INNER MOMENTUM FOR DIFFERENTIALLY PRIVATE MUON

Anonymous authors
Paper under double-blind review

ABSTRACT

Differentially private training clips each per-example gradient before adding noise. This clipping is radial for each example, yet unequal clipping factors can distort the relative singular-vector geometry of their average. Muon is particularly exposed to this effect, since its update is an approximate polar factor $UV^\top$ that depends only on the singular vectors that clipping can shift. To curb this degradation, we propose averaging each sampled example's Muon gradient over the current model and a short history of recent models before clipping. The clipped batch matrix then separates into a common rescaling and a covariance residual R between sampled gradients and clipping values, with $\|R\|_F \leq \sigma_\lambda \sigma_G$, bounding the clipping-induced distortion directly. We further show that a finite Newton–Schulz iteration preserves the polar factor of its input under these spectral conditions, confirming that our correction survives orthogonalization. In private GPT-2 fine-tuning on E2E and DART at $\epsilon \in \{1, 2, 4, 8\}$, DP-Muon-IM improves BLEU and ROUGE-L over DP-Muon in every seed-matched comparison, and non-private diagnostics show 2–4% lower pre-noise polar error.

1 INTRODUCTION

Training a neural network can reveal information about individual training records. Differential privacy (DP) bounds this leakage by ensuring that a model's output distribution changes only slightly depending on whether any single record was used in training (Dwork and Roth, 2014). A standard approach in private deep learning is DP stochastic gradient descent (DPSGD). It clips each per-example gradient before averaging, which limits the influence of one record, and adds Gaussian noise calibrated to that limit (Abadi et al., 2016). Both operations can affect utility. Noise perturbs the averaged gradient, while clipping can change its direction by rescaling per-example gradients unequally (Chen et al., 2020).

This work studies how clipping affects Muon (Jordan et al., 2024), an optimizer that approximately orthogonalizes matrix-valued momentum using a small number of Newton–Schulz iterations. Muon has recently been used alongside AdamW (Loshchilov and Hutter, 2019) to train modern large language models, including DeepSeek-V4 (DeepSeek-AI, 2026) and Kimi K2 via MuonClip (Kimi Team, 2025), with Muon-based training outperforming AdamW baselines on small transformers and Moonlight (Liu et al., 2025). Motivated by these gains, we ask whether Muon's advantages persist under the per-example gradient clipping and Gaussian noise required for differentially private training. In fact, per-example clipping affects Muon by changing the relative weights of the gradients in the batch average. As shown in Fig. 1, when clipping factors differ across examples, the clipped average can have different singular-vector bases, and hence a different polar factor (Chen et al., 2020; Xiao et al., 2023). DP-Muon (Kim and Fan, 2026) adds Gaussian noise to this clipped average and then applies momentum and Newton–Schulz. Because the Newton–Schulz iteration preserves the polar factor of its input in exact arithmetic, it cannot correct clipping-induced changes to that polar factor.

Instead, we propose to smooth each sampled example's Muon gradient across the current and recent private model states before clipping. Under the stated variance and temporal-covariance assumptions, averaging can tighten a bound on the expected clipping-residual norm. We measure the realized residual and polar error separately. The resulting method, DP-Muon-IM, adapts the inner-momentum approach of Xiao et al. (2023) to realize that intuition. We further formalize this idea by decomposing the clipped average into a common-rescaling term, which preserves the desired polar factor, and a

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.

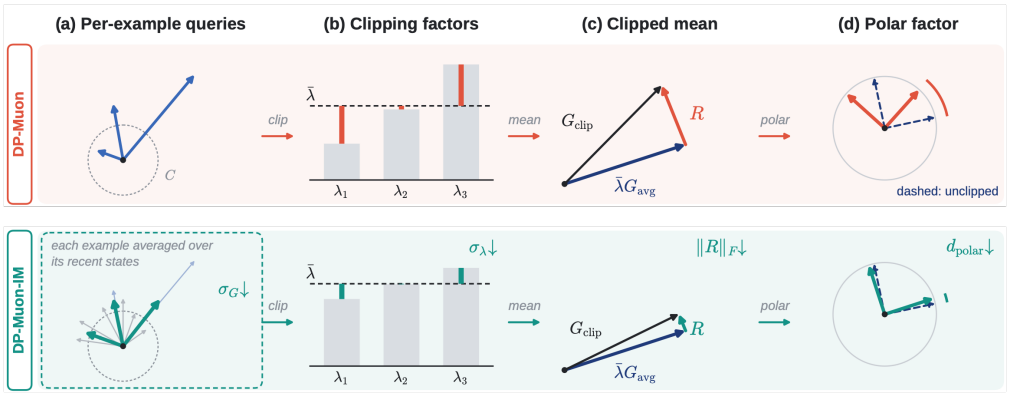

Figure 1: **Intuition for DP-Muon with inner momentum (DP-Muon-IM). (a)** DP-Muon-IM averages each example's Muon gradients at the current and recent private model states before flat clipping. **(b-c)** Unequal clipping factors introduce a residual R relative to a common rescaling of the batch mean. IM averages per-example gradients before this reweighting. **(d)** The residual can change the polar factor. We measure residual size and polar error separately; the utility experiments evaluate the complete method.

residual governed by variation in the gradients and clipping factors. Privacy accounting remains unchanged: each example contributes a single query clipped to the same radius as in the baseline Gaussian mechanism, with the same sampling procedure and release schedule. In GPT-2 fine-tuning on E2E and DART, DP-Muon-IM improves BLEU and ROUGE-L over DP-Muon.

Our contributions are as follows.

- Section 3 explains how clipping can alter the matrix geometry relevant to Muon by decomposing the clipped average into a common rescaling and a residual R, with $\|R\|_{\mathrm{F}} \leq \sigma_\lambda \sigma_G$. We also show that the stated finite Newton–Schulz iteration preserves its immediate input's polar factor in exact arithmetic, so it cannot remove polar error present at that stage.
- Section 4 introduces DP-Muon-IM, which averages each sampled example's gradients at current and recent private model states before clipping. The sampling rule, clipping radius, noise calibration, release schedule, and baseline per-run privacy accounting remain unchanged.
- Section 5 presents private GPT-2 experiments showing that DP-Muon-IM improves both BLEU and ROUGE-L scores on E2E and DART compared with DP-Muon. We report E2E and DART results at target $\epsilon \in \{1, 2, 4, 8\}$, and a single-seed history-length ablation quantifies the tradeoff between generation quality, training time, and memory.

Organization. Section 2 introduces notation and background. Section 3 analyzes clipping-induced distortion, and Section 4 describes DP-Muon-IM. Section 5 reports the experiments. Sections 6–8 cover related work, discussion, and conclusions.

2 Preliminaries

2.1 Problem setup and notation

Let $\mathcal{D} = \{x_i\}_{i=1}^n$ be the private training set, let θ denote the model parameters, and let $\ell(\theta; x_i)$ be the loss on example x_i. Training seeks to minimize the empirical risk

$$F(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; x_i). \tag{1}$$

At training step t, the optimizer samples a minibatch S_t (also called a *lot* in DP training), and we write $m = |S_t|$ for its realized size. The per-example gradient is

$$G_t^{(i)} = \nabla_\theta \ell(\theta_t; x_i). \tag{2}$$

Throughout we treat a gradient as a vector over all trainable parameters. However, algorithmically Muon operates on matrix representations of specific subsets of parameters. Thus let $\mathcal{W}$ be the set of

hidden-layer matrices operated on by Muon, with each matrix corresponding to a specific subset of parameter indices. For any $W \in \mathcal{W}$, the corresponding block of $G_t^{(i)}$ is implicitly converted to its matrix form,

$$G_{t,W}^{(i)} = \nabla_W \ell(\theta_t; x_i) \in \mathbb{R}^{n_{\text{out}}^W \times n_{\text{in}}^W}. \tag{3}$$

Muon acts separately on each $W \in \mathcal{W}$. Non-Muon parameters are omitted from the blockwise notation. If present, their gradients remain in $G_t^{(i)}$, receive the same clipping-and-noise mechanism, and are updated by the auxiliary optimizer. The norm $\|G_t^{(i)}\|_2$ treats all parameter blocks as one vector.

2.2 Differential privacy and private optimization

Two datasets are neighbors if they differ by the addition or removal of one record. A randomized algorithm $\mathcal{M}$ is (ϵ, δ)-differentially private (Dwork and Roth, 2014) if, for every neighboring pair $\mathcal{D}, \mathcal{D}'$ and every measurable output set $\mathcal{E}$,

$$\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{E}] \le e^{\epsilon} \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{E}] + \delta, \tag{4}$$

where ϵ is the privacy budget and δ allows a small additive relaxation; smaller values give stronger privacy.

To achieve this in optimization, given a clipping radius $C > 0$, flat clipping bounds each example's contribution by rescaling its complete gradient:

$$\text{clip}_C(G_t^{(i)}) = \lambda_t^{(i)} G_t^{(i)}, \qquad \lambda_t^{(i)} = \min\left\{1, \frac{C}{\|G_t^{(i)}\|_2}\right\}, \tag{5}$$

with $\lambda_t^{(i)} = 1$ when $G_t^{(i)} = 0$. For matrix-level equations, $\text{clip}_C(G_{t,W}^{(i)})$ denotes the W block of $\text{clip}_C(G_t^{(i)})$, with the clipping factor still determined by the full gradient norm. Letting B be the (expected) lot size, the private release is

$$\widetilde{G}_t = \frac{1}{B}\left(\sum_{i \in S_t} \text{clip}_C(G_t^{(i)}) + \sigma_{\text{DP}} C Z_t\right), \qquad Z_t \sim \mathcal{N}(0, I), \tag{6}$$

with $\widetilde{G}_{t,W}$ denoting the matrix block of $\widetilde{G}_t$ for $W \in \mathcal{W}$. Clipping controls sensitivity, Gaussian noise provides privacy, and all subsequent optimizer operations are post-processing.

2.3 Muon and DP-Muon

Muon updates each hidden-layer matrix separately (Jordan et al., 2024). Fix $W \in \mathcal{W}$, and let $\widetilde{G}_{t,W}$ be its block of the private release in Eq. (6). DP-Muon (Kim and Fan, 2026) applies

$$\begin{aligned} M_{t,W} &= \mu M_{t-1,W} + \widetilde{G}_{t,W}, & N_{t,W} &= \widetilde{G}_{t,W} + \mu M_{t,W}, \\ O_{t,W} &= \text{Newton-Schulz}_J(N_{t,W}), & W_{t+1} &= W_t - \eta O_{t,W}. \end{aligned} \tag{7}$$

Here $M_{t,W}$ is the outer-momentum buffer, $N_{t,W}$ is the Nesterov direction, and $O_{t,W}$ is the orthogonalized direction. The operator Newton-Schulz_J applies J Newton–Schulz steps; Section 3.2 gives the exact iteration. We call $M_{t,W}$ *outer momentum* to distinguish it from the per-example *inner momentum* introduced next. Inner momentum acts before clipping, whereas outer momentum acts on the private release. Section 4 gives the complete training procedure.

2.4 Metrics

For nonzero matrices X and Y, define the cosine error

$$d_{\cos}(X, Y) = 1 - \frac{\langle X, Y \rangle_{\text{F}}}{\|X\|_{\text{F}} \|Y\|_{\text{F}}}. \tag{8}$$

This is cosine distance after vectorizing the matrices. For a reduced singular value decomposition $X = U\Sigma V^\top$ over the positive singular values, define $\text{polar}(X) = UV^\top$ and $\text{polar}(0) = 0$. Individual singular vectors are not unique within a repeated-singular-value subspace, but $UV^\top$ is invariant to the corresponding joint change of basis. It is the partial isometry mapping the right singular subspace to the left singular subspace. The polar error is

$$d_{\text{polar}}(X, Y) = \|\text{polar}(X) - \text{polar}(Y)\|_{\text{F}}. \tag{9}$$

Cosine error compares all entries, whereas polar error compares this basis-invariant map used by Muon. We report both because two matrices can differ in entrywise cosine similarity while sharing a polar factor.

3 GEOMETRIC EFFECTS OF GRADIENT CLIPPING

Muon's Newton–Schulz update preserves the input's polar factor while transforming its singular values. This makes flat clipping, as used in standard DP-SGD, a poor fit for Muon. The reason is that different examples receive different clipping factors, clipping reweights each example's contribution. In other words, the straightforward *rescale-then-average* mechanism that makes DP-SGD's privacy guarantee simple actually corrupts the geometric object that drives Muon's update.

We make this concrete in three steps. First, in §3.1 we decompose the clipped batch average into a common-rescaling term, which would leave the polar factor unchanged, and a residual term R that captures exactly the distortion caused by unequal per-example clipping factors. Second, in §3.2 we show that Newton–Schulz orthogonalization, which computes the polar factor of a matrix, does not undo this distortion. Third, in §B we give a minimal two-example numerical example that makes the resulting change in update direction explicit and quantifiable.

3.1 THE CLIPPING RESIDUAL

Fix a step t and a block $W \in \mathcal{W}$. Write G_i for the matrix supplied by example i for this block before clipping, and let λ_i be the clipping factor computed from its complete query. For DP-Muon, $G_i = G_{t,W}^{(i)}$. We suppress t and W below.

For a batch S_t of size m, define

$$G_{\text{avg}} = \frac{1}{m}\sum_{i\in S_t} G_i, \qquad G_{\text{clip}} = \frac{1}{m}\sum_{i\in S_t} \lambda_i G_i, \qquad \bar{\lambda} = \frac{1}{m}\sum_{i\in S_t} \lambda_i. \tag{10}$$

If every example has the same clipping factor, then $G_{\text{clip}} = \bar{\lambda} G_{\text{avg}}$, so clipping cannot change the singular vectors or polar factor of the mean. Unequal factors add the residual below.

Define the variation in clipping factors and matrix gradients by

$$\sigma_\lambda^2 = \frac{1}{m}\sum_{i\in S_t} (\lambda_i - \bar{\lambda})^2, \qquad \sigma_G^2 = \frac{1}{m}\sum_{i\in S_t} \|G_i - G_{\text{avg}}\|_{\text{F}}^2 . \tag{11}$$

Proposition 1 (Residual identity and bound)**.** *The clipped mean can be written as*

$$G_{\text{clip}} = \bar{\lambda} G_{\text{avg}} + R, \qquad R = \frac{1}{m}\sum_{i\in S_t} (\lambda_i - \bar{\lambda})(G_i - G_{\text{avg}}), \tag{12}$$

where the residual satisfies $\|R\|_{\text{F}} \le \sigma_\lambda \sigma_G$.

The proof is in Appendix A.1.

The term $\bar{\lambda} G_{\text{avg}}$ preserves the singular vectors and polar factor, so any clipping-induced change to them must arise from R. Not every component of R changes the polar factor, however, and the bound in Proposition 1 is only a worst-case envelope; it need not predict the observed polar discrepancy tightly. The residual vanishes when the clipping factors are equal ($\sigma_\lambda = 0$) or the per-example matrices are equal ($\sigma_G = 0$). The same argument applies to per-matrix clipping using the block's own clipping factor. Section 5.3 reports polar error, and Appendix D.1 reports the scale-normalized clipping residual for historical and current-only queries.

Appendix B gives a worked two-example calculation of the clipping factors, residual, and polar error.

3.2 WHAT NEWTON–SCHULZ PRESERVES

Muon uses a small number of Newton–Schulz iterations, each a polynomial in $XX^\top$ applied to X:

$$X^+ = aX + b(XX^\top)X + c(XX^\top)^2 X, \qquad (a, b, c) = (3.4445, -4.7750, 2.0315). \tag{13}$$

Let $\text{Newton-Schulz}_J(X)$ denote J such steps, starting from $X/(\|X\|_{\text{F}} + \varepsilon)$ with $\varepsilon \ge 0$.[1] Observe that each step acts only on singular values: if $X = U\Sigma V^\top$ in reduced SVD form, then $X^+ =$

[1]These coefficients, from Jordan et al. (2024), are chosen so that small singular values grow quickly; the iterates approach an approximately orthogonal matrix rather than converging exactly to the polar factor.

$U\,p(\Sigma)\,V^\top$, where $p(\sigma) = a\sigma + b\sigma^3 + c\sigma^5$ acts entrywise on Σ. The singular vectors are therefore unchanged, and the polar factor is preserved as long as no singular value changes sign. Proposition 2 makes this precise.

Proposition 2 (Newton–Schulz preserves the polar factor)**.** *Suppose $c > 0$ and $b^2 < 4ac$. Then, in exact arithmetic, for every nonzero matrix X, every $\varepsilon \geq 0$, and every finite $J \geq 1$,* $\mathrm{polar}\big(\text{Newton-Schulz}_J(X)\big) = \mathrm{polar}(X)$. *Consequently, for nonzero X and Y,*

$$d_{\mathrm{polar}}\big(\text{Newton-Schulz}_J(X), \text{Newton-Schulz}_J(Y)\big) = d_{\mathrm{polar}}(X, Y). \tag{14}$$

Muon's coefficients satisfy both conditions: $4ac - b^2 = 5.189382 > 0$.

The proof is in Appendix A.2.

For a constant clipping factor across samples, $\lambda_i = \bar{\lambda}$, we have $R = 0$ in Eq. (12). Then G_{clip} is a positive rescaling of G_{avg}, so their polar factors, and hence their Muon updates, coincide. Any change in the polar factor due to non-constant clipping is therefore captured by the residual R. Note that Eq. (14) concerns only the matrices entering Newton–Schulz. In DP-Muon (Kim and Fan, 2026), Gaussian noise and outer momentum are applied first and may change the discrepancy. Even a common clipping rescaling can change the signal-to-noise ratio or the balance between the current gradient and the momentum buffer.

4 Inner momentum before clipping

Section 3 showed that unequal per-example clipping changes the clipped mean's polar factor only through the residual R, with $\|R\| \leq \sigma_\lambda \sigma_G$ (Proposition 1). To reduce this distortion, we average each example's gradients over the current and recent private model states *before* clipping, while leaving the rest of DP-Muon unchanged. We call this *inner momentum*, distinguishing it from Muon's outer momentum, which averages noisy batch gradients after clipping.

The section has three parts. §4.1 defines the per-example average. §4.2 is the central result: averaging shrinks the envelope of R by a factor $\rho + (1-\rho)S_K$, where ρ measures how correlated an example's gradient fluctuations are across states and S_K is the sum of squared averaging weights. The factor is below one when $\rho < 1$ and at least two weights are nonzero. §4.3 gives the resulting algorithm, its cost, and its privacy guarantee. The envelope result bounds R and does not prove the realized R is smaller, we measure the resulting polar error directly in Section 5.

4.1 Averaging one example across model states

Let $K_t = \min\{t, K_{\max}\}$. For $i \in S_t$ and each Muon matrix block $W \in \mathcal{W}$, evaluate

$$G^{(i)}_{t-j,W} = \nabla_W \ell(\theta_{t-j}; x_i), \qquad j = 0, \ldots, K_t, \tag{15}$$

where $j = 0$ is the current model and every term uses the same example x_i. With decay $\gamma \in [0, 1]$ and normalized weights $\alpha_j = \gamma^j / \sum_{\ell=0}^{K_t} \gamma^\ell$, form

$$H^{(i)}_{t,W} = \sum_{j=0}^{K_t} \alpha_j \, G^{(i)}_{t-j,W}. \tag{16}$$

If $K_t = 0$ or $\gamma = 0$, then $H^{(i)}_{t,W} = G^{(i)}_{t,W}$ and we recover DP-Muon. For each example, the blocks $H^{(i)}_{t,W}$ are combined with the non-Muon blocks, which are averaged in the same way with their own history length and decay (K_A, γ_A), and clipped as one query, with a scalar factor λ_i computed from all trainable coordinates. Proposition 1 then applies with $G_i = H^{(i)}_{t,W}$.

We normalize by $\sum_j \gamma^j$, unlike the unnormalized sum of Xiao et al. (2023), for two reasons. First, $H^{(i)}_{t,W}$ stays on the scale of a single gradient, so the clipping radius C keeps its meaning. An unnormalized sum would inflate query norms by up to $\sum_j \gamma^j$ and force stronger clipping. Second, the weights sum to one, which is what the analysis in §4.2 uses.

4.2 AVERAGING SHRINKS THE RESIDUAL ENVELOPE

Proposition 1 bounds R by the product $\sigma_\lambda \sigma_G$, so the question is whether averaging lowers both factors. Let $g_j(z) = \nabla_\theta \ell(\theta_{t-j}; z) \in \mathbb{R}^d$ be example z's full-query gradient over all trainable coordinates, including Muon and non-Muon parameters. Define $\mu_{j,u} = \mathbb{E}\, g_j(z)$ and $\xi_{j,u}(z) = g_j(z) - \mu_{j,u}$. With the same weights for all coordinates, the averaged query is $u_i = \sum_j \alpha_j g_j(z_i)$. Let $S_K = \sum_j \alpha_j^2$, which is 1 without averaging and $1/(K+1)$ for equal weights. We assume, for constants $\tau_u \geq 0$ and $\rho \in [0, 1]$, that $\mathbb{E}\|\xi_{j,u}\|_2^2 \leq \tau_u^2$ and $\mathbb{E}\langle \xi_{j,u}, \xi_{\ell,u}\rangle \leq \rho \tau_u^2$ for $j \neq \ell$ on the same example. The examples in a batch are independent. Set

$$s_{K,u}^2 = \tau_u^2 \big(\rho + (1-\rho) S_K\big). \tag{17}$$

Here τ_u bounds the root-mean-square fluctuation of a full-query gradient, and ρ bounds its cross-state covariance relative to τ_u^2.

Proposition 3 (Residual envelope under averaging)**.** *Let $\lambda_i = \min\{1, C/\|u_i\|_2\}$ be the flat-clipping factor of example i's averaged query u_i, and let R_W be the residual of Proposition 1 for a block W of the query. Then*

$$\mathbb{E}\,\|R_W\|_F \;\leq\; \frac{m-1}{m}\,\frac{s_{K,u}^2}{C} \;=\; \frac{m-1}{m}\,\frac{\tau_u^2}{C}\,\big(\rho + (1-\rho) S_K\big). \tag{18}$$

The current-only query ($K = 0$) satisfies the same bound with $s_{0,u} = \tau_u$. Averaging therefore shrinks the envelope by the factor $\rho + (1-\rho)S_K$, which is below one whenever $\rho < 1$ and at least two weights are nonzero.

Why it holds. The proof has two steps, each answering one factor of Proposition 1. *(i) σ_G shrinks.* The weights sum to one, so the population means $\mu_{j,u}$ cancel when one example's averaged gradient is compared with the batch mean. The remaining fluctuations combine with variance at most $\tau_u^2(\rho + (1-\rho)S_K)$. *(ii) σ_λ is controlled by the same dispersion.* The map $n \mapsto \min\{1, C/n\}$ is $1/C$-Lipschitz, and $|\,\|u_i\| - \|\bar{u}\|\,| \leq \|u_i - \bar{u}\|$, so $\sigma_\lambda \leq \sigma_u / C$, where σ_u is the dispersion of the complete queries. A block is a subset of coordinates, so $\sigma_{G,W} \leq \sigma_u$, and hence $\|R_W\|_F \leq \sigma_\lambda \sigma_{G,W} \leq \sigma_u^2 / C$, which is quadratic in the dispersion. The full proof is in Appendix F.

Separate histories. Proposition 3 uses one set of weights for all coordinates. Suppose instead that each parameter group b (for example, Muon and AdamW blocks) uses its own weights, with constants τ_b, ρ_b and squared-weight sum $S_{K,b}$. Set $s_b^2 = \tau_b^2\big(\rho_b + (1-\rho_b) S_{K,b}\big)$ and $s_u^2 = \sum_b s_b^2$. The same argument, combined with Cauchy–Schwarz, gives

$$\mathbb{E}\,\|R_W\|_F \leq \frac{m-1}{m}\,\frac{s_u\, s_W}{C}, \tag{19}$$

where s_W is the constant of the group containing W. A group without history ($S_{K,b} = 1$) contributes an unreduced term to s_u.

What it does and does not say. It bounds R from above and does not prove that the realized R is smaller. No lower bound on the current-only residual is available in general. The reduction also needs $\rho < 1$, meaning an example's gradient fluctuations at nearby states must not be perfectly correlated. This is plausible with dropout or other evaluation randomness and with parameter drift. Staleness does not enter Eq. (18): averaging over older states affects the residual only through fluctuations, and affects the bias of the averaged gradient through a drift term D_K that grows with K and γ. Appendix F (Theorem 6) shows how the two combine in a polar-error bound. This is the trade-off that $K_{\max}$ and γ control.

4.3 TRAINING PROCEDURE

Algorithm 1 gives the update for one Muon block $W \in \mathcal{W}$, where $\mathcal{W}$ is the set of Muon matrices. We drop the block subscript and apply the same steps to every block. The only change from DP-Muon is line 4, which replaces the current gradient $G_t^{(i)}$ by the averaged $H_t^{(i)}$ of Eq. (16). Clipping, noise, outer momentum, and Newton–Schulz are unchanged. Because Newton–Schulz preserves the polar factor (Proposition 2), any correction must act before it, so we leave it unchanged.

Algorithm 1 DP-Muon with inner momentum (DP-Muon-IM).

Require: Iterations T, learning rate η, momentum μ, and NS steps J
Require: History $K_{\max}, \gamma$ and privacy parameters $C, \sigma_{\mathrm{DP}}, q, B$
$M_{-1} \leftarrow 0$.
for $t = 0, \dots, T-1$ **do**
  Poisson-sample $S_t \subset \mathcal{D}$ with rate q; $K_t \leftarrow \min\{t, K_{\max}\}$.
  $H_t^{(i)} \leftarrow \dfrac{\sum_{j=0}^{K_t} \gamma^j \nabla_W \ell(\theta_{t-j}; x_i)}{\sum_{j=0}^{K_t} \gamma^j}, \qquad i \in S_t.$
  $\widetilde{G}_t \leftarrow \dfrac{1}{B}\left(\sum_{i \in S_t} \mathrm{clip}_C(H_t^{(i)}) + \sigma_{\mathrm{DP}} C Z_t\right), \quad Z_t \sim \mathcal{N}(0, I_{\dim(W)}).$
  $M_t \leftarrow \mu M_{t-1} + \widetilde{G}_t$.
  $N_t \leftarrow \widetilde{G}_t + \mu M_t$.
  $O_t \leftarrow \text{Newton-Schulz}_J(N_t)$.
  $W_{t+1} \leftarrow W_t - \eta O_t$.
end for
return θ_T

Cost. The history holds at most $K_{\max}$ pre-update model states, and each sampled example needs up to $K_t + 1$ gradient evaluations, so the per-step cost grows linearly in $K_{\max}$.

Privacy. History averaging adds no private releases. To analyze privacy at each step, we hold the previous noisy releases fixed, which also fixes the current and stored model states. Each sampled example then contributes one query clipped at C, and Gaussian noise is added once to the batch sum. The baseline accounting applies whenever the sampling rule, clipping radius, noise calibration, release denominator, and number of releases are unchanged. Theorem 4 makes this precise (see Appendix E for the proof).

Theorem 4 (Privacy of DP-Muon-IM)**.** *Algorithm 1 is (ϵ, δ)-DP under add/remove adjacency whenever a valid Poisson-sampled Gaussian accountant certifies this guarantee for T updates with sampling rate q and noise multiplier σ_{DP}. The guarantee holds for any fixed public history length $K_{\max}$ and decay $\gamma \in [0, 1]$.*

5 Experiments

We fine-tune GPT-2 on two table-to-text generation tasks and compare DP-Muon-IM with DP-Muon across privacy budgets. We also compare with published Muon-based private optimizers, and study how the history length affects generation quality, training cost, and polar error.

Data and model. We fully fine-tune GPT-2-small on E2E (Novikova et al., 2017) and DART (Nan et al., 2021). This is a standard benchmark for private fine-tuning (Li et al., 2022; Zhang et al., 2025) and the setting used to evaluate DP-Muon (Kim and Fan, 2026). Training/validation/test splits contain $42{,}061/4{,}672/4{,}693$ records on E2E and $62{,}659/6{,}980/12{,}552$ on DART.

Optimizers and training. Our baseline, DP-Muon, disables both inner histories ($K = 0$). DP-Muon-IM adds separate Muon and AdamW histories, with lengths K_M, K_A and decays γ_M, γ_A, extending Section 4. Each per-example query is clipped at $C = 1$ before noise is added.

We train for ten epochs on one H100 80GB, with nominal logical/physical batches of $1024/32$. Muon uses momentum 0.95, Nesterov updates, and five Newton–Schulz steps. Learning rates decay linearly without warm-up or weight decay. Hyperparameter details for our runs are provided in Appendix C.

Privacy and selection. Our utility runs target $\epsilon \in \{1, 2, 4, 8\}$ with $\delta = 1/(2N)$, where N is the number of training records. We calibrate noise using PRV accounting (Gopi et al., 2021). We select configurations by validation loss on seed 37 and evaluate seeds 37–39.

Evaluation. We compute BLEU (Papineni et al., 2002) and multi-reference ROUGE-L (Lin, 2004). Both scores are reported on a 0–100 scale; higher is better. E2E has 630 test prompts and $4{,}693$ references; DART has $5{,}097$ prompt groups and $12{,}551$ nonempty references.

5.1 Comparison with existing methods

Table 1 compares Muon-based optimizers on E2E at target $\epsilon = 8$. We take the DP-Muon and DP-Muon-BC scores from Kim and Fan (2026); DP-Muon-BC adds a correction for privacy-noise bias

Table 1: E2E test results for GPT-2 fine-tuning at target $\epsilon = 8$.

Method	**Source**	**BLEU** ↑	**ROUGE-L** ↑
DP-Muon	Kim and Fan (2026)	63.78	66.96
DP-Muon-BC	Kim and Fan (2026)	63.66	66.71
KF-DP-Muon-IM	This work	64.77	67.37
KF-DP-Muon	This work	64.95	67.36
DP-Muon	This work	64.79	67.58
DP-Muon-IM	This work	65.65	67.80

Table 2: Test generation after ten epochs: mean ± standard deviation over seeds 37–39.

		DP-Muon		DP-Muon-IM	
Task	Target ϵ	BLEU ↑	ROUGE-L ↑	BLEU ↑	ROUGE-L ↑
E2E	1	57.64 ± 1.97	62.71 ± 0.21	60.21 ± 2.65	64.41 ± 1.02
	2	63.05 ± 0.34	65.89 ± 0.62	63.76 ± 0.26	66.43 ± 0.60
	4	64.10 ± 0.29	66.72 ± 0.18	64.83 ± 0.57	67.03 ± 0.39
	8	64.74 ± 0.05	67.31 ± 0.25	65.34 ± 0.39	67.60 ± 0.17
DART	1	32.34 ± 0.59	50.20 ± 0.23	33.19 ± 0.42	50.91 ± 0.16
	2	34.88 ± 0.37	52.27 ± 0.12	35.94 ± 0.29	53.09 ± 0.13
	4	36.48 ± 0.35	53.52 ± 0.27	37.16 ± 0.33	54.24 ± 0.17
	8	37.44 ± 0.33	54.35 ± 0.30	37.95 ± 0.19	54.79 ± 0.12

after Newton–Schulz. Our KF variants apply DiSK's simplified Kalman filter to the private release before the Muon and AdamW updates (Zhang et al., 2025). KF-DP-Muon disables inner history; KF-DP-Muon-IM adds history averaging. DP-Muon and DP-Muon-IM use per-metric maxima over seeds 37–39.

5.2 Utility across privacy budgets

DP-Muon-IM improves BLEU and ROUGE-L over DP-Muon in all 24 seed-matched comparisons (two datasets, four privacy budgets, three seeds; Table 2). Mean BLEU gains are 0.60–2.57 points on E2E and 0.50–1.06 on DART. Selected E2E configurations use $K_M = 4$, $\gamma_M = \gamma_A = 0.5$, and $K_A = 3$ for $\epsilon \in \{1, 2, 4\}$ or $K_A = 2$ for $\epsilon = 8$. DART uses $(K_M, K_A) = (3, 3)$ and $(\gamma_M, \gamma_A) = (0.7, 0.5)$ throughout.

Appendix C.1 reports a fixed-learning-rate ablation with Muon-only history, AdamW-only history, both, and neither.

5.3 Effect of history length

Utility and cost. We vary $K_M = K_A = K \in \{0, \ldots, 4\}$ on E2E at $\epsilon \approx 4$ for ten epochs, with Muon/AdamW learning rates $0.002/0.0035$ and decays $\gamma_M = \gamma_A = 0.5$ for $K > 0$; $K = 0$ is DP-Muon without history. One historical state provides most of the BLEU gain: $K = 0, 1, 4$ yield validation BLEU scores of 66.97, 67.76, and 67.82 in 36.9, 59.0, and 123.7 minutes, respectively, with nonmonotonic gains across K (Figure 2, left). Peak GPU allocation rises from 26.70 GiB without history to 49.85 GiB with history. Timings exclude setup, evaluation, and checkpoint I/O.

Polar error. We measure polar error with Muon-only history, $\gamma = 0.9$ (Figure 2, right). Each K follows a separate trajectory. Within each trajectory, we compare IM and current-only clipped means at the same model state against a shared unclipped current-gradient reference, before noise and outer momentum. Error is the root-sum-square polar-factor distance across 48 Muon matrices, using 512 query records and 2,048 disjoint held-out reference records. Just before the 10th optimizer update, IM reduces polar error by 1.76% at $K = 1$ and 3.71% at $K = 4$ relative to the current-only estimate.

Clipping residual. With $K = 4$ and $\gamma = 0.9$, historical averaging lowers the normalized clipping residual in all 21 seed–checkpoint comparisons. Mean paired reductions range from 25.0% to 46.1% across checkpoints (Appendix D.1). These raw-query diagnostics precede noise and outer momentum.

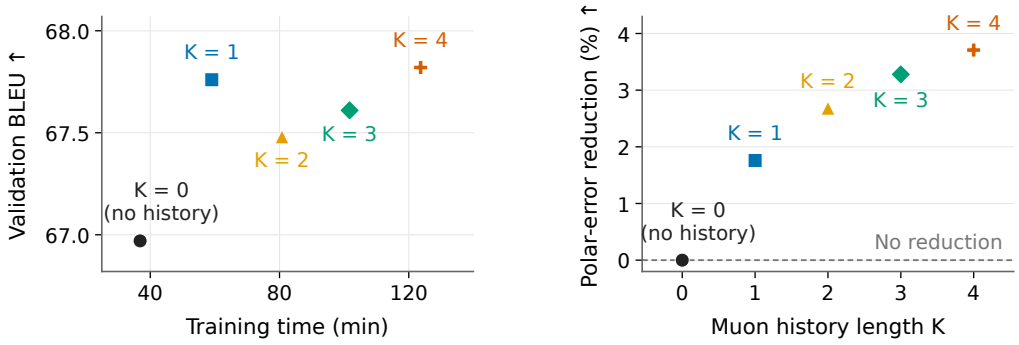

Figure 2: Effect of history length in separate utility and geometry experiments. **Left:** validation BLEU versus training time after ten epochs. **Right:** pre-noise polar-error reduction before optimizer update step 10, with Muon-only history.

6 Related Work

Clipping bias and pre-clipping averaging. Geometric analyses show that clipping can bias private optimization by changing the aggregate gradient (Chen et al., 2020). Xiao et al. (2023) relate this bias to per-example sampling noise, introduce inner–outer momentum, and report cosine similarity between unclipped and averaged-clipped gradients. Adaptive clipping and error feedback offer other treatments of clipping bias (Bu et al., 2023; Zhang et al., 2024); DP-PMLF combines per-sample momentum with post-release filtering (Xu et al., 2026). We adapt pre-clipping averaging to Muon's matrix updates and study the resulting clipping residual (Section 3.1).

Muon under differential privacy. Muon applies momentum and a finite Newton–Schulz polynomial to hidden-layer matrix gradients (Jordan et al., 2024). DP-Muon studies private matrix optimization and Gaussian perturbations under this transform (Kim and Fan, 2026); Kim and Oh (2026) analyze optimization with finitely many Newton–Schulz steps. Section 3.2 characterizes the polar geometry preserved by the iteration used here.

7 Discussion

The E2E and DART results support averaging per-example gradients before clipping. The residual decomposition explains why this stage matters: unequal clipping can change the batch mean's polar factor, while the Newton–Schulz iteration used here preserves its immediate input's polar factor in exact arithmetic.

The geometry probes compare clipped IM and current-only means against the same current-gradient reference. With history length $K = 4$ and decay $\gamma = 0.9$, IM reduces polar error by 2.4–3.7% across measured checkpoints (Appendix D). These measurements precede Gaussian noise and outer momentum; they do not measure the final Muon update.

The E2E ablation favors a short history when compute is limited: $K = 1$ came within 0.06 validation BLEU points of $K = 4$ in less than half the training time (Figure 2). This result favors a short history in the tested setting. The extra gradient evaluations and stored model states remain costs to address in larger-scale experiments.

8 Conclusion

We identify a covariance residual caused by unequal clipping and show that the Newton–Schulz iteration used here preserves its input's polar factor in exact arithmetic. DP-Muon-IM averages per-example gradients at current and recent private model states before clipping, retaining baseline per-run privacy accounting under the stated conditions. In GPT-2 fine-tuning, DP-Muon-IM achieves higher BLEU and ROUGE-L than DP-Muon on E2E and DART at target $\epsilon \in \{1, 2, 4, 8\}$.

AI Use Statement

We used large language models for language editing, LaTeX support, bibliographic checks, and mathematical consistency. We reviewed all AI-assisted work and take full responsibility for the paper.

REFERENCES

Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 308–318. Association for Computing Machinery, 2016. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Automatic clipping: Differentially private deep learning made easier and stronger. In *Advances in Neural Information Processing Systems*, volume 36, pages 41727–41764, 2023. doi: 10.52202/075280-1808. First posted as arXiv:2206.07136 in 2022.

Xiangyi Chen, Zhiwei Steven Wu, and Mingyi Hong. Understanding gradient clipping in private SGD: A geometric perspective. In *Advances in Neural Information Processing Systems*, volume 33, pages 13773–13782, Red Hook, NY, USA, 2020. Curran Associates Inc. URL https://proceedings.neurips.cc/paper/2020/hash/9ecff5455677b38d19f49ce658ef0608-Abstract.html.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. arXiv preprint arXiv:2606.19348, 2026. URL https://arxiv.org/abs/2606.19348.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–487, August 2014. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11631–11642. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/6097d8f3714205740f30debe1166744e-Paper.pdf.

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

Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton–Schulz. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4c0129f1f7a036af5ed2409f1f9f8131-Abstract-Conference.html. Also available as arXiv:2601.19156.

Jihwan Kim and Chenglin Fan. DP-Muon: Differentially private optimization via matrix-orthogonalized momentum. arXiv preprint arXiv:2605.12994, 2026. URL https://arxiv.org/abs/2605.12994.

Kimi Team. Kimi K2: Open agentic intelligence. arXiv preprint arXiv:2507.20534, 2025. URL https://arxiv.org/abs/2507.20534v2. Revised February 2026.

Ren-Cang Li. New perturbation bounds for the unitary polar factor. *SIAM Journal on Matrix Analysis and Applications*, 16(1):327–332, 1995. doi: 10.1137/S0895479893256359.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=bVuP3ltATMz.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng

Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. arXiv preprint arXiv:2502.16982, 2025. URL https://arxiv.org/abs/2502.16982.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. DART: Open-domain structured data record to text generation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.37. URL https://aclanthology.org/2021.naacl-main.37/.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. The E2E dataset: New challenges for end-to-end generation. In Kristiina Jokinen, Manfred Stede, David DeVault, and Annie Louis, editors, *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5525. URL https://aclanthology.org/W17-5525/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Hanshen Xiao, Zihang Xiang, Di Wang, and Srinivas Devadas. A theory to instruct differentially-private learning via clipping bias reduction. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 2170–2189, 2023. doi: 10.1109/SP46215.2023.10179409.

Xincheng Xu, Thilina Ranbaduge, Qing Wang, Thierry Rakotoarivelo, and David Smith. Enhancing DPSGD via per-sample momentum and low-pass filtering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(32):27341–27349, 2026. doi: 10.1609/aaai.v40i32.39951.

Xinwei Zhang, Zhiqi Bu, Zhiwei Steven Wu, and Mingyi Hong. Differentially private SGD without clipping bias: An error-feedback approach. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uFbWHyTlPn.

Xinwei Zhang, Zhiqi Bu, Borja Balle, Mingyi Hong, Meisam Razaviyayn, and Vahab Mirrokni. DiSK: Differentially private optimizer with simplified kalman filter for noise reduction. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c7793beb7f32b559df55d48370f2b8ae-Abstract-Conference.html.

A PROOFS

A.1 RESIDUAL IDENTITY AND BOUND

We use the means and dispersions defined in Eqs. (10) and (11).

Proof of Proposition 1. Expanding R and using $\sum_{i\in S_t}(\lambda_i - \bar{\lambda}) = 0$ gives

$$R = \frac{1}{m}\sum_{i\in S_t} \lambda_i G_i - \bar{\lambda} G_{\text{avg}} = G_{\text{clip}} - \bar{\lambda} G_{\text{avg}}.$$

The triangle inequality and Cauchy–Schwarz yield

$$\|R\|_{\text{F}} \le \frac{1}{m}\sum_{i\in S_t} |\lambda_i - \bar{\lambda}|\ \|G_i - G_{\text{avg}}\|_{\text{F}} \le \sigma_\lambda \sigma_G. \qquad \square$$

A.2 NEWTON–SCHULZ POLAR-FACTOR PRESERVATION

Recall $p(\sigma) = a\sigma + b\sigma^3 + c\sigma^5$, the scalar polynomial associated with Eq. (13).

Proof of Proposition 2. Completing the square gives $p(\sigma) = \sigma\left[c(\sigma^2 + b/2c)^2 + (4ac - b^2)/4c\right] > 0$ for all $\sigma > 0$. Starting from a reduced SVD $X = U\Sigma V^\top$, the normalization and each step replace Σ by another positive diagonal matrix while keeping U and V fixed. The new diagonal entries need not be in decreasing order, but sorting them only permutes the columns of U and V jointly, which leaves $UV^\top$ unchanged. Hence every iterate has a reduced SVD with polar factor $UV^\top$, and by the definition in Section 2.4, $\text{polar}(\text{Newton-Schulz}_J(X)) = \text{polar}(X)$. $\square$

B ILLUSTRATIVE EXAMPLE OF CLIPPING DISTORTION

Consider two per-example gradient matrices:

$$G_1 = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}, \qquad G_2 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

With clipping radius $C = \sqrt{2}$, their clipping factors are 0.1 and 1. The means before and after clipping are

$$G_{\text{avg}} = \begin{bmatrix} 5 & -0.5 \\ 0.5 & 5 \end{bmatrix}, \qquad G_{\text{clip}} = \begin{bmatrix} 0.5 & -0.5 \\ 0.5 & 0.5 \end{bmatrix}.$$

The mean clipping factor is $\bar{\lambda} = 0.55$, so the residual is

$$R = G_{\text{clip}} - 0.55 G_{\text{avg}} = \begin{bmatrix} -2.25 & -0.225 \\ 0.225 & -2.25 \end{bmatrix}.$$

Here $\sigma_\lambda = 0.45$ and $\sigma_G = \sqrt{202}/2$, giving

$$\|R\|_{\text{F}} = 0.225\sqrt{202} = \sigma_\lambda \sigma_G \approx 3.198.$$

The bound is attained in this two-example construction, but it need not be tight in general. More importantly, G_{clip} is not a scalar multiple of G_{avg}: clipping changes the aggregate polar factor even though it preserves the polar factor of either example.

In the metrics of Section 2.4, the resulting errors are

$$d_{\cos}(G_{\text{clip}}, G_{\text{avg}}) = 1 - \frac{11}{\sqrt{202}} \approx 0.226, \qquad d_{\text{polar}}(G_{\text{clip}}, G_{\text{avg}}) = 2\sqrt{1 - \frac{11}{\sqrt{202}}} \approx 0.951.$$

In this two-dimensional special case, the polar factors happen to be planar rotations by 5.7° and 45°. In higher dimensions, the polar factor instead describes a partial isometry between the right and left singular subspaces.

C UTILITY EXPERIMENT DETAILS

Parameter groups and updates. Muon updates tensors with at least two dimensions, excluding token and position embeddings and the language-model head. AdamW updates those excluded parameters, vectors, and biases. For an $r \times c$ Muon matrix, the learning rate is the selected base rate times $0.2\sqrt{\max(r, c)}$ (Liu et al., 2025). The five Newton–Schulz steps use coefficients $(3.4445, -4.7750, 2.0315)$, with no extra output-level bias correction. AdamW retains standard moment bias correction and numerical $\varepsilon = 10^{-8}$. Training dropout is active; historical evaluations use a separate stochastic stream, and evaluation disables dropout.

Selected histories and learning rates. Table 3 gives the base learning rates for the utility runs. At $\epsilon \in \{1, 2, 4\}$, E2E uses $(K_M, \gamma_M) = (4, 0.5)$ and $(K_A, \gamma_A) = (3, 0.5)$. DART uses $(K_M, \gamma_M) = (3, 0.7)$ and $(K_A, \gamma_A) = (3, 0.5)$ at every budget. DP-Muon (No-IM) disables both histories. Both methods in this table use AdamW $\beta_1 = 0.9$. The E2E $\epsilon = 8$ campaign fixes $(K_M, K_A) = (4, 2)$ and $\gamma_M = \gamma_A = 0.5$ for IM; its baseline disables both histories. Both use Muon/AdamW learning rates $0.002/0.0035$, retained from the earlier validation search without retuning.

At E2E $\epsilon = 8$, both methods are trained afresh at $\delta = 1/84122$. Across seeds 37–39, IM gives 65.34 ± 0.39 BLEU and 67.60 ± 0.17 ROUGE-L, versus 64.74 ± 0.05 and 67.31 ± 0.25 for the baseline. Table 1 reports per-metric maxima: $65.65/67.80$ for IM and $64.79/67.58$ for the baseline, all from seed 39. Seed 37 informed the earlier search; seeds $38, 39$ retain the previously used confirmation identities.

Table 3: Validation-selected base learning rates for DP-Muon (No-IM) and DP-Muon-IM; each pair is (Muon, AdamW). Both methods retain the same standard optimizer moments.

Task	Target ϵ	DP-Muon (No-IM)	DP-Muon-IM
E2E	1	$(0.0016, 0.0028)$	$(0.0016, 0.0035)$
	2	$(0.0016, 0.0035)$	$(0.0016, 0.0035)$
	4	$(0.0016, 0.0035)$	$(0.0020, 0.0035)$
	8	$(0.0020, 0.0035)$	$(0.0020, 0.0035)$
DART	1	$(0.00105, 0.00168)$	$(0.00105, 0.00168)$
	2	$(0.0013125, 0.00168)$	$(0.0013125, 0.0021)$
	4	$(0.0013125, 0.0021)$	$(0.0013125, 0.0021)$
	8	$(0.0013125, 0.0021)$	$(0.0013125, 0.002688)$

Private release. E2E uses Poisson sampling with $q = 1/41$ for 410 releases; DART uses $q = 1/61$ for 610 releases. The implementation uses $\lfloor B \rfloor$ rather than B in Eq. (6), giving release denominators 1025 and 1027, respectively. These differ from the nominal logical batch size 1024. Clipping uses the full concatenated per-example query, and each logical update makes one Gaussian release, not one release per microbatch. Table 4 lists the verified per-run calibration.

Table 4: Per-run PRV calibration for our utility runs, with $\delta = 1/(2N)$ at every budget. Achieved ϵ is rounded.

Task	Target ϵ	Noise multiplier σ_{DP}	Achieved ϵ	δ
E2E	1	2.05078125	0.996290	$1/84122$
	2	1.2646484375	1.999911	$1/84122$
	4	0.889892578125	3.996703	$1/84122$
	8	0.67626953125	7.998976	$1/84122$
DART	1	1.767578125	0.993732	$1/125318$
	2	1.123046875	1.998849	$1/125318$
	4	0.8197021484375	3.993658	$1/125318$
	8	0.63720703125	7.995195	$1/125318$

DiSK settings. Each DiSK closure forms its temporal query. DiSK combines the base and extrapolated queries before clipping and noise, then filters the private release before Muon/AdamW. The filter uses κ times the previous state plus $(1 - \kappa)$ times the new input.

The single-seed KF rows in Table 1 form a separate joint ablation using learning rates $0.002/0.0021$, $\kappa = 0.5$, and DiSK extrapolation coefficient 0.8. The IM row retains the same histories. The IM setting was selected by validation loss from 30 new trials and one reused anchor. Both KF studies use $\sigma_{\mathrm{DP}} = 0.67626953125$, $\delta = 1/84122$, and achieved $\epsilon \approx 7.998976$. DiSK's extrapolation coefficient is distinct from either history decay.

C.1 Optimizer-group ablation

We switch history on or off separately for Muon and AdamW, fixing their base learning rates at 0.002 and 0.0035. Enabled histories use $(K_M, \gamma_M) = (4, 0.5)$ and $(K_A, \gamma_A) = (3, 0.5)$. All four arms retain Muon momentum 0.95, AdamW $\beta_1 = 0.9$ and the E2E $\epsilon = 4$ privacy calibration in Table 4.

Table 5: Optimizer-group ablation on E2E at target $\epsilon = 4$ and $\delta = 1/84122$. NLL is measured in nats per target token.

History enabled	Test NLL ↓	BLEU ↑	ROUGE-L ↑
Neither	1.2842	63.66	66.90
Muon only	1.2779	63.76	66.95
AdamW only	1.2686	64.37	67.12
Both	1.2586	65.30	67.47

D Geometric error across training

The corrected diagnostic uses consistent per-record mean target-token loss for current and historical gradients. It compares IM with a current-only query at the same state of each IM-trained trajectory, before noise and outer momentum. Both are compared with the shared finite raw-current reference described in Section 5.3. This is not a comparison of independently trained IM and baseline models.

Runs use E2E, GPT-2, batch size 8, learning rate 5×10^{-5}, Muon $(K, \gamma) = (4, 0.9)$, and no AdamW history. Full-query clipping uses $C = 1$. The three training seeds are $42, 43, 44$; probes occur before updates $10, 50, 100, 250, 500, 1000, 2000$. Thus index 10 means nine completed updates. Means accumulate in FP32 and polar factors use CPU FP64 SVD.

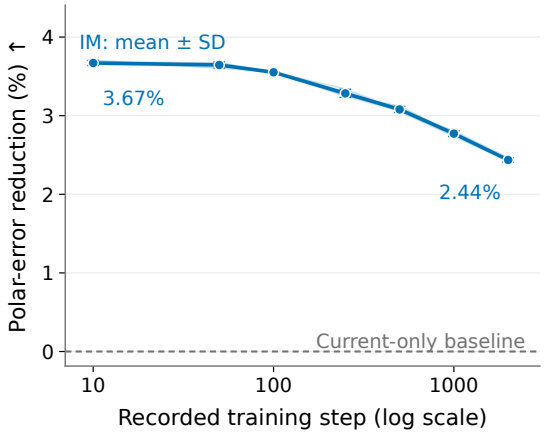

Figure 3: Corrected three-seed polar diagnostic at $(K, \gamma) = (4, 0.9)$. Relative reduction is computed within each seed as $100(1 - E_{\mathrm{IM}}/E_{\mathrm{current}})$, then averaged. Faint lines show individual seeds; bars and shading are pointwise sample SD, not confidence intervals.

Both cosine and polar error are lower with IM in all 21 seed–checkpoint pairs. Mean polar-error reduction decreases from 3.67% before update 10 to 2.44% before update 2000 (Figure 3); mean

cosine-error reductions range from 3.20% to 5.98%. These distances are relative to a finite reference, not a population gradient or the noisy Newton–Schulz update.

The one-seed K sweep in Figure 2 uses the same raw-query diagnostic design at index 10. It pairs IM and current-only queries within each trajectory; the $K = 4$ point reuses the corrected seed-42 run above. The $K = 0$–3 points are completed early probes, not final-training results. The two panels of that figure therefore differ in history groups, normalization, seed, batch size, and training horizon.

D.1 Normalized clipping residual

We measure the clipping residual on the same $m = 512$ fixed held-out query records at each checkpoint. Let $H_W^{(i)}$ be record i's query for Muon block W: a history average for IM and a single current-state evaluation for the control. Write $G_{\mathrm{avg},W} = m^{-1}\sum_i H_W^{(i)}$ and $G_{\mathrm{clip},W} = m^{-1}\sum_i \lambda_i H_W^{(i)}$. Clipping factors come from the complete raw query at $C = 1$. We report

$$R_W = G_{\mathrm{clip},W} - \bar{\lambda} G_{\mathrm{avg},W}, \qquad \eta_R = \frac{\left(\sum_{W\in\mathcal{W}} \|R_W\|_{\mathrm{F}}^2\right)^{1/2}}{\bar{\lambda}\left(\sum_{W\in\mathcal{W}} \|G_{\mathrm{avg},W}\|_{\mathrm{F}}^2\right)^{1/2}}. \tag{20}$$

The norms aggregate the same 48 Muon matrices. Each arm uses its own mean clipping factor $\bar{\lambda}$ and unclipped query mean in the denominator. These measurements precede Gaussian noise and outer momentum.

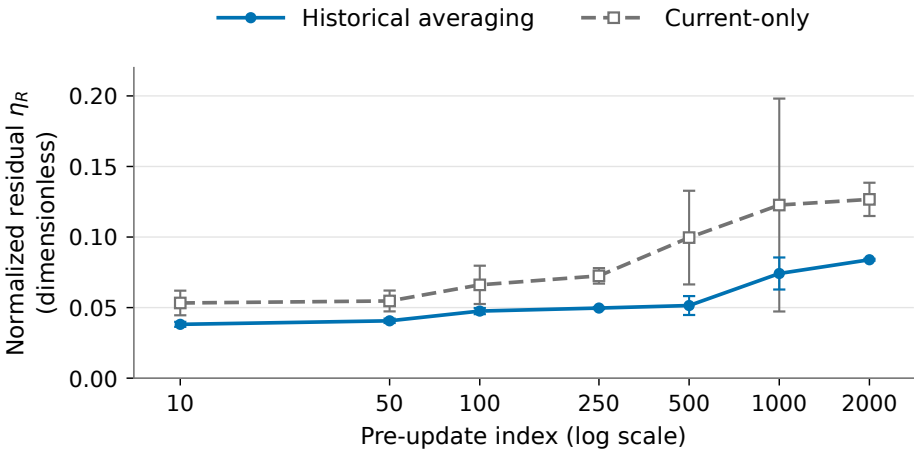

Figure 4: Normalized clipping residual η_R on E2E for historical averaging ($K = 4$, $\gamma = 0.9$) and a single current-state evaluation at the same model states. Points show means over training seeds 42–44; bars show sample standard deviation, not confidence intervals. All seven pre-update checkpoints are shown. Clipping uses the complete raw query at $C = 1$; the residual metric covers Muon coordinates.

Historical averaging gives a smaller η_R in all 21 seed–checkpoint pairs (Figure 4). We compute the percentage reduction within each seed as $100(1 - \eta_{R,\mathrm{IM}}/\eta_{R,\mathrm{current}})$, then average over seeds. These means range from 24.99% to 46.10% across checkpoints.

E Privacy and implementation conditions

Fix a public initialization and public values of the sampling rate $q \in (0, 1]$, clipping radius $C > 0$, noise multiplier $\sigma_{\mathrm{DP}} > 0$, release denominator $b > 0$, and update count T. Current and stored model states and optimizer states must be computed only from preceding noisy releases and public settings. At each update, sample every record independently with probability q, independently of previous sampling draws. Sampling decisions remain internal. The accountant must bound adaptive composition under add/remove adjacency.

Complete-query mechanism. Write $\mathcal{T}_{t-1} = (\widetilde{G}_0, \ldots, \widetilde{G}_{t-1})$ for the preceding noisy releases, with $\mathcal{T}_{-1}$ empty. Given these releases, let $u_t(x; \mathcal{T}_{t-1}, \omega) \in \mathbb{R}^d$ be record x's query over all trainable coordinates. Here ω is fresh data-independent randomness, drawn independently for each record and update and independently of sampling, Gaussian noise, and preceding releases. Historical evaluations within one record's query may share randomness. The query may combine different Muon and AdamW histories and apply per-record preprocessing. It must not use unaccounted private statistics from other records. Define

$$v_t(x; \mathcal{T}_{t-1}, \omega) = \mathrm{clip}_C\big(u_t(x; \mathcal{T}_{t-1}, \omega)\big), \qquad \|v_t(x; \mathcal{T}_{t-1}, \omega)\|_2 \le C. \tag{21}$$

The complete release at step t is

$$\widetilde{G}_t = \frac{1}{b}\left(\sum_{x \in S_t} v_t(x; \mathcal{T}_{t-1}, \omega_{t,x}) + \sigma_{\mathrm{DP}} C Z_t\right), \qquad Z_t \sim \mathcal{N}(0, I_d), \tag{22}$$

where $b > 0$ is fixed and public. Gaussian draws are independent across updates and independent of sampling and query randomness. Algorithm 1 displays one Muon block of this release with $b = B$; the utility implementation uses $b = \lfloor B \rfloor$. Neither choice changes the privacy bound when the denominator is held fixed across neighboring datasets. Independent Gaussian draws for individual coordinates or blocks implement the same joint release; they are not separately clipped mechanisms.

Privacy argument. Consider neighboring datasets $\mathcal{D}$ and $\mathcal{D}' = \mathcal{D} \cup \{x_\star\}$. Fix a common value of $\mathcal{T}_{t-1}$. With the public initialization and settings fixed, the current model and stored histories are the same in both conditional executions. This does not assert that their unconditional training trajectories coincide.

For every common record, the conditional query law in Eq. (21) is the same on both datasets. Couple the sampling and local randomness of these common records. The additional record is selected independently with probability q and contributes a vector of norm at most C. Thus the conditional update is a Poisson-sampled Gaussian mechanism with noise multiplier σ_{DP}: the add/remove sensitivity of the unsampled sum is at most C, irrespective of how many historical gradients were used within each query. Division by public b is post-processing.

This per-step privacy bound holds uniformly over preceding transcripts. Adaptive composition gives the complete transcript $(\widetilde{G}_0, \ldots, \widetilde{G}_{T-1})$ the (ϵ, δ) guarantee certified by the accountant (Abadi et al., 2016; Gopi et al., 2021). Updating the history, applying Muon momentum and Newton–Schulz, and applying AdamW are post-processing of these releases. The final model and any checkpoints computed solely from the transcript inherit the same guarantee (Dwork and Roth, 2014). □

Calibration and output scope. The privacy analysis holds calibration fixed across neighboring datasets. Any reference dataset size used to choose q, T, δ, or b must therefore be a public design value, not recomputed from the private dataset's cardinality. The guarantee uses add/remove adjacency; replace-one adjacency has an unsampled sensitivity bound of $2C$ and requires the corresponding accounting.

This is a per-run guarantee for fixed public settings, not a privacy bound for the full experimental search. Private validation, hyperparameter selection, additional training runs, or a privately trained initialization require separate accounting. Releasing raw gradients, sampling decisions, or additional data-dependent diagnostics is not covered by this guarantee. In particular, the geometric probes' diagnostic releases and non-Poisson training protocol are not covered by the utility runs' reported budgets. This privacy argument requires no smoothness, gradient-decorrelation, or polar-rank assumptions.

F AVERAGING AND POLAR ERROR

Proposition 1 in Section 3 separates the clipped mean into a common rescaling and a residual caused by unequal clipping. Here we use this decomposition to bound polar error relative to the current population gradient, accounting for sampling error and the use of historical gradients. The analysis concerns the clipped mean before Gaussian noise and outer momentum, not the final Muon update.

Setup. Fix model states $\theta_0, \dots, \theta_K$ and one Muon block $W \in \mathbb{R}^{r\times c}$, where θ_0 is the current state. For an example z, let $G_j(z)$ be its raw gradient for this block at state j, and define $\mu_j = \mathbb{E}G_j(z)$ and $\xi_j(z) = G_j(z) - \mu_j$. Expectations include any local evaluation randomness, with the same conditional law given z at every state. Draw $m \geq 2$ examples independently and reuse each example across states. Each example and all its evaluation draws form one independent, identically distributed tuple; the states and weights are fixed independently of these tuples.

With fixed weights $\alpha_j \geq 0$, $\sum_{j=0}^{K} \alpha_j = 1$, shared across the block's coordinates, set

$$H^{(i)} = \sum_{j=0}^{K} \alpha_j G_j(z_i), \qquad N = \frac{1}{m}\sum_{i=1}^{m} \lambda_i H^{(i)}, \qquad S_K = \sum_{j=0}^{K} \alpha_j^2. \tag{23}$$

As in Section 3, $\lambda_i > 0$ is computed by flat clipping of example i's complete raw query, not by clipping this block separately. Write $\bar{\lambda} = m^{-1}\sum_i \lambda_i$ and $\sigma_\lambda^2 = m^{-1}\sum_i (\lambda_i - \bar{\lambda})^2$. The current-only query has $K = 0$ and $S_0 = 1$; equal weights give $S_K = 1/(K+1)$.

We assume, for fixed constants $\tau_W \geq 0$, $\rho \in [0,1]$, and $c_\lambda \geq 0$,

1. $\mathbb{E}\,\|\xi_j\|_{\mathrm{F}}^2 \leq \tau_W^2$ for every j;
2. $\mathbb{E}\,\langle \xi_j, \xi_\ell \rangle_{\mathrm{F}} \leq \rho\tau_W^2$ for $j \neq \ell$, evaluated on the same example;
3. $\sigma_\lambda \leq c_\lambda \bar{\lambda}$ almost surely.

Here τ_W bounds fluctuations of block W, whereas τ_u in Section 4.2 bounds fluctuations of the full query. The blockwise covariance bound ρ here need not equal the full-query value used there. Assumption (1) is the blockwise version of Assumption 3.1 in Xiao et al. (2023). Assumption (2) allows temporal correlation: independent centered fluctuations satisfy $\rho = 0$, but independence is not required. The choice $\rho = 1$ follows from (1) by Cauchy–Schwarz and does not require identical fluctuations. Assumption (3) and the full-rank reference condition below are additional assumptions, not results of Xiao et al. (2023).

Polar perturbation. For square nonsingular inputs, the following inequality follows from classical polar-factor perturbation bounds (Li, 1995). The proof below also covers rectangular matrices and rank-deficient X using the partial-polar convention in Section 2.4.

Lemma 5 (Polar perturbation with a full-rank reference)**.** *Let $T, X \in \mathbb{R}^{r\times c}$ with $\mathrm{rank}(T) = \min(r,c)$. Then*

$$d_{\mathrm{polar}}(X,T) \leq \frac{2\,\|X - T\|_{\mathrm{F}}}{\sigma_{\min}(T)}. \tag{24}$$

Here $\sigma_{\min}(T) > 0$ is the smallest of the $\min(r,c)$ singular values of T.

Proof. First suppose $r \geq c$. Let $Q = \mathrm{polar}(T)$, $U = \mathrm{polar}(X)$, and write $T = QP$ with $P = (T^\top T)^{1/2} \succeq sI_c$, where $s = \sigma_{\min}(T) > 0$. Since $Q^\top Q = I_c$ and $U^\top U \preceq I_c$,

$$A = I_c - \tfrac{1}{2}(Q^\top U + U^\top Q) \succeq 0, \qquad \|Q - U\|_{\mathrm{F}}^2 \leq 2\,\mathrm{tr}(A).$$

The polar factor U, including when X is rank deficient, maximizes $\langle X, V\rangle_{\mathrm{F}}$ over $\|V\|_{\mathrm{op}} \leq 1$. As Q is feasible, $\langle X, Q - U\rangle_{\mathrm{F}} \leq 0$. Hence

$$\frac{s}{2}\,\|Q - U\|_{\mathrm{F}}^2 \leq \mathrm{tr}(PA) = \langle T, Q - U\rangle_{\mathrm{F}} \leq \langle T - X, Q - U\rangle_{\mathrm{F}} \leq \|T - X\|_{\mathrm{F}}\,\|Q - U\|_{\mathrm{F}}\,.$$

Dividing by $\|Q - U\|_{\mathrm{F}}$ proves the claim for $r \geq c$ when it is nonzero; the zero case is immediate. For $r < c$, apply this result to $T^\top$ and $X^\top$. Transposition preserves Frobenius norms and singular values, and $\mathrm{polar}(X^\top) = \mathrm{polar}(X)^\top$. □

Theorem 6 (Polar-error bound under averaging)**.** *Under Assumptions (1)–(3), suppose $\mu_0 \in \mathbb{R}^{r\times c}$ has rank $\min(r,c)$. Define*

$$s_{K,W} = \tau_W\sqrt{\rho + (1-\rho)S_K}, \qquad D_K = \left\| \sum_{j=0}^{K} \alpha_j \mu_j - \mu_0 \right\|_{\mathrm{F}}. \tag{25}$$

Then

$$\mathbb{E}\, d_{\mathrm{polar}}(N, \mu_0) \le \frac{2}{\sigma_{\min}(\mu_0)} \left(D_K + \frac{s_{K,W}}{\sqrt{m}} + c_\lambda \sqrt{\frac{m-1}{m}}\, s_{K,W} \right). \tag{26}$$

The three terms account for staleness, sampling error, and unequal clipping, respectively. No rank condition is imposed on N.

Proof. Let $Y_i = \sum_j \alpha_j \xi_j(z_i)$ and $\bar{Y} = m^{-1} \sum_i Y_i$. Applying Proposition 1 with $G_i = H^{(i)}$ gives

$$\frac{N}{\bar{\lambda}} - \mu_0 = \left(\sum_j \alpha_j \mu_j - \mu_0 \right) + \bar{Y} + \frac{R}{\bar{\lambda}}. \tag{27}$$

The first term has norm D_K. For the second, Assumptions (1)–(2) and the nonnegative weights imply

$$\mathbb{E}\, \|Y_i\|_{\mathrm{F}}^2 \le \tau_W^2 \sum_j \alpha_j^2 + \rho \tau_W^2 \sum_{j \ne \ell} \alpha_j \alpha_\ell = \tau_W^2 \big(S_K + \rho(1 - S_K) \big) = s_{K,W}^2.$$

The Y_i are independent with mean zero, so $\mathbb{E}\, \big\|\bar{Y}\big\|_{\mathrm{F}} \le s_{K,W}/\sqrt{m}$. For the last term, write $\sigma_H^2 = m^{-1} \sum_i \big\|H^{(i)} - \bar{H}\big\|_{\mathrm{F}}^2$. Proposition 1 and (3) give

$$\frac{\|R\|_{\mathrm{F}}}{\bar{\lambda}} \le c_\lambda \sigma_H, \qquad \mathbb{E}\sigma_H \le \sqrt{\mathbb{E}\sigma_H^2} = \sqrt{\frac{m-1}{m} \mathbb{E}\, \|Y_i\|_{\mathrm{F}}^2} \le \sqrt{\frac{m-1}{m}}\, s_{K,W}.$$

Finally, $\mathrm{polar}(N) = \mathrm{polar}(N/\bar{\lambda})$ because $\bar{\lambda} > 0$. Taking norms and expectations in Eq. (27) and applying Lemma 5 proves Eq. (26). □

When the bound improves. Without averaging, $D_0 = 0$ and $s_{0,W} = \tau_W$. For $\tau_W > 0$, averaging gives $s_{K,W} < \tau_W$ when $\rho < 1$ and at least two weights are nonzero ($S_K < 1$). At the same current state and batch size, use one constant c_λ satisfying (3) for both the averaged and current-only queries. With $a_m = m^{-1/2} + c_\lambda \sqrt{(m-1)/m}$, the right-hand side of Eq. (26) is no larger than the current-only bound whenever

$$D_K \le a_m (\tau_W - s_{K,W}). \tag{28}$$

Strict inequality gives a strictly smaller bound. This compares upper bounds, not actual errors. When $\rho = 1$, the variance terms in this bound do not decrease.

If each per-example loss is β-smooth in the full parameter vector, uniformly over the local randomness, then the block mean is β-Lipschitz and

$$D_K \le \beta \sum_{j=0}^{K} \alpha_j \|\theta_j - \theta_0\|_2. \tag{29}$$

This is the smoothness condition used by Xiao et al. (2023), with the common conditional evaluation-randomness law specified above. Nearby states control staleness, but do not by themselves imply temporal decorrelation.

Proof of Proposition 3. Use the full-query gradients and moment assumptions with τ_u from Section 4.2. Let $u_i = \sum_j \alpha_j g_j(z_i)$, $Y_{i,u} = \sum_j \alpha_j \xi_{j,u}(z_i)$, and $\sigma_u^2 = m^{-1} \sum_i \|u_i - \bar{u}\|_2^2$. The map $n \mapsto \min\{1, C/n\}$ is $1/C$-Lipschitz: its derivative is 0 for $n < C$ and $-C/n^2 \in [-1/C, 0)$ for $n > C$. The variance of the λ_i is at most their second moment about $\lambda(\|\bar{u}\|)$, and $|\, \|u_i\| - \|\bar{u}\|\, | \le \|u_i - \bar{u}\|$, so $\sigma_\lambda \le \sigma_u / C$. Proposition 1 gives $\|R_W\|_F \le \sigma_\lambda \sigma_{H,W}$, and $\sigma_{H,W} \le \sigma_u$ because a block is a sub-collection of coordinates. Hence $\|R_W\|_F \le \sigma_u^2 / C$. The full-query moment bounds give $\mathbb{E}\|Y_{i,u}\|_2^2 \le s_{K,u}^2$. Since the means $\sum_j \alpha_j \mu_{j,u}$ cancel in $u_i - \bar{u}$, the same sample-variance identity used in Theorem 6 yields $\mathbb{E}\, \sigma_u^2 = \frac{m-1}{m} \mathbb{E}\|Y_{1,u}\|_2^2 \le \frac{m-1}{m} s_{K,u}^2$. □

If the full-query moment bounds in Section 4.2 hold with equality and $\tau_u > 0$, then $\mathbb{E}\, \sigma_u^2 = \frac{m-1}{m} s_{K,u}^2$ exactly. The ratio of the expected empirical variances for averaged and current-only queries is then $\rho + (1 - \rho) S_K$.