL https://github.com/KellerJordan/modded-nanogpt.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cecista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024b.

Pooria Joulani, Anant Raj, Andras Gyorgy, and Csaba Szepesvári. A simpler approach to accelerated optimization: iterative averaging meets optimism. In *International conference on machine learning*, pp. 4984–4993. PMLR, 2020.

Ali Kavis, Kfir Yehuda Levy, Francis Bach, and Volkan Cevher. UnixGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025. URL https://arxiv.org/abs/2503.12645.

Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

Tim Large, Yang Liu, Minyoung Huh, Hyojin Bahng, Phillip Isola, and Jeremy Bernstein. Scalable optimization in the modular norm. In *Advances in Neural Information Processing Systems*, 2024.

Kfir Yehuda Levy. Online to offline conversions, universality and adaptive minibatch sizes. In *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

Kfir Yehuda Levy, Alp Yurtsever, and Volkan Cevher. Online adaptive methods, universality and acceleration. In *NeurIPS '18: Proceedings of the 32nd International Conference of Neural Information Processing Systems*, 2018.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.19529.

Martin Marek, Sanae Lotfi, Aditya Somasundaram, Andrew Wilson, and Micah Goldblum. Small batch size training for language models: When vanilla sgd works, and why gradient accumulation is wasteful. *Advances in Neural Information Processing Systems*, 38:148837–148862, 2026.

Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *Journal of Machine Learning Research*, 21 (105):1–49, 2020.

Yuji Nakatsukasa and Roland W Freund. Computing fundamental matrix decompositions accurately via the matrix sign function in two iterations: The power of zolotarev's functions. *siam REVIEW*, 58(3):461–493, 2016.

Yuji Nakatsukasa, Zhaojun Bai, and François Gygi. Optimizing halley's iteration for computing the matrix polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, 31(5):2700–2720, 2010.

Yu Nesterov and Vladimir Shikhman. Quasi-monotone subgradient methods for nonsmooth convex minimization. *Journal of Optimization Theory and Applications*, 165(3):917–940, 2015.

Francesco Orabona. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2019. URL http://arxiv.org/abs/1912.13213.

Tianyu Pang, Yujie Fang, Zihang Liu, Shenyang Deng, Lei Hsiung, Shuhua Yu, and Yaoqing Yang. HTmuon: Improving muon via heavy-tailed spectral correction. *arXiv preprint arXiv:2603.10067*, 2026.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *International Conference on Machine Learning*, 2025.

Thomas Pethick, Wanyun Xie, Roman Machacek, and Volkan Cevher. Optimistic dual averaging unifies modern optimizers. *arXiv preprint arXiv:2605.11172*, 2026.

I.F. Pinelis and A.I. Sakhanenko. Remarks on inequalities for large deviation probabilities. *Theory of Probability and its Applications*, 30:143, 1986.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pp. 993–1019. PMLR, 2013.

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

Maria-Eleni Sfyraki and Jun-Kun Wang. Lions and muons: Optimization via stochastic frank-wolfe. *arXiv preprint arXiv:2506.04192*, 2025. URL https://arxiv.org/abs/2506.04192.

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

Zakhar Shumaylov, Nathaël Da Costa, Peter Zaika, Bálint Mucsányi, Alex Massucco, Yoav Gelberg, Carola-Bibiane Schönlieb, Yarin Gal, and Philipp Hennig. Muon is not that special: Random or inverted spectra work just as well. *arXiv preprint arXiv:2605.11181*, 2026.

Nati Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2645–2653, 2011.

Teodora Srećković, Jonas Geiping, and Antonio Orvieto. Is your batch size the problem? revisiting the adam-sgd gap in language modeling. *arXiv preprint arXiv:2506.12543*, 2025.

Karthik Sridharan and Ambuj Tewari. Convex games in banach spaces. In *Proceedings of the 23rd Conference on Learning Theory*, pp. 1–13, 2010. URL https://dblp.org/rec/conf/colt/SridharanT10.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.

Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. Powersgd: Practical low-rank gradient compression for distributed optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303, 1965.

Nuri Mert Vural, Lu Yu, Krishna Balasubramanian, Stanislav Volgushev, and Murat A. Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 65–102. PMLR, 2022. URL https://proceedings.mlr.press/v178/vural22a.html.

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

APPENDIX CONTENTS

A Preliminaries **14**

B Technical lemmas **15**

C Convergence rate **17**

C.1 Non-accelerated 17

C.2 Accelerated 20

C.3 Near $p = \infty$ 22

D Oracle complexity **23**

D.1 Accelerated 23

D.2 Accelerated batch size rule for a fixed geometry 23

D.3 Heavy-tailed noise and batching 23

A PRELIMINARIES

We use the following notation. For a norm $\|\cdot\|$ on the primal space, the dual norm is

$$\|g\|_* := \sup_{\|x\|\le 1} \langle g, x\rangle .$$

For a proper function $h : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$, its domain is

$$\operatorname{dom} h := \{x \in \mathcal{X} : h(x) < \infty\}.$$

Constraints are supported by taking the reference function to be $h(x) = \psi(x) + \iota_{\mathcal{C}}(x)$, where $\iota_{\mathcal{C}}$ is the indicator of the constraint set, equal to 0 on $\mathcal{C}$ and $+\infty$ otherwise. Then $\operatorname{dom} h = \mathcal{C} \cap \operatorname{dom} \psi$, and the dual map can be written as

$$\partial h^*(-\gamma m) = \arg\min_{x\in\mathcal{X}}\{\gamma \langle m, x\rangle + h(x)\} = \arg\min_{x\in\mathcal{C}}\{\gamma \langle m, x\rangle + \psi(x)\}.$$

For a matrix X, let $\sigma_i(X)$ denote its singular values and define the Schatten-p norm by

$$\|X\|_{S_p} := \left(\textstyle\sum_i \sigma_i(X)^p\right)^{1/p}, \qquad \|X\|_{S_\infty} := \max_i \sigma_i(X).$$

The dual of $\|\cdot\|_{S_p}$ is $\|\cdot\|_{S_q}$ when $q = p/(p-1)$.

Additional p-uniformly convex norms. The same framework is not restricted to vector ℓ_p and Schatten-p norms (see Ball et al. (1994) for Schatten-p norms). For a matrix X with rows X_i, define the mixed row norm

$$\|X\|_{2,p} := \left(\textstyle\sum_i \|X_i\|_2^p\right)^{1/p}, \qquad \|X\|_{2,\infty} := \max_i \|X_i\|_2 .$$

The endpoint $\|\cdot\|_{2,\infty}$ is the operator norm $\|\cdot\|_{2\to\infty}$. For $p > 2$, $h(X) = \frac{1}{p}\left\|X - Z^0\right\|_{2,p}^p$ is p-uniformly convex with respect to $\|\cdot\|_{2,p}$ with admissible constant $\mu = 2^{2-p}$, since it is an ℓ_p sum of row norms. Thus the results below also apply to the mixed-norm geometry after replacing $\|\cdot\|_p$ and its dual by $\|\cdot\|_{2,p}$ and $\|\cdot\|_{2,q}$.

For positive quantities A and B, $A \lesssim_\Theta B$ means $A \le C(\Theta)B$ for a constant depending only on the parameters Θ, and $A \gtrsim_\Theta B$ means $B \lesssim_\Theta A$. We write $A \asymp_\Theta B$ when both $A \lesssim_\Theta B$ and $A \gtrsim_\Theta B$ hold, and $A = O_\Theta(B)$ with the same meaning as $A \lesssim_\Theta B$.

B TECHNICAL LEMMAS

Our proof builds on Defazio et al. (2024). However, rather than combining primal extrapolation with an adaptive version of Optimistic Mirror Descent, we use Optimistic Dual Averaging (Rakhlin & Sridharan, 2013) as the underlying no-regret algorithm. The ODA analysis is based on a non-adaptive version of Rakhlin & Sridharan (2013), which is classical (see, e.g., Orabona (2019, Sec. 7.12)). We furthermore generalize the result to uniform convex reference functions, Hölder smoothness and a heavy-tailed type noise assumption, which are essential for our regime discussion.

Assumption 6 ((L_ν, ν)-Hölder smooth)**.** *For some $\nu \in (0, 1]$, the function f is differentiable and satisfies*

$$\|\nabla f(u) - \nabla f(v)\|_* \le L_\nu \|u - v\|^\nu \qquad \text{for all } u, v \in \mathcal{X}.$$

When specializing Assumption 6 to $\|\cdot\|_p$, we write $L_{\nu,p}$ for the corresponding Hölder smoothness constant and use the shorthand $L_p := L_{1,p}$ in the smooth case.

Lemma B.1 (ODA regret with a uniformly convex reference)**.** *Consider SODA, with the gradient term $\nabla f(y^k, \xi_k)$ replaced by g^k. Choose positive weights $a_0, \dots, a_{n-1}$ and write*

$$A_k := \textstyle\sum_{i=0}^k a_i, \qquad A_{-1} := 0.$$

We use the convention $g^{-1} = 0$ and set $z^0 \in \partial h^*(0)$. For $k = 0, \dots, n-2$, choose

$$\alpha_k = \tfrac{a_k}{A_k}, \quad \bar\alpha_k = \tfrac{a_{k+1}}{A_{k+1}}, \quad \gamma_k = \eta \tfrac{A_k}{1-\bar\alpha_k}$$

for some $\eta > 0$. Let $p \ge 2$, let $q = p/(p-1)$, and let h be p-uniformly convex with constant μ in the sense of Definition 1. Then, for every $x \in \mathcal{X}$,

$$\textstyle\sum_{k=0}^{n-1} a_k \langle g^k, z^k - x\rangle \le \frac{h(x) - \inf_{u\in\mathcal{X}} h(u)}{\eta} + \frac{\eta^{q-1}}{q\mu^{q-1}} \sum_{k=0}^{n-1} a_k^q \left\|g^k - g^{k-1}\right\|_*^q. \tag{3}$$

Proof. By the choice of γ_k and $\bar\alpha_k$ we have, for $k = 0, \dots, n-2$,

$$\gamma_k(1 - \bar\alpha_k) = \eta A_k, \quad \text{and} \quad \gamma_k \bar\alpha_k = \eta a_{k+1}.$$

Combined with the choice of α_k, the dual update can be written as

$$z^k \in \partial h^*\left(-\eta \textstyle\sum_{i=0}^{k-1} a_i g^i - \eta a_k g^{k-1}\right), \qquad k = 0, \dots, n-1,$$

where the sum is empty when $k = 0$ and the case $k = 0$ is exactly $z^0 \in \partial h^*(0)$ because $g^{-1} = 0$. Set

$$\theta^k := -\eta \textstyle\sum_{i=0}^k a_i g^i, \qquad \theta^{-1} := 0, \qquad \hat\theta^k := \theta^{k-1} - \eta a_k g^{k-1}.$$

Thus $z^k \in \partial h^*(\hat\theta^k)$ and

$$\theta^k = \hat\theta^k - \eta a_k (g^k - g^{k-1}).$$

Since h is p-uniformly convex, h^* is q-uniformly smooth (Vural et al., 2022, Prop. 1):

$$h^*(\theta + \Delta) \le h^*(\theta) + \langle \nabla h^*(\theta), \Delta\rangle + \tfrac{1}{q\mu^{q-1}} \|\Delta\|_*^q.$$

Hence

$$h^*(\theta^k) \le h^*(\hat\theta^k) - \eta a_k \langle g^k - g^{k-1}, z^k\rangle + \tfrac{\eta^q a_k^q}{q\mu^{q-1}} \left\|g^k - g^{k-1}\right\|_*^q.$$

By Fenchel–Young,

$$\langle \hat\theta^k, z^k\rangle = h(z^k) + h^*(\hat\theta^k), \qquad \langle \theta^k, x\rangle \le h(x) + h^*(\theta^k).$$

Therefore

$$\begin{aligned}
\eta a_k \langle g^k, z^k - x\rangle &= \eta a_k \langle g^k - g^{k-1}, z^k\rangle + \eta a_k \langle g^{k-1}, z^k\rangle - \eta a_k \langle g^k, x\rangle \\
&= \eta a_k \langle g^k - g^{k-1}, z^k\rangle + \langle \theta^{k-1} - \hat\theta^k, z^k\rangle + \langle \theta^k - \theta^{k-1}, x\rangle \\
&= \eta a_k \langle g^k - g^{k-1}, z^k\rangle + \langle \theta^{k-1}, z^k\rangle - \langle \hat\theta^k, z^k\rangle + \langle \theta^k - \theta^{k-1}, x\rangle \\
&= \eta a_k \langle g^k - g^{k-1}, z^k\rangle + \langle \theta^{k-1}, z^k\rangle - h(z^k) - h^*(\hat\theta^k) + \langle \theta^k - \theta^{k-1}, x\rangle \\
&\le h^*(\theta^{k-1}) - h^*(\theta^k) + \tfrac{\eta^q a_k^q}{q\mu^{q-1}} \left\|g^k - g^{k-1}\right\|_*^q + \langle \theta^k - \theta^{k-1}, x\rangle.
\end{aligned}$$

Summing over $k = 0, \dots, n-1$ yields

$$\eta \textstyle\sum_{k=0}^{n-1} a_k \langle g^k, z^k - x\rangle \le h^*(0) - h^*(\theta^{n-1}) + \langle \theta^{n-1}, x\rangle + \frac{\eta^q}{q\mu^{q-1}} \sum_{k=0}^{n-1} a_k^q \left\|g^k - g^{k-1}\right\|_*^q.$$

Finally, $h^*(0) = -\inf_{u\in\mathcal{X}} h(u)$ and $\langle \theta^{n-1}, x\rangle - h^*(\theta^{n-1}) \le h(x)$, so dividing by η proves the claim. □

Lemma B.2 (Hölder co-coercivity)**.** *Suppose Assumptions 1 and 6 hold. For differentiable f, write*

$$D_f(u,v) := f(u) - f(v) - \langle \nabla f(v), u - v\rangle .$$

Then, for all $u, v \in \mathcal{X}$,

$$D_f(u,v) \geq \tfrac{\nu}{1+\nu} L_\nu^{-1/\nu} \left\|\nabla f(u) - \nabla f(v)\right\|_*^{(1+\nu)/\nu} .$$

Proof. Assumption 6 implies, by integration, that f is $(L_\nu, 1+\nu)$-uniformly smooth:

$$f(u) \leq f(v) + \langle \nabla f(v), u - v\rangle + \tfrac{L_\nu}{1+\nu} \left\|u - v\right\|^{1+\nu} ,$$

and hence, by uniform smoothness/convexity duality (Vural et al., 2022, Prop. 1),

$$D_{f^*}(s,t;\zeta) \geq \tfrac{\nu}{1+\nu} L_\nu^{-1/\nu} \left\|s - t\right\|_*^{(1+\nu)/\nu} \qquad \text{for all } s, t \text{ and } \zeta \in \partial f^*(t). \tag{4}$$

Since $u \in \partial f^*(\nabla f(u))$,

$$\begin{aligned} D_f(u,v) &\overset{\text{(a)}}{=} D_{f^*}(\nabla f(v), \nabla f(u); u) \\ &\overset{\text{(b)}}{\geq} \tfrac{\nu}{1+\nu} L_\nu^{-1/\nu} \left\|\nabla f(u) - \nabla f(v)\right\|_*^{(1+\nu)/\nu} . \end{aligned}$$

Here (a) is Bregman duality and (b) applies (4) with $(s,t,\zeta) = (\nabla f(v), \nabla f(u), u)$. □

Lemma B.3 (Online-to-batch Hölder refinement)**.** *Let an online no-regret algorithm generate $z^k = \mathcal{A}_k(g^0, \dots, g^{k-1}) \in \mathcal{X}$ before receiving the feedback $g^k = \nabla f(y^k, \xi_k) \in \mathcal{X}^*$. Define, for $k = 1, \dots, n-1$, the wrapper*

$$\begin{aligned} x^k &= (1 - \lambda_{k-1}) x^{k-1} + \lambda_{k-1} z^k, \\ y^k &= (1 - \bar{\lambda}_{k-1}) x^k + \bar{\lambda}_{k-1} z^k, \end{aligned}$$

with initialization $x^0 = y^0 = z^0$. Given positive weights $a_0, \dots, a_{n-1}$, choose

$$\lambda_{k-1} = \tfrac{a_k}{A_k}, \qquad 0 \leq \bar{\lambda}_{k-1} \leq c_\nu \lambda_{k-1}, \qquad A_k = \textstyle\sum_{i=0}^k a_i, \qquad A_{-1} = 0, \qquad k = 1, \dots, n-1.$$

Suppose Assumptions 1, 3 and 6 hold. Let

$$r_\nu := \tfrac{1+\nu}{\nu}.$$

Then there exists a constant $c_\nu > 0$, depending only on ν, with $c_1 = 1/6$, such that, for every $x \in \mathcal{X}$, the wrapped sequence satisfies

$$\begin{aligned} A_{n-1}\, \mathbb{E}[f(x^{n-1}) - f(x)] \leq \mathbb{E}\left[\textstyle\sum_{k=0}^{n-1} a_k \langle g^k, z^k - x\rangle\right] \\ - \tfrac{c_\nu}{L_\nu^{1/\nu}} \textstyle\sum_{k=1}^{n-1} A_{k-1} \mathbb{E}\left[\left\|\nabla f(y^k) - \nabla f(y^{k-1})\right\|_*^{r_\nu}\right] . \end{aligned} \tag{5}$$

Proof. For differentiable f, write

$$D_f(u,v) := f(u) - f(v) - \langle \nabla f(v), u - v\rangle .$$

For $k = 0, \dots, n-1$, following Defazio et al. (2024, Thm. 5), a careful expansion gives

$$\begin{aligned} A_k\big(f(x^k) - f(x)\big) - A_{k-1}\big(f(x^{k-1}) - f(x)\big) = a_k \langle \nabla f(y^k), z^k - x\rangle \\ - \tfrac{a_k}{\bar{\lambda}_{k-1}} D_f(y^k, x^k) - \tfrac{a_k(1-\bar{\lambda}_{k-1})}{\bar{\lambda}_{k-1}} D_f(x^k, y^k) \\ - A_{k-1} D_f(x^{k-1}, x^k) - a_k D_f(x, y^k), \end{aligned}$$

with the convention that the terms involving A_{-1} vanish, and that the two terms involving $\bar{\lambda}_{k-1}^{-1}$ vanish when $k = 0$ or $\bar{\lambda}_{k-1} = 0$, since then $y^k = x^k$.

By construction, z^k and hence x^k, y^k are $\mathcal{F}_{k-1}$-measurable. Therefore, using the conditional unbiasedness in Assumption 3,

$$\mathbb{E}\big[\langle g^k, z^k - x\rangle \mid \mathcal{F}_{k-1}\big] = \langle \nabla f(y^k), z^k - x\rangle .$$

Summing over $k = 0, \dots, n-1$ and taking expectations gives

$$\begin{aligned} A_{n-1}\mathbb{E}[f(x^{n-1}) - f(x)] \le{}& \mathbb{E}\left[\textstyle\sum_{k=0}^{n-1} a_k \langle g^k, z^k - x\rangle\right] \\ &- \textstyle\sum_{k=0}^{n-1}\left(\frac{a_k}{\bar\lambda_{k-1}}\mathbb{E}D_f(y^k, x^k) + \frac{a_k(1-\bar\lambda_{k-1})}{\bar\lambda_{k-1}}\mathbb{E}D_f(x^k, y^k)\right) \\ &- \textstyle\sum_{k=0}^{n-1}\left(A_{k-1}\mathbb{E}D_f(x^{k-1}, x^k) + a_k\mathbb{E}D_f(x, y^k)\right). \end{aligned} \tag{6}$$

Assumptions 1 and 6 imply the generalized co-coercivity inequality (Lemma B.2)

$$D_f(u, v) \ge \tfrac{\nu}{1+\nu} L_\nu^{-1/\nu} \left\|\nabla f(u) - \nabla f(v)\right\|_*^{(1+\nu)/\nu}.$$

For $k \ge 1$, let

$$\begin{aligned} U_k &:= \left\|\nabla f(y^k) - \nabla f(x^k)\right\|_*^{r_\nu}, \\ V_k &:= \left\|\nabla f(x^k) - \nabla f(x^{k-1})\right\|_*^{r_\nu}, \\ W_k &:= \left\|\nabla f(y^k) - \nabla f(y^{k-1})\right\|_*^{r_\nu}. \end{aligned}$$

The relation $x^0 = y^0$ gives $U_0 = 0$. Also,

$$\begin{aligned} W_k &= \left\|\nabla f(y^k) - \nabla f(x^k) + \nabla f(x^k) - \nabla f(x^{k-1}) + \nabla f(x^{k-1}) - \nabla f(y^{k-1})\right\|_*^{r_\nu} \\ &\le K_\nu(U_k + V_k + U_{k-1}), \end{aligned} \tag{7}$$

where $K_\nu = 3^{r_\nu - 1}$. Moreover, $\bar\lambda_{k-1} \le c_\nu a_k / A_k$ implies $a_k/\bar\lambda_{k-1} \ge A_k/c_\nu$. Let

$$b_\nu := \tfrac{\nu}{1+\nu}, \qquad \beta_\nu := \tfrac{b_\nu}{L_\nu^{1/\nu}}.$$

Choose $c_\nu > 0$ sufficiently small, depending only on ν, and set $C_\nu := c_\nu K_\nu$. Then, for $k \ge 1$,

$$\begin{aligned} &-\tfrac{a_k}{\bar\lambda_{k-1}} D_f(y^k, x^k) - A_{k-1} D_f(x^{k-1}, x^k) \\ &\quad\text{(a)}\ \le -\tfrac{\beta_\nu A_k}{c_\nu} U_k - \beta_\nu A_{k-1} V_k \\ &\quad\text{(b)}\ \le -\tfrac{2C_\nu A_k}{L_\nu^{1/\nu}} U_k - \tfrac{C_\nu A_{k-1}}{L_\nu^{1/\nu}} V_k \\ &\quad\text{(c)}\ \le -\tfrac{C_\nu A_{k-1}}{L_\nu^{1/\nu}}(U_k + V_k + U_{k-1}) - \tfrac{C_\nu A_k}{L_\nu^{1/\nu}} U_k + \tfrac{C_\nu A_{k-1}}{L_\nu^{1/\nu}} U_{k-1} \\ &\quad\text{(d)}\ \le -\tfrac{c_\nu A_{k-1}}{L_\nu^{1/\nu}} W_k - \tfrac{C_\nu A_k}{L_\nu^{1/\nu}} U_k + \tfrac{C_\nu A_{k-1}}{L_\nu^{1/\nu}} U_{k-1}. \end{aligned} \tag{8}$$

Here (a) uses the co-coercivity lower bound and $a_k/\bar\lambda_{k-1} \ge A_k/c_\nu$; (b) uses the choice of c_ν satisfying $K_\nu c_\nu \le b_\nu$ and $2K_\nu c_\nu^2 \le b_\nu$; (c) uses $A_{k-1} \le A_k$; and (d) uses (7) and $C_\nu = c_\nu K_\nu$. Adding back the additional nonpositive Bregman terms therefore gives, for each $k \ge 1$,

$$\begin{aligned} &-\tfrac{a_k}{\bar\lambda_{k-1}} D_f(y^k, x^k) - \tfrac{a_k(1-\bar\lambda_{k-1})}{\bar\lambda_{k-1}} D_f(x^k, y^k) - A_{k-1} D_f(x^{k-1}, x^k) - a_k D_f(x, y^k) \\ &\quad\le -\tfrac{c_\nu A_{k-1}}{L_\nu^{1/\nu}} W_k - \tfrac{C_\nu A_k}{L_\nu^{1/\nu}} U_k + \tfrac{C_\nu A_{k-1}}{L_\nu^{1/\nu}} U_{k-1}, \end{aligned}$$

Summing over $k = 1, \dots, n-1$ telescopes the last two terms:

$$-\textstyle\sum_{k=1}^{n-1} A_k U_k + \sum_{k=1}^{n-1} A_{k-1} U_{k-1} = -A_{n-1} U_{n-1} \le 0,$$

because $U_0 = 0$. Dropping this nonpositive remainder and combining with (6) gives (5). For $\nu = 1$, we have $K_1 = 3$ and $b_1 = 1/2$, so the choice $c_1 = 1/6$ satisfies $K_1 c_1 \le b_1$ and $2K_1 c_1^2 \le b_1$. □

C Convergence Rate

C.1 Non-accelerated

Theorem C.1 (Non-accelerated parameterization under Hölder smoothness)**.** *Let $p > 2$, let $q = p/(p-1)$, and let $\nu \in (0, 1]$. Let $x^\star \in \arg\min_{x \in \operatorname{dom} h} f(x)$ and $D_\star := h(x^\star) - \inf h$. Consider SODA. For every $k = 0, \dots, n-1$, choose*

$$\alpha_k = \tfrac{1}{k+1}, \qquad \bar\alpha_k = \lambda_k = \tfrac{1}{k+2}, \qquad \bar\lambda_k \le c_\nu \lambda_k, \qquad \gamma_k = \eta(k+2),$$

where $c_\nu > 0$ is the constant from Lemma B.3. Suppose Assumptions 1, 3, 4 and 6 hold with $q = p/(p-1)$. Assume also that h is p-uniformly convex with constant μ in the sense of Definition 1. Define

$$\kappa_{\mathrm{na}} := \min\left\{1, \tfrac{1+(p+1)\nu}{p}\right\}.$$

Set

$$\ell_n := \begin{cases} \log(\mathrm{e}n), & \nu = (p-1)/(p+1), \\ 1, & \text{otherwise}. \end{cases}$$

Choose

$$\eta \asymp_{p,\nu} \min\left\{ \tfrac{\mu^{(1+\nu)/p} D_\star^{(p-1-\nu)/p}}{\tau_q^{1+\nu} L_\nu n^{1-\kappa_{\mathrm{na}}} \ell_n^{(p-1-\nu)/p}}, \tfrac{\mu^{1/p} D_\star^{(p-1)/p}}{\sigma_q\, n^{(p-1)/p}} \right\}.$$

Then, for every $n \geq 1$,

$$\mathbb{E}[f(x^{n-1}) - f(x^\star)] = O_{p,\nu}\left(\tfrac{\tau_q^{1+\nu} L_\nu D_\star^{(1+\nu)/p}}{\mu^{(1+\nu)/p} n^{\kappa_{\mathrm{na}}}} + \tfrac{\sigma_q D_\star^{1/p}}{\mu^{1/p} n^{1/p}} \right),$$

up to an additional factor $(\log(\mathrm{e}n))^{(p-1-\nu)/p}$ when $\nu = (p-1)/(p+1)$. In particular, for $h(x) = \frac{1}{p}\left\|x - z^0\right\|_p^p$ and $R_p := \left\|x^\star - z^0\right\|_p$, taking $\mu = 2^{2-p}$ gives

$$\mathbb{E}[f(x^{n-1}) - f(x^\star)] = O_{p,\nu}\left(\tfrac{\tau_q^{1+\nu} L_{\nu,p} R_p^{1+\nu}}{n^{\kappa_{\mathrm{na}}}} + \tfrac{\sigma_q R_p}{n^{1/p}} \right),$$

with the same boundary logarithm.

Proof. Choose the weights $a_k \equiv 1$, so that $A_k = k+1$. Combining Lemma B.1 with Lemma B.3 gives

$$\begin{aligned} n\,\mathbb{E}[f(x^{n-1}) - f(x^\star)] \leq{}& \tfrac{D_\star}{\eta} + \tfrac{\eta^{q-1}}{q\mu^{q-1}} \textstyle\sum_{k=0}^{n-1} \mathbb{E}\left\|g^k - g^{k-1}\right\|_*^q \\ & - \tfrac{c_\nu}{L_\nu^{1/\nu}} \textstyle\sum_{k=1}^{n-1} k\,\mathbb{E}\left[\left\|\nabla f(y^k) - \nabla f(y^{k-1})\right\|_*^{r_\nu}\right]. \end{aligned}$$

We next split this last sum into stochastic and deterministic-gradient variation terms. For $k \geq 1$, raising Assumption 4 to the qth power and using $(a+b)^q \lesssim_q a^q + b^q$ gives

$$\mathbb{E}\left\|g^k - g^{k-1}\right\|_*^q \lesssim_q \tau_q^q \mathbb{E}\left\|\nabla f(y^k) - \nabla f(y^{k-1})\right\|_*^q + \sigma_q^q.$$

For $k = 0$, using the boundary conventions $g^{-1} = 0$ and $y^{-1} = x^\star$, Assumption 4 gives the bound on the startup term

$$\mathbb{E}\left\|g^0 - g^{-1}\right\|_*^q \lesssim_q \tau_q^q \mathbb{E}\left\|\nabla f(y^0) - \nabla f(x^\star)\right\|_*^q + \sigma_q^q.$$

Absorbing only q-dependent constants, we obtain

$$\begin{aligned} n\,\mathbb{E}[f(x^{n-1}) - f(x^\star)] \lesssim_q{}& \tfrac{D_\star}{\eta} + \tfrac{\eta^{q-1}\sigma_q^q n}{q\mu^{q-1}} \\ & + \tfrac{\tau_q^q \eta^{q-1}}{q\mu^{q-1}} \textstyle\sum_{k=1}^{n-1} \mathbb{E}\left[\left\|\nabla f(y^k) - \nabla f(y^{k-1})\right\|_*^q\right] \\ & - \tfrac{c_\nu}{L_\nu^{1/\nu}} \textstyle\sum_{k=1}^{n-1} k\,\mathbb{E}\left[\left\|\nabla f(y^k) - \nabla f(y^{k-1})\right\|_*^{r_\nu}\right] \\ & + \tfrac{\tau_q^q \eta^{q-1}}{q\mu^{q-1}} \mathbb{E}\left[\left\|\nabla f(y^0) - \nabla f(x^\star)\right\|_*^q\right]. \end{aligned} \tag{9}$$

The last term of (9) is dominated by the first term under the stated choice of η, since uniform convexity and choice of initialization gives $\left\|y^0 - x^\star\right\| = O_p((D_\star/\mu)^{1/p})$, and Assumption 6 gives

$$\left\|\nabla f(y^0) - \nabla f(x^\star)\right\|_*^q \leq L_\nu^q \left\|y^0 - x^\star\right\|^{\nu q} = O_p\left(L_\nu^q (D_\star/\mu)^{\nu q/p}\right).$$

Hence the last term of (9) is at most

$$O_p\left(\tfrac{\tau_q^q \eta^{q-1} L_\nu^q D_\star^{\nu q/p}}{\mu^{q-1+\nu q/p}} \right).$$

To compare with the first term $D_\star/\eta$, it is enough that

$$\tfrac{\tau_q^q \eta^{q-1} L_\nu^q D_\star^{\nu q/p}}{\mu^{q-1+\nu q/p}} \lesssim_p \tfrac{D_\star}{\eta} \quad\Longleftrightarrow\quad \eta^q \lesssim_p \tfrac{\mu^{q-1+\nu q/p} D_\star^{1-\nu q/p}}{\tau_q^q L_\nu^q}.$$

Taking qth roots, and using $q = p/(p-1)$, this becomes

$$\eta \lesssim_p \frac{\mu^{(1+\nu)/p} D_\star^{(p-1-\nu)/p}}{\tau_q L_\nu},$$

which is enforced by the deterministic part of the chosen stepsize since $\tau_q \geq 1$.

To deal with the two remaining terms of (9), for $k \geq 1$,

$$\sup_{s \geq 0} \{A s^q - B s^{r_\nu}\} \leq C_{p,\nu} A^{r_\nu/(r_\nu - q)} B^{-q/(r_\nu - q)} \qquad \text{for all } A, B > 0,$$

where $C_{p,\nu}$ is a finite constant depending only on p and ν. This follows by optimizing over s. Since $r_\nu = (1+\nu)/\nu > q = p/(p-1)$, the supremum is finite and the nonzero maximizer satisfies $s^{r_\nu - q} = qA/(r_\nu B)$. Applying this bound with

$$A = \frac{\tau_q^q \eta^{q-1}}{\mu^{q-1}}, \qquad B = \frac{k}{L_\nu^{1/\nu}},$$

yields

$$\sup_{s \geq 0} \left\{ \frac{\tau_q^q \eta^{q-1}}{\mu^{q-1}} s^q - \frac{k}{L_\nu^{1/\nu}} s^{r_\nu} \right\} \leq C_{p,\nu} \frac{\tau_q^{p(1+\nu)/(p-1-\nu)} L_\nu^{p/(p-1-\nu)}}{\mu^{(1+\nu)/(p-1-\nu)}} \eta^{(1+\nu)/(p-1-\nu)} k^{-p\nu/(p-1-\nu)}. \quad (10)$$

The exponents in this expression come from

$$r_\nu - q = \frac{p-1-\nu}{\nu(p-1)}, \qquad \frac{r_\nu}{r_\nu - q} = \frac{(1+\nu)(p-1)}{p-1-\nu}, \qquad \frac{q}{r_\nu - q} = \frac{p\nu}{p-1-\nu}.$$

To keep the summed bound readable, define

$$\theta_\nu := \frac{1+\nu}{p-1-\nu}, \qquad \zeta_\nu := \frac{p\nu}{p-1-\nu}, \qquad \mathfrak{S}_{p,\nu,n} := 1 + \textstyle\sum_{k=1}^{n-1} k^{-\zeta_\nu}, \qquad M_\nu := \frac{L_\nu^{p/(p-1-\nu)}}{\mu^{(1+\nu)/(p-1-\nu)}},$$

With this notation, (10) applies pointwise with

$$s_k := \left\| \nabla f(y^k) - \nabla f(y^{k-1}) \right\|_*.$$

Combining (10) with the corresponding positive and negative variation terms in (9) gives, for each $k \geq 1$,

$$\frac{\tau_q^q \eta^{q-1}}{q \mu^{q-1}} \mathbb{E}[s_k^q] - \frac{c_\nu}{L_\nu^{1/\nu}} k \, \mathbb{E}[s_k^{r_\nu}] \leq O_{p,\nu}\left(\tau_q^{p(1+\nu)/(p-1-\nu)} M_\nu \eta^{\theta_\nu} k^{-\zeta_\nu} \right).$$

Thus the positive and negative variation sums in (9) contribute at most the sum of these remainders. Summing over k and dividing by n gives

$$\mathbb{E}[f(x^{n-1}) - f(x^\star)] \leq O_{p,\nu}\left(\frac{D_\star}{\eta n} + \frac{\eta^{q-1} \sigma_q^q}{\mu^{q-1}} + \tau_q^{p(1+\nu)/(p-1-\nu)} M_\nu \eta^{\theta_\nu} \frac{1}{n} \mathfrak{S}_{p,\nu,n} \right). \quad (11)$$

Equating the deterministic and Hölder-variation terms in (11), and multiplying by n, gives

$$\frac{D_\star}{\eta} \asymp_{p,\nu} \tau_q^{p(1+\nu)/(p-1-\nu)} M_\nu \eta^{\theta_\nu} \mathfrak{S}_{p,\nu,n} \quad \Longleftrightarrow \quad \eta^{1+\theta_\nu} \asymp_{p,\nu} \frac{D_\star}{\tau_q^{p(1+\nu)/(p-1-\nu)} M_\nu \mathfrak{S}_{p,\nu,n}}.$$

Since

$$1 + \theta_\nu = \frac{p}{p-1-\nu},$$

this gives

$$\eta_{\mathrm{det}} \asymp_{p,\nu} \frac{\mu^{(1+\nu)/p} D_\star^{(p-1-\nu)/p}}{\tau_q^{1+\nu} L_\nu \mathfrak{S}_{p,\nu,n}^{(p-1-\nu)/p}},$$

where the power $\tau_q^{1+\nu}$ comes from

$$\frac{p(1+\nu)}{p-1-\nu} \cdot \frac{p-1-\nu}{p} = 1 + \nu.$$

Substituting this value of η_{det} into the deterministic term $D_\star/(\eta n)$ in (11) gives

$$O_{p,\nu}\left(\frac{\tau_q^{1+\nu} L_\nu D_\star^{(1+\nu)/p}}{\mu^{(1+\nu)/p}} \frac{\mathfrak{S}_{p,\nu,n}^{(p-1-\nu)/p}}{n} \right). \quad (12)$$

The cruder $O(n^{-\kappa_{\mathrm{na}}})$ rate then follows by applying the standard integral comparison

$$\mathfrak{S}_{p,\nu,n} = 1 + \textstyle\sum_{k=1}^{n-1} k^{-\zeta_\nu} \lesssim_{p,\nu} \begin{cases} n^{1-\zeta_\nu}, & \zeta_\nu < 1, \\ \log(en), & \zeta_\nu = 1, \\ 1, & \zeta_\nu > 1, \end{cases}$$

where the boundary $\zeta_\nu = 1$ is exactly $\nu = (p-1)/(p+1)$. Consequently, if $\zeta_\nu < 1$, then

$$\frac{\mathfrak{S}_{p,\nu,n}^{(p-1-\nu)/p}}{n} \lesssim_{p,\nu} n^{-\kappa_{\mathrm{na}}},$$

while for $\zeta_\nu > 1$ this factor is $O_{p,\nu}(n^{-1})$, and at $\zeta_\nu = 1$ it is $n^{-1}(\log(en))^{(p-1-\nu)/p}$. Thus the deterministic contribution is

$$O_{p,\nu}\left(\frac{\tau_q^{1+\nu} L_\nu D_\star^{(1+\nu)/p}}{\mu^{(1+\nu)/p} n^{\kappa_{\mathrm{na}}}}\right),$$

up to the logarithmic factor at the boundary.

Equating the deterministic and stochastic terms in (11) gives

$$\frac{D_\star}{\eta n} \asymp_p \frac{\eta^{q-1}\sigma_q^q}{\mu^{q-1}}, \qquad \text{hence} \qquad \eta^q \asymp_p \frac{\mu^{q-1} D_\star}{\sigma_q^q n}.$$

Using $q = p/(p-1)$, this yields

$$\eta_{\mathrm{stoch}} \asymp_p \frac{\mu^{1/p} D_\star^{(p-1)/p}}{\sigma_q n^{(p-1)/p}},$$

Plugging this into the deterministic term of (11) gives the stochastic rate $\sigma_q D_\star^{1/p}/(\mu^{1/p} n^{1/p})$. □

C.2 Accelerated

Theorem C.2 (Accelerated parameterization under Hölder smoothness)**.** *Let $p > 2$, let $q = p/(p-1)$, and let $\nu \in (0, 1]$. Let $x^\star \in \arg\min_{x \in \mathrm{dom}\, h} f(x)$ and $D_\star := h(x^\star) - \inf h$. Consider SODA. For every $k = 0, \dots, n-1$, choose*

$$\alpha_k = \tfrac{2}{k+2}, \qquad \bar{\alpha}_k = \lambda_k = \tfrac{2}{k+3}, \qquad \bar{\lambda}_k \le c_\nu \lambda_k, \qquad \gamma_k = \eta \tfrac{(k+2)(k+3)}{2},$$

where $c_\nu > 0$ is the constant from Lemma B.3. Suppose Assumptions 1, 3, 4 and 6 hold with $q = p/(p-1)$. Assume also that h is p-uniformly convex with constant μ in the sense of Definition 1. Define

$$\kappa_{\mathrm{acc}} := \tfrac{1+(p+1)\nu}{p}.$$

Choose

$$\eta \asymp_{p,\nu} \min\left\{ \frac{\mu^{(1+\nu)/p} D_\star^{(p-1-\nu)/p}}{\tau_q^{1+\nu} L_\nu n^{2-\kappa_{\mathrm{acc}}}}, \frac{\mu^{1/p} D_\star^{(p-1)/p}}{\sigma_q\, n^{(2p-1)/p}} \right\}.$$

Then, for every $n \ge 1$,

$$\mathbb{E}[f(x^{n-1}) - f(x^\star)] = O_{p,\nu}\left(\frac{\tau_q^{1+\nu} L_\nu D_\star^{(1+\nu)/p}}{\mu^{(1+\nu)/p} n^{\kappa_{\mathrm{acc}}}} + \frac{\sigma_q D_\star^{1/p}}{\mu^{1/p} n^{1/p}} \right).$$

In particular, for $h(x) = \frac{1}{p}\left\|x - z^0\right\|_p^p$ and $R_p := \left\|x^\star - z^0\right\|_p$, taking $\mu = 2^{2-p}$ gives

$$\mathbb{E}[f(x^{n-1}) - f(x^\star)] = O_{p,\nu}\left(\frac{\tau_q^{1+\nu} L_{\nu,p} R_p^{1+\nu}}{n^{\kappa_{\mathrm{acc}}}} + \frac{\sigma_q R_p}{n^{1/p}} \right).$$

Proof. Choose $a_k = k+1$, so that $A_k = (k+1)(k+2)/2$. Combining Lemma B.1 with Lemma B.3 gives

$$\begin{aligned} A_{n-1}\mathbb{E}[f(x^{n-1}) - f(x^\star)] \le{}& \tfrac{D_\star}{\eta} + \tfrac{\eta^{q-1}}{q\mu^{q-1}} \textstyle\sum_{k=0}^{n-1} (k+1)^q \mathbb{E}\left\|g^k - g^{k-1}\right\|_*^q \\ & - \tfrac{c_\nu}{L_\nu^{1/\nu}} \textstyle\sum_{k=1}^{n-1} A_{k-1} \mathbb{E}\left[\left\|\nabla f(y^k) - \nabla f(y^{k-1})\right\|_*^{r_\nu}\right] \end{aligned}$$

We next split the regret noise term. For $k \ge 1$, raising Assumption 4 to the qth power and using $(a+b)^q \lesssim_q a^q + b^q$ gives

$$\mathbb{E}\left\|g^k - g^{k-1}\right\|_*^q \lesssim_q \tau_q^q \mathbb{E}\left\|\nabla f(y^k) - \nabla f(y^{k-1})\right\|_*^q + \sigma_q^q.$$

For $k = 0$, using the boundary conventions $g^{-1} = 0$ and $y^{-1} = x^\star$, Assumption 4 gives

$$\mathbb{E}\left\|g^0 - g^{-1}\right\|_*^q \lesssim_q \tau_q^q \mathbb{E}\left\|\nabla f(y^0) - \nabla f(x^\star)\right\|_*^q + \sigma_q^q.$$

Since $\sum_{k=0}^{n-1}(k+1)^q \lesssim_q n^{q+1}$, absorbing only q-dependent constants gives

$$\begin{aligned} A_{n-1}\mathbb{E}[f(x^{n-1}) - f(x^\star)] \lesssim_q \frac{D_\star}{\eta} + \frac{\eta^{q-1}\sigma_q^q n^{q+1}}{q\mu^{q-1}} \\ + \frac{\tau_q^q \eta^{q-1}}{q\mu^{q-1}} \sum_{k=1}^{n-1}(k+1)^q \mathbb{E}\left[\left\|\nabla f(y^k) - \nabla f(y^{k-1})\right\|_*^q\right] \\ - \frac{c_\nu}{L_\nu^{1/\nu}} \sum_{k=1}^{n-1} A_{k-1}\mathbb{E}\left[\left\|\nabla f(y^k) - \nabla f(y^{k-1})\right\|_*^{r_\nu}\right] \\ + \frac{\tau_q^q \eta^{q-1}}{q\mu^{q-1}} \mathbb{E}\left[\left\|\nabla f(y^0) - \nabla f(x^\star)\right\|_*^q\right]. \end{aligned} \tag{13}$$

The last term of (13) is dominated by the first term by the same argument as in the non-accelerated proof, since the deterministic branch of η is again at most

$$\frac{\mu^{(1+\nu)/p} D_\star^{(p-1-\nu)/p}}{\tau_q L_\nu}.$$

It remains to control the positive and negative variation sums. Since $(k+1)^q \lesssim_q k^q$ and $A_{k-1} \gtrsim k^2$ for $k \geq 1$, it is enough to use the same scalar bound as in the non-accelerated proof:

$$\sup_{s \geq 0}\{As^q - Bs^{r_\nu}\} \leq C_{p,\nu} A^{r_\nu/(r_\nu - q)} B^{-q/(r_\nu - q)} \qquad \text{for all } A, B > 0.$$

Here $C_{p,\nu}$ is a finite constant depending only on p and ν. Applying this bound with

$$A = \frac{\tau_q^q \eta^{q-1}}{\mu^{q-1}} k^q, \qquad B = \frac{k^2}{L_\nu^{1/\nu}},$$

yields

$$\sup_{s \geq 0}\left\{\frac{\tau_q^q \eta^{q-1}}{\mu^{q-1}} k^q s^q - \frac{k^2}{L_\nu^{1/\nu}} s^{r_\nu}\right\} \leq C_{p,\nu} \frac{\tau_q^{p(1+\nu)/(p-1-\nu)} L_\nu^{p/(p-1-\nu)}}{\mu^{(1+\nu)/(p-1-\nu)}} \eta^{(1+\nu)/(p-1-\nu)} k^{p(1-\nu)/(p-1-\nu)}. \tag{14}$$

The exponent of k in (14) comes from

$$q\frac{r_\nu}{r_\nu - q} - 2\frac{q}{r_\nu - q} = \frac{p(1-\nu)}{p-1-\nu}.$$

Define, as before,

$$\theta_\nu := \frac{1+\nu}{p-1-\nu}, \qquad \xi_\nu := \frac{p(1-\nu)}{p-1-\nu}, \qquad M_\nu := \frac{L_\nu^{p/(p-1-\nu)}}{\mu^{(1+\nu)/(p-1-\nu)}}.$$

With this notation, (14) applies pointwise with

$$s_k := \left\|\nabla f(y^k) - \nabla f(y^{k-1})\right\|_*.$$

Combining (14) with $(k+1)^q \lesssim_q k^q$ and $A_{k-1} \gtrsim k^2$ gives, for each $k \geq 1$,

$$\frac{\tau_q^q \eta^{q-1}}{q\mu^{q-1}}(k+1)^q \mathbb{E}[s_k^q] - \frac{c_\nu}{L_\nu^{1/\nu}} A_{k-1}\mathbb{E}[s_k^{r_\nu}] \leq O_{p,\nu}\left(\tau_q^{p(1+\nu)/(p-1-\nu)} M_\nu \eta^{\theta_\nu} k^{\xi_\nu}\right).$$

Thus the positive and negative variation sums in (13) contribute at most the sum of these remainders. Since

$$\sum_{k=1}^{n-1} k^{\xi_\nu} \lesssim_{p,\nu} n^{1+\xi_\nu}, \qquad A_{n-1} = \Theta(n^2),$$

dividing by A_{n-1} gives

$$\mathbb{E}[f(x^{n-1}) - f(x^\star)] \leq O_{p,\nu}\left(\frac{D_\star}{\eta n^2} + \tau_q^{p(1+\nu)/(p-1-\nu)} M_\nu \eta^{\theta_\nu} n^{\xi_\nu - 1} + \frac{\eta^{q-1}\sigma_q^q n^{q-1}}{\mu^{q-1}}\right). \tag{15}$$

Equating the deterministic and Hölder-variation terms in (15) gives

$$\frac{D_\star}{\eta n^2} \asymp_{p,\nu} \tau_q^{p(1+\nu)/(p-1-\nu)} M_\nu \eta^{\theta_\nu} n^{\xi_\nu - 1} \quad \Longleftrightarrow \quad \eta^{1+\theta_\nu} \asymp_{p,\nu} \frac{D_\star}{\tau_q^{p(1+\nu)/(p-1-\nu)} M_\nu n^{1+\xi_\nu}}.$$

Since

$$1 + \theta_\nu = \frac{p}{p-1-\nu}, \qquad (1+\xi_\nu)\frac{p-1-\nu}{p} = 2 - \kappa_{\mathrm{acc}},$$

this yields

$$\eta_{\mathrm{det}} \asymp_{p,\nu} \frac{\mu^{(1+\nu)/p} D_\star^{(p-1-\nu)/p}}{\tau_q^{1+\nu} L_\nu n^{2-\kappa_{\mathrm{acc}}}}, \qquad \kappa_{\mathrm{acc}} = \frac{1+(p+1)\nu}{p}.$$

Substituting this value of η_{det} into $D_\star/(\eta n^2)$ gives the deterministic rate in the theorem.

Equating the deterministic and stochastic terms in (15) gives

$$\frac{D_\star}{\eta n^2} \asymp_p \frac{\eta^{q-1}\sigma_q^q n^{q-1}}{\mu^{q-1}}, \qquad \text{hence} \qquad \eta^q \asymp_p \frac{\mu^{q-1} D_\star}{\sigma_q^q n^{q+1}}.$$

Using $q = p/(p-1)$, this gives

$$\eta_{\mathrm{stoch}} \asymp_p \frac{\mu^{1/p} D_\star^{(p-1)/p}}{\sigma_q n^{(2p-1)/p}},$$

and substitution into $D_\star/(\eta n^2)$ gives the stochastic rate $\sigma_q D_\star^{1/p}/(\mu^{1/p} n^{1/p})$. □

Effective stepsize. For the accelerated choice in Corollary 3.2, write

$$C_{\mathrm{det}} := \frac{R_p^{p-2}}{L_p}, \qquad C_{\mathrm{stoch}} := \frac{R_p^{p-1}}{\sigma_q}.$$

The corresponding horizon-free stepsize is obtained by replacing the horizon n in η by the current iteration k:

$$\eta_k \asymp_{p,\tau_q} \min\left\{C_{\mathrm{det}} k^{-(p-2)/p}, C_{\mathrm{stoch}} k^{-(2p-1)/p}\right\}.$$

Since $\gamma_k \asymp \eta_k k^2$, this gives

$$\gamma_k^{q-1} \asymp_{p,\tau_q} \min\left\{C_{\mathrm{det}}^{1/(p-1)} k^{(p+2)/(p(p-1))}, C_{\mathrm{stoch}}^{1/(p-1)} k^{1/(p(p-1))}\right\}.$$

The effective scale $\lambda_k \gamma_k^{q-1}$, with $\lambda_k \asymp k^{-1}$, therefore has endpoints

$$p = 2: \qquad \lambda_k \gamma_k^{q-1} \asymp_{\tau_2} \min\left\{\frac{k}{L_2}, \frac{R_2}{\sigma_2 k^{1/2}}\right\}, \qquad p \to \infty: \qquad \lambda_k \gamma_k^{q-1} \asymp k^{-1}.$$

C.3 Near $p = \infty$

Corollary C.3 (Near $p = \infty$ non-accelerated parameterization)**.** *Let d denote the ambient dimension, or the effective rank in the Schatten case, and set*

$$p_d := \max\{3, \lceil \log(\mathrm{e}d) \rceil\}, \qquad q_d := \frac{p_d}{p_d - 1}.$$

Let $x^\star \in \arg\min_{x \in \mathrm{dom}\, h} f(x)$ and $R_\infty := \left\|x^\star - z^0\right\|_\infty$. Run SODA with the finite p_d geometry $h(x) = \frac{1}{p_d}\left\|x - z^0\right\|_{p_d}^{p_d}$ and the non-accelerated parameters from Theorem C.1 with $\nu = 1$. Suppose the assumptions of Theorem C.1, except for Assumption 6, hold for this choice of p_d, q_d, and h. Assume instead that f is L_∞-smooth with respect to $\|\cdot\|_\infty$. Then

$$\mathbb{E}[f(x^{n-1}) - f(x^\star)] \le C_{\tau_{q_d}} \left(\frac{L_\infty R_\infty^2 \min\{\log(\mathrm{e}n), \log(\mathrm{e}^3 d)\}}{n} + \frac{\sigma_{q_d} R_\infty}{n^{1/p_d}}\right).$$

Proof. We apply Theorem C.1 with $\nu = 1$ and $p = p_d$. The radius and smoothness constant can be controlled by their ∞-norm analogues. Specifically, the radius satisfies

$$\left\|x^\star - z^0\right\|_{p_d} \le d^{1/p_d} \left\|x^\star - z^0\right\|_\infty \le \mathrm{e}R_\infty.$$

The smoothness conversion uses the dual-norm inequality $\|\cdot\|_{q_d} \le \|\cdot\|_1$. Therefore

$$\|\nabla f(x) - \nabla f(y)\|_{q_d} \le \|\nabla f(x) - \nabla f(y)\|_1 \le L_\infty \|x - y\|_\infty \le L_\infty \|x - y\|_{p_d}.$$

Thus the finite-p_d theorem applies with $R_{p_d} \le \mathrm{e}R_\infty$ and $L_{p_d} \le L_\infty$.

We still need to track constants depending on p_d in the bound. For $\nu = 1$ and $p = p_d$, the constants depending on $q_d \in [1, 3/2]$ and $\mu = 2^{2-p_d}$ are uniformly bounded. What remains is the direct dependency on p_d in the deterministic contribution from (12),

$$\mathfrak{S}_{p_d,1,n}^{(p_d-2)/p_d}, \qquad \mathfrak{S}_{p_d,1,n} := 1 + \textstyle\sum_{k=1}^{n-1} k^{-p_d/(p_d-2)}.$$

Since $p_d/(p_d - 2) = 1 + 2/(p_d - 2)$, an integral test gives

$$\textstyle\sum_{k=1}^{n-1} k^{-p_d/(p_d-2)} \le 1 + \int_1^\infty x^{-1-2/(p_d-2)}\, dx = 1 + \frac{p_d - 2}{2}.$$

The same sum is also bounded by the harmonic series, $\sum_{k=1}^{n-1} k^{-1} \le \log(\mathrm{e}n)$. Since

$$\mathfrak{S}_{p_d,1,n} \le 1 + \min\{\log(\mathrm{e}n), 1 + \tfrac{p_d - 2}{2}\},$$

we have

$$\mathfrak{S}_{p_d,1,n}^{(p_d-2)/p_d} \lesssim \min\{\log(\mathrm{e}n), p_d\} \lesssim \min\{\log(\mathrm{e}n), \log(\mathrm{e}^3 d)\}.$$

□

Remark 2. This corollary uses a finite approximation to ∞-norm. This matches practice for Schatten-∞, where the matrix sign operation is typically not computed exactly, but approximated by an iterative solver such as Newton–Schulz.

We remark that it is also possible to obtain a similar result for the accelerated parameterization in Theorem C.2. However, we do not pursue this here as it does not lead to an improvement in the asymptotic rate, since the $O(1/n)$ term is optimal for $p = \infty$ in the smooth convex deterministic setting (Guzmán & Nemirovski, 2018, Cor. 1).

D ORACLE COMPLEXITY

D.1 ACCELERATED

Corollary D.1 (Accelerated oracle complexity with less heavy-tailed noise)**.** *Let $p \geq 2$, set $q = p/(p-1)$, and suppose $p \geq p_r$. Suppose Assumption 5 holds. Run the accelerated parameterization in Corollary 3.2 with gradient feedback $g^k = \bar{g}_B(y^k)$. Then Corollary 3.2 applies with its noise parameter σ_q replaced by $d^{1/q-1/r}\sigma_r B^{-1/p_r}$, giving*

$$\mathbb{E}[f(x^{n-1}) - f(x^\star)] \lesssim \tfrac{A_p}{n^{1+2/p}} + \tfrac{S_{r,p}}{B^{1/p_r} n^{1/p}},$$

where $R_p := \left\|x^\star - z^0\right\|_p$, $A_p = L_p R_p^2$, and $S_{r,p} := \sigma_r d^{1/q-1/r} R_p$. Hence accuracy ϵ is reached by taking

$$n \asymp \left(\tfrac{A_p}{\epsilon}\right)^{p/(p+2)}, \qquad B \asymp \left(\tfrac{S_{r,p}}{\epsilon}\right)^{p_r} \left(\tfrac{\epsilon}{A_p}\right)^{p_r/(p+2)},$$

and therefore with total stochastic gradient budget

$$N = nB = O\left(\left(\tfrac{A_p}{\epsilon}\right)^{p/(p+2)} + \left(\tfrac{S_{r,p}}{\epsilon}\right)^{p_r} \left(\tfrac{A_p}{\epsilon}\right)^{(p-p_r)/(p+2)}\right).$$

Equivalently, for a fixed oracle budget $N = nB$,

$$\epsilon_p^{\mathrm{acc}}(n, N) \lesssim \tfrac{A_p}{n^{1+2/p}} + \tfrac{S_{r,p}}{N^{1/p_r}} n^{1/p_r - 1/p}, \qquad p \geq p_r.$$

Proof. The mini-batch moment bound follows from the same von Bahr–Esseen and norm comparison argument used in Corollary 3.3 stating that the averaged feedback has effective q-moment scale $d^{1/q-1/r}\sigma_r B^{-1/p_r}$. Substituting this scale into Corollary 3.2 gives the above rate. The stated choices of n and B make the deterministic and stochastic terms at most constants times ϵ, and the total stochastic gradient budget is $N = nB$. Substituting $B = N/n$ gives the form where N is fixed. □

D.2 ACCELERATED BATCH SIZE RULE FOR A FIXED GEOMETRY

The fixed oracle budget result in Corollary D.1 also gives an accelerated analogue of the batch size rule in (2). Balancing the smooth and stochastic terms in

$$\epsilon_p^{\mathrm{acc}}(n, N) \lesssim \tfrac{A_p}{n^{1+2/p}} + \tfrac{S_{r,p}}{N^{1/p_r}} n^{1/p_r - 1/p}$$

gives

$$n_p^{\mathrm{acc}} \asymp \left(\tfrac{A_p}{S_{r,p}}\right)^{1/(1+1/p_r+1/p)} N^{\frac{1/p_r}{1+1/p_r+1/p}},$$
$$B_p^{\mathrm{acc}} = \tfrac{N}{n_p^{\mathrm{acc}}} \asymp \left(\tfrac{S_{r,p}}{A_p}\right)^{1/(1+1/p_r+1/p)} N^{\frac{1+1/p}{1+1/p_r+1/p}}.$$

At $p = p_r$, the stochastic term is independent of n after substituting $B = N/n$, so this balance gives the largest batch size, or equivalently the smallest iteration count, that remains optimal up to constants. For bounded variance, $p_r = 2$, the batch exponent is

$$B_p^{\mathrm{acc}} \asymp N^{\beta_{\mathrm{acc}}(p)}, \qquad \beta_{\mathrm{acc}}(p) = \tfrac{2(p+1)}{3p+2}, \qquad \beta_{\mathrm{acc}}(2) = \tfrac{3}{4}, \quad \lim_{p\to\infty} \beta_{\mathrm{acc}}(p) = \tfrac{2}{3}.$$

The corresponding scaling rules are summarized in Table 4, with the batch exponent plotted alongside. We see that the accelerated parameterization allows taking larger batch sizes than the non-accelerated version (*c.f.* Table 2).

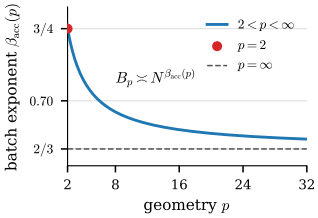
D.3 HEAVY-TAILED NOISE AND BATCHING

In the noisy setting, the accelerated complexity result in Corollary D.1 gives the same kind of d-N separation. Under ∞-smoothness, the only additional estimates are

$$A_p = L_p R_p^2 \lesssim L_\infty R_\infty^2 d^{2/p}, \qquad S_{r,p} = \sigma_r d^{1/q-1/r} R_p \lesssim \sigma_r R_\infty d^{1/p_r}.$$

Table 4: Accelerated batch size scaling rules for a fixed geometry under bounded variance. Here $N = nB$ is the total number of stochastic gradient oracle calls.

Geometry	Iterations	Batch size	Rate
$p = 2$	$n_2 \asymp N^{1/4}$	$B_2 \asymp N^{3/4}$	$\epsilon_N \asymp N^{-1/2}$
$2 < p < \infty$	$n_p \asymp N^{p/(3p+2)}$	$B_p \asymp N^{2(p+1)/(3p+2)}$	$\epsilon_N \asymp N^{-(p+2)/(3p+2)}$
$p \to \infty$	$n_\infty \asymp N^{1/3}$	$B_\infty \asymp N^{2/3}$	$\epsilon_N \asymp N^{-1/3}$



Write $A := L_\infty R_\infty^2$ and $S := \sigma_r R_\infty$. Thus, after setting $B = N/n$, fixing the budget N in Corollary D.1 gives a rate of

$$\epsilon_p^{\mathrm{stoch}}(n, N) \lesssim \frac{Ad^{2/p}}{n^{1+2/p}} + \frac{Sd^{1/p_r}}{N^{1/p_r}}\, n^{1/p_r - 1/p}, \qquad p \geq p_r.$$

At the smallest admissible geometry, taking $p = p_r$ and $B = 1$ gives

$$\epsilon_{p_r}^{\mathrm{stoch}}(N) \lesssim A\, d^{-1} \left(\tfrac{d}{N}\right)^{1+2/p_r} + S \left(\tfrac{d}{N}\right)^{1/p_r}.$$

At the other endpoint $p \to \infty$, fixing the budget N and optimizing over n gives

$$\epsilon_\infty^{\mathrm{stoch}}(N) \lesssim A\, d^{-1} \left(\tfrac{d}{N}\right) + A^{1/(1+p_r)} S^{p_r/(1+p_r)} \left(\tfrac{d}{N}\right)^{1/(1+p_r)}.$$

Thus the accelerated term is clearly better for $p = p_r$ once $N \gtrsim d$. The stochastic terms give the same comparison once

$$\tfrac{d}{N} \lesssim \left(\tfrac{A}{S}\right)^{p_r}, \qquad \text{equivalently} \qquad N \gtrsim d \left(\tfrac{S}{A}\right)^{p_r}.$$

Combining the deterministic and stochastic comparisons, the smallest admissible geometry $p = p_r$ is favored when

$$N \gtrsim d \max\left\{1, \left(\tfrac{S}{A}\right)^{p_r}\right\}.$$

This is again the low-dimensional regime $N \gtrsim d$, as summarized in Table 3.