# DP-MUON: DIFFERENTIALLY PRIVATE OPTIMIZATION VIA MATRIX-ORTHOGONALIZED MOMENTUM

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We study differentially private (DP) training with Muon, a matrix-valued optimizer that updates hidden-layer weights using momentum followed by Newton-Schulz orthogonalization. While DP-SGD is well understood, the interaction between per-example clipping, Gaussian noise, momentum, and nonlinear orthogonalization in Muon has not been systematically analyzed. We formulate DP-Muon, a private Muon procedure that clips per-example gradients globally, adds Gaussian noise to the clipped minibatch average, and then applies momentum and Newton-Schulz orthogonalization as post-processing. We prove that DP-Muon inherits the privacy guarantee certified by standard subsampled-Gaussian accounting, with no additional privacy cost from Muon-specific post-processing. On the optimization side, we establish finite-horizon and vanishing stationarity guarantees under global clipping, with bounds that separate optimization error, clipping residual, privacy noise, and Newton-Schulz approximation error. We further show that Gaussian privacy noise, although centered before the nonlinear Newton-Schulz map, induces a systematic bias in the expected update after this nonlinear transformation. This motivates DP-MuonBC, an adaptive bias-corrected variant designed to cancel the leading bias induced by the current Gaussian perturbation while preserving the same privacy guarantee. Experiments on E2E at four privacy budgets show that the tested Muon configurations improve test NLL over Adam baselines, with further gains for the selected DP-MuonBC configuration.

## 1 INTRODUCTION

Differential privacy (DP) (Dwork et al., 2006; Dwork & Roth, 2014) has become a standard framework for training models on sensitive data, with DP-SGD and its variants serving as the dominant baseline in modern deep learning (Abadi et al., 2016; Song et al., 2013; Bassily et al., 2014; Mironov, 2017; Wang et al., 2019). DP has also seen broad deployment in industry, including applications to telemetry collection, browser privacy, and official statistics (Úlfar Erlingsson et al., 2014; D. P. T. Apple, 2017; Ding et al., 2017; Abowd, 2018). The basic recipe of DP is by now familiar: sample a minibatch, clip per-example gradients to control sensitivity, add Gaussian noise, and compose the privacy loss over training steps using an accountant.

DP-SGD has been extended in many directions, including adaptive and preconditioned methods (Koskela & Honkela, 2020; Zhou et al., 2021), low-dimensional and parameter-efficient updates (Yu et al., 2021a;b), and clipping-aware or variance-reduced variants (Wang et al., 2017; Andrew et al., 2021; Pichapati et al., 2019; Li et al., 2022). Recent analysis further shows that, without explicit bias correction, DP-Adam can collapse toward a rescaled momentum form of DP-SGD under privacy noise (Tang et al., 2024). By contrast, much less is known for private optimizers that exploit matrix structure.

A recent example is *Muon* (Jordan et al., 2024), an optimizer for hidden-layer matrix parameters that replaces the raw update direction with an orthogonalized momentum matrix, typically computed in practice using a short Newton-Schulz iteration. Empirically, this practical SVD-free implementation performs strongly at scale, making Muon a promising alternative to conventional vector-based optimizers (Jordan et al., 2024; Liu et al., 2025). On the theory side, several recent works have begun to analyze the convergence behavior of Muon (Shen et al., 2025; Sato et al., 2025; Kim & Oh, 2026). On the privacy side, however, the picture is even less developed.



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.

This paper studies a natural *DP-Muon* procedure that adds Gaussian noise to the globally clipped per-example average before the momentum and Newton–Schulz steps. This ordering has two advantages. First, it retains the standard DP-SGD accounting structure: clipping controls the sensitivity of the released lot average, Gaussian perturbation gives a private per-step mechanism, and privacy amplification and composition yield the full-run guarantee (Abadi et al., 2016; Balle et al., 2018; Wang et al., 2019). Second, it preserves the defining structure of practical Muon, in which a matrix momentum buffer is transformed by a finite Newton–Schulz iteration, rather than replacing the update with an exact SVD-based polar factor.

Our privacy mechanism applies a single global clipping operation to the full per-example gradient, including both Muon-managed matrix blocks and auxiliary parameters, followed by one joint Gaussian release. For this joint mechanism, we establish a privacy guarantee under standard sampled-Gaussian accounting whose parameters depend on the sampling rule, number of steps, and noise multiplier, but not explicitly on the ambient parameter dimension. We then analyze a Muon-managed matrix block and establish finite-horizon and vanishing stationarity guarantees for its noisy Newton–Schulz update. The optimization bounds isolate two quantities that are especially relevant in the private setting: an explicit *clipping residual*, which captures the bias introduced by global clipping, and a *Newton–Schulz quality parameter*, which captures the approximation error from using a finite number of orthogonalization steps.

**Our contributions.** First, we formulate DP-Muon using global per-example clipping and a joint Gaussian release, and prove that its subsequent Muon and auxiliary optimizer updates incur no additional privacy cost. Second, we characterize the output bias induced by applying Newton–Schulz to noisy momentum and introduce DP-MuonBC, for which an appropriate correction coefficient cancels the leading fresh-noise conditional bias. Third, we establish finite-horizon and vanishing stationarity guarantees for DP-Muon and DP-MuonBC on a Muon-managed matrix block, and identify a regime in which bias correction improves the matched stationarity certificate. Finally, experiments on E2E compare these methods under the same global clipping mechanism and four privacy budgets, with the tested Muon configurations achieving lower test NLL than the Adam baselines.

**Organization.** Section 2 defines DP-Muon and establishes its privacy guarantee via joint sampled-Gaussian accounting. Section 3 introduces DP-MuonBC and analyzes the conditional output bias induced by applying Newton–Schulz to noisy momentum. Section 4 establishes finite-horizon and vanishing stationarity guarantees for DP-Muon and DP-MuonBC under global clipping and gives a regime in which bias correction improves the matched stationarity certificate. Section 5 reports experiments on E2E.

## 2 DP-MUON

**Setting and sampling.** Let $\mathcal{D} = \{z_i\}_{i=1}^N$ be the dataset, and let $\mathcal{B}_t$ denote the lot sampled at step $t$ according to a fixed public sampling rule. Under fixed-size sampling, $|\mathcal{B}_t| = B$ and $\gamma := B/N$.

Under Poisson sampling, $\gamma$ denotes the inclusion probability and $B$ is a fixed normalization constant, typically $B = \gamma N$, rather than the realized lot size $|\mathcal{B}_t|$. The privacy accountant must correspond to the actual sampling rule, adjacency notion, noise scale, and number of iterations, with the sampling parameters and normalization held fixed across adjacent datasets.

Let $\mathcal{W}$ denote the set of hidden-layer matrix parameters updated by Muon. Parameters outside $\mathcal{W}$ are omitted from the matrix notation but are included in the same global per-example clipping operation and joint Gaussian release. Their auxiliary optimizer uses the corresponding coordinates of this privatized release.

**Global clipping and joint release.** Choose a clipping threshold $C > 0$. For a full per-example gradient $G$, define

$$\mathrm{clip}_C(G) := \frac{G}{\max\{1, \|G\|_2/C\}}.$$

**Algorithm 1** DP-Muon with global clipping and joint accounting

**Require:** Dataset $\mathcal{D} = \{z_i\}_{i=1}^N$, iterations $T$, normalization $B$.
**Require:** Muon matrices $\mathcal{W}$, global clipping bound $C > 0$, noise multiplier $\sigma > 0$.
**Require:** Learning rates $\{\eta_t\}_{t=1}^T$, momentum $\beta \in [0, 1)$.
**Require:** Newton-Schulz steps $q \geq 1$, polynomial $p_\kappa$ from equation 3, $\kappa \geq 1$, smooth regularizer $\gamma_{\text{ns}} > 0$.

Initialize $M_0(W) \leftarrow 0$ for all $W \in \mathcal{W}$.
**for** $t = 1, \ldots, T$ **do**
  Sample $\mathcal{B}_t$ using the public sampling rule.
  Compute full per-sample gradients $G_t(z) \leftarrow \nabla_\theta \ell(\theta_{t-1}; z)$, $z \in \mathcal{B}_t$.
  Clip

$$H_t(z) \leftarrow \frac{G_t(z)}{\max\{1, \|G_t(z)\|_2 / C\}}, \qquad z \in \mathcal{B}_t.$$

  Release

$$\widetilde{G}_t \leftarrow \frac{1}{B} \sum_{z \in \mathcal{B}_t} H_t(z) + Z_t, \qquad Z_t \sim \mathcal{N}\left(0, \frac{\sigma^2 C^2}{B^2} I_{d_{\text{all}}}\right).$$

  Set $\mathcal{R}_t \leftarrow \widetilde{G}_t$ and extract $\widetilde{G}_t(W)$ for each $W \in \mathcal{W}$.
  **for** each $W \in \mathcal{W}$ **do**
    $M_t(W) \leftarrow \beta M_{t-1}(W) + \widetilde{G}_t(W)$.

$$Y_{t,0}(W) \leftarrow \frac{\mathcal{T}_W(M_t(W))}{\alpha_t(W)}, \qquad \alpha_t(W) \leftarrow \sqrt{\gamma_{\text{ns}}^2 + \|\mathcal{T}_W(M_t(W))\|_F^2}.$$

    **for** $j = 1, \ldots, q$ **do**

$$Y_{t,j}(W) \leftarrow p_\kappa\big(Y_{t,j-1}(W) Y_{t,j-1}(W)^\top\big)\, Y_{t,j-1}(W).$$

    **end for**
    $O_t(W) \leftarrow \mathcal{T}_W^{-1}\big(Y_{t,q}(W)\big)$.
  **end for**
  Simultaneously update, for all $W \in \mathcal{W}$,

$$W_t \leftarrow W_{t-1} - \eta_t O_t(W).$$

  Update auxiliary parameters using their coordinates of $\widetilde{G}_t$ and private optimizer state.
**end for**

At step $t$, for each $z \in \mathcal{B}_t$, let

$$G_t(z) := \nabla_\theta \ell(\theta_{t-1}; z), \qquad H_t(z) := \text{clip}_C\big(G_t(z)\big).$$

The clipping norm is taken over all trainable coordinates, including those updated by the auxiliary optimizer. For a Muon-managed matrix block $W$, let $G_t(z; W)$ and $H_t(z; W)$ denote the corresponding block coordinates. Since $\|H_t(z)\|_2 \leq C$, DP-Muon forms a single joint Gaussian release

$$\widetilde{G}_t = \frac{1}{B} \sum_{z \in \mathcal{B}_t} H_t(z) + Z_t, \qquad Z_t \sim \mathcal{N}\left(0, \frac{\sigma^2 C^2}{B^2} I_{d_{\text{all}}}\right),$$

where $d_{\text{all}}$ is the total number of trainable coordinates and the noises $Z_t$ are independent across steps. We write $\mathcal{R}_t := \widetilde{G}_t$ for the full per-step release and $\widetilde{G}_t(W)$ for its coordinates associated with block $W$. For each example, the same clipping multiplier is applied across all parameter blocks, and all blocks are obtained from the same sampled lot and joint release. The base sensitivity of this joint query is $C/B$ under add/remove adjacency and $2C/B$ under replace-one adjacency; the privacy accountant must use the corresponding adjacency model.

**Muon post-processing.** After forming $\mathcal{R}_t$, DP-Muon updates the matrix momentum buffers and applies a finite Newton–Schulz iteration to obtain the Muon update directions. Because these operations depend only on the privatized release and public hyperparameters, they incur no additional privacy cost by post-processing. For notational convenience, for a matrix block $W$ of shape $m_W \times n_W$, let $\mathcal{T}_W(A) = A^\top$ when $m_W > n_W$ and $\mathcal{T}_W(A) = A$ otherwise; $\mathcal{T}_W^{-1}$ applies the corresponding reverse convention.

### 2.1 Algorithm

Algorithm 1 summarizes the complete DP-Muon procedure. Each iteration forms a single joint Gaussian release from the globally clipped per-example gradients, after which the momentum, Newton–Schulz, and parameter updates are post-processing of the privatized release.

### 2.2 Privacy Guarantee via Joint Accounting

Let $\mathcal{R}_{1:T} := (\mathcal{R}_1, \dots, \mathcal{R}_T)$ be the full noisy gradient transcript. The accountant is applied to the mechanism $D \mapsto \mathcal{R}_{1:T}$, not to a single matrix block in isolation. It must account for the actual globally clipped joint Gaussian release at each step.

**Theorem 2.1** (Privacy of DP-Muon)**.** *Assume add/remove adjacency. At each step $t$, clip the full per-example gradients to Euclidean norm $C$ and release*

$$\widetilde{G}_t = \frac{1}{B} \sum_{z \in \mathcal{B}_t} H_t(z) + Z_t, \qquad Z_t \sim \mathcal{N}\left(0, \frac{\sigma^2 C^2}{B^2} I_{d_{\text{all}}}\right).$$

*Let $\mathcal{R}_t = \widetilde{G}_t$ be the resulting joint Gaussian release. If a valid accountant for the actual sampling rule and this joint transcript certifies $D \mapsto \mathcal{R}_{1:T}$ as $(\varepsilon, \delta)$-differentially private, then Algorithm 1, including the final parameters and any released intermediate Muon quantities that are post-processing of the transcript, is also $(\varepsilon, \delta)$-differentially private.*

*Non-Muon parameter updates using the same release incur no additional privacy cost. For replace-one adjacency, the base sensitivity is $2C/B$, and a valid accountant for that sampling and adjacency model is required.*

The proof follows from global sensitivity, valid accounting of the sampled joint Gaussian transcript, adaptive composition, and post-processing; see Appendix A.

## 3 DP-MuonBC: Output-Level Bias Correction

In DP-Muon, Gaussian privacy noise is added to the clipped minibatch average before the momentum and Newton-Schulz steps. The momentum recursion is linear, so the injected Gaussian perturbation remains additive at the momentum-buffer level. The nontrivial bias appears only after applying the nonlinear Newton-Schulz map that produces the Muon direction.

This section isolates this output-level effect and introduces DP-MuonBC, a bias-corrected post-processing variant of DP-Muon. Algorithm 2 summarizes the complete DP-MuonBC procedure. We condition on the realized training history immediately before the current Gaussian noise is drawn. Only this fresh noise is independent of the current clipped pre-noise state; previous clipped gradients and iterates may depend on all previous privacy noises. The privacy guarantee remains exact because DP-MuonBC is a randomized post-processing of the privatized gradients.

**Adaptive clipped-gradient model.** Fix one Muon-managed matrix parameter with shape $m \times n$, and write $d = mn$. Let $\bar{G}_t$ be its globally clipped-but-unnoised minibatch mean. The privatized gradient at step $t$ is

$$\widetilde{G}_t = \bar{G}_t + Z_t, \qquad Z_t \sim \mathcal{N}(0, \nu^2 I_d), \qquad \nu^2 := \frac{\sigma^2 C^2}{B^2},$$

where matrices are vectorized when interpreting the Gaussian covariance, and the noises are independent across steps. Let $\mu = \beta$ be the momentum parameter and define the current pre-noise and

**Algorithm 2** DP-MuonBC with global clipping, antithetic output correction, and joint accounting

**Require:** Dataset $\mathcal{D} = \{z_i\}_{i=1}^N$, iterations $T$, normalization $B$.
**Require:** Muon matrices $\mathcal{W}$, global clipping bound $C > 0$, noise multiplier $\sigma > 0$.
**Require:** Learning rates $\{\eta_t\}_{t=1}^T$, momentum $\beta \in [0, 1)$, probe count $J \geq 1$, correction strength $\lambda \geq 0$.
**Require:** For each $W \in \mathcal{W}$, the smooth Newton-Schulz map $Q_{\kappa,q}^{(W)}$ induced by the transpose convention, polynomial $p_\kappa$, and regularizer $\gamma_{\mathrm{ns}}$ in Algorithm 1.

Initialize $M_0(W) \leftarrow 0$ for all $W \in \mathcal{W}$.
**for** $t = 1, \ldots, T$ **do**
  Sample $\mathcal{B}_t$ using the public sampling rule.
  Compute full per-sample gradients $G_t(z) \leftarrow \nabla_\theta \ell(\theta_{t-1}; z)$, $z \in \mathcal{B}_t$.
  Clip

$$H_t(z) \leftarrow \frac{G_t(z)}{\max\{1, \|G_t(z)\|_2 / C\}}, \qquad z \in \mathcal{B}_t.$$

  Release

$$\widetilde{G}_t \leftarrow \frac{1}{B} \sum_{z \in \mathcal{B}_t} H_t(z) + Z_t, \qquad Z_t \sim \mathcal{N}\left(0, \frac{\sigma^2 C^2}{B^2} I_{d_{\mathrm{all}}}\right).$$

  Set $\mathcal{R}_t \leftarrow \widetilde{G}_t$, set $\rho_t \leftarrow \sigma C / B$, and extract $\widetilde{G}_t(W)$ for each $W \in \mathcal{W}$.
  **for** each $W \in \mathcal{W}$ **do**
    $M_t(W) \leftarrow \beta M_{t-1}(W) + \widetilde{G}_t(W)$.
    $O_t^{(0)}(W) \leftarrow Q_{\kappa,q}^{(W)}(M_t(W))$.
    Draw data-independent probes $U_{t,1}(W), \ldots, U_{t,J}(W)$ with $\mathrm{vec}(U_{t,j}(W)) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{m_W n_W})$, independently of the history and privacy noises.

$$O_t^{(2)}(W) \leftarrow \frac{1}{2J} \sum_{j=1}^{J} \left[ Q_{\kappa,q}^{(W)}(M_t(W) + \rho_t U_{t,j}(W)) + Q_{\kappa,q}^{(W)}(M_t(W) - \rho_t U_{t,j}(W)) \right].$$

    $O_t^\lambda(W) \leftarrow (1 + \lambda) O_t^{(0)}(W) - \lambda O_t^{(2)}(W)$.
  **end for**
  Simultaneously update, for all $W \in \mathcal{W}$,

$$W_t \leftarrow W_{t-1} - \eta_t O_t^\lambda(W).$$

  Update auxiliary parameters using their coordinates of $\widetilde{G}_t$ and private optimizer state.
**end for**

private momenta by

$$H_t = \mu M_{t-1} + \bar{G}_t, \qquad M_t = \mu M_{t-1} + \widetilde{G}_t, \qquad M_0 = 0.$$

Here $\mathcal{F}_{t-1}$ includes the complete past, including earlier probes and any auxiliary updates. Let $\mathcal{R}_t^{\mathrm{grad}}$ collect any current gradient-evaluation randomness, drawn independently of the current privacy noise, and set $\mathcal{G}_t = \sigma(\mathcal{F}_{t-1}, \mathcal{B}_t, \mathcal{R}_t^{\mathrm{grad}})$. For deterministic per-example losses, $\mathcal{R}_t^{\mathrm{grad}}$ is omitted. Then $H_t$ is $\mathcal{G}_t$-measurable and

$$M_t = H_t + \rho_t E_t, \qquad E_t \mid \mathcal{G}_t \sim \mathcal{N}(0, I_d), \qquad \rho_t^2 = \nu^2.$$

Thus $\rho_t$ is the fresh noise scale at the input of the matrix map, not the accumulated momentum-noise scale. The reference $H_t$ already includes past noisy updates; it is not a momentum buffer from a separate nonprivate trajectory.

**Newton-Schulz output bias.** Let $Q_{\kappa,q}$ denote the $q$-step Newton-Schulz output map used to produce the Muon direction. Along the adaptive trajectory, the DP-Muon direction is

$$O_t^{(0)} := Q_{\kappa,q}(M_t).$$

Since $Q_{\kappa,q}$ is nonlinear, even centered Gaussian noise in $M_t$ induces a bias after applying $Q_{\kappa,q}$. Let

$$(P_\rho F)(X) := \mathbb{E}_{U\sim\mathcal{N}(0,I_d)}[F(X+\rho U)]$$

be the Gaussian heat operator. Then

$$\mathbb{E}[O_t^{(0)} \mid \mathcal{G}_t] = (P_{\rho_t} Q_{\kappa,q})(H_t),$$

where the expectation is conditional on the adaptive pre-noise information. The positive smooth regularizer $\gamma_{\mathrm{ns}}$ in Algorithm 1 gives globally bounded derivatives of every fixed order (Appendix B). The heat expansion gives

$$(P_{\rho_t} Q_{\kappa,q})(H_t) = Q_{\kappa,q}(H_t) + \frac{\rho_t^2}{2}\Delta Q_{\kappa,q}(H_t) + O(\rho_t^4),$$

where $\Delta$ is the entrywise Laplacian, applied componentwise. Hence the leading privacy-induced bias is a matrix-valued curvature term at the output of the Newton-Schulz map. It is not a scalar additive bias in the momentum buffer.

**DP-MuonBC.** DP-MuonBC cancels the leading heat-smoothing term by an additional randomized post-processing step. Let

$$U_{t,1},\dots,U_{t,J} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,I_d)$$

be data-independent Gaussian probe matrices, independent of the history, current lot, and privacy noises. Define the antithetic probe average

$$O_t^{(2)} := \frac{1}{2J}\sum_{j=1}^{J}\left[Q_{\kappa,q}(M_t+\rho_t U_{t,j}) + Q_{\kappa,q}(M_t-\rho_t U_{t,j})\right],$$

and define the bias-corrected direction

$$O_t^{\mathrm{bc}} := 2O_t^{(0)} - O_t^{(2)}.$$

The corresponding DP-MuonBC update is

$$W_t = W_{t-1} - \eta_t O_t^{\mathrm{bc}}.$$

The correction is applied at the output of the Newton-Schulz map, rather than to the momentum buffer. This is essential because the momentum noise is already centered before the nonlinear map. The leading bias is created by the curvature of $Q_{\kappa,q}$, so DP-MuonBC estimates the locally heat-smoothed output and extrapolates back toward the unsmoothed Newton-Schulz direction. The antithetic probes cancel first-order perturbations while preserving the same expectation.

**Theorem 3.1** (Adaptive cancellation of the leading output bias)**.** *For the smooth Newton-Schulz map $Q_{\kappa,q}$ in Algorithm 1, along the actual adaptive trajectory,*

$$\mathbb{E}[O_t^{\mathrm{bc}} \mid \mathcal{G}_t] = \big(2P_{\rho_t} - P_{\sqrt{2}\rho_t}\big)Q_{\kappa,q}(H_t).$$

*Moreover, for fixed map parameters, the following expansion holds uniformly in the pre-noise state as $\rho_t \to 0$:*

$$\mathbb{E}[O_t^{\mathrm{bc}} \mid \mathcal{G}_t] = Q_{\kappa,q}(H_t) - \frac{\rho_t^4}{4}\Delta^2 Q_{\kappa,q}(H_t) + O(\rho_t^6).$$

*In particular,*

$$\mathbb{E}[O_t^{\mathrm{bc}} \mid \mathcal{G}_t] = Q_{\kappa,q}(H_t) + O(\rho_t^4),$$

*so DP-MuonBC removes the leading $O(\rho_t^2)$ output-level bias in the adaptive conditional expectation.*

The theorem explains the role of the extrapolation $2O_t^{(0)} - O_t^{(2)}$: the expectation of $O_t^{(0)}$ corresponds to one application of Gaussian heat smoothing, while the expectation of $O_t^{(2)}$ corresponds to an additional independent smoothing step. The linear combination $2P_{\rho_t} - P_{\sqrt{2}\rho_t}$ cancels the leading second-order heat term and leaves only fourth-order bias. The full proof, including the smoothness argument and the finite-probe Monte Carlo error, is deferred to Appendix B.

The experiments also tune a strength $\lambda$ in $O_t^{\lambda} = (1+\lambda)O_t^{(0)} - \lambda O_t^{(2)}$. The fourth-order claim above is for $\lambda = 1$; Appendix B.7 gives the arbitrary-strength and probe-scale identities.

**Privacy.** In the actual adaptive algorithm, DP-MuonBC uses the same privatized gradients $\widetilde{G}_t$ as DP-Muon. All subsequent operations, including evaluation of $Q_{\kappa,q}$, generation of data-independent probes, formation of $O_t^{\mathrm{bc}}$, and the parameter update, are randomized post-processing of already privatized quantities. Therefore, by the post-processing property of differential privacy, DP-MuonBC satisfies exactly the same $(\varepsilon, \delta)$-DP guarantee as DP-Muon for any number of probes $J$.

## 4 Stationarity Bounds for DP-Muon and DP-MuonBC with Global Clipping

We analyze one Muon-target matrix block $W_t \in \mathbb{R}^{m\times n}$, with $r = \min\{m, n\}$ and $d = mn$. The other parameters are held fixed in this section, and the update uses a constant stepsize and no weight decay. Let

$$f(W) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(W; z)], \qquad g_t = \nabla f(W_{t-1}), \qquad D_0 = f(W_0) - \inf_W f(W) < \infty.$$

The natural stationarity measure for Muon is the nuclear norm $\|g_t\|_*$, whose dual norm is the operator norm of the update direction.

We use the global clipping mechanism of Section 2. Let $\nu = \sigma C/B$, and define

$$\mathfrak{B}_T(C) := \sqrt{r} \sup_{1\le t\le T} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\big(\|\mathbf{G}_t^{(i)}\|_2 - C\big)_+\right], \qquad v_C^2 := \frac{C^2}{B} + d\nu^2.$$

For

$$w_{t,j} := \frac{\beta^{t-j}}{\sum_{k=0}^{t-1}\beta^k}, \qquad 1 \le j \le t,$$

define

$$\Gamma_T(\beta) := \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{j=1}^{t} w_{t,j}^2\right)^{1/2}.$$

The term $\mathfrak{B}_T(C)$ captures clipping bias, while $v_C\Gamma_T(\beta)$ captures sampling and privacy noise after momentum averaging.

We assume operator–nuclear smoothness,

$$\|\nabla f(X) - \nabla f(Y)\|_* \le L\|X - Y\|_{\mathrm{op}},$$

together with the Poisson-sampling and independent-noise conditions stated in Appendix C.

**DP-Muon.** Let $\varepsilon_q \in [0, 1)$ denote the deterministic expected finite-step Newton–Schulz approximation error defined in Appendix C, and set $\chi_q = (1 - \varepsilon_q)^{-1}$.

**Theorem 4.1** (Finite-horizon stationarity of DP-Muon)**.** *For every horizon $T \ge 1$,*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|g_t\|_* \le \chi_q\left[\frac{D_0}{\eta T} + \frac{L\eta}{2} + (2 - \varepsilon_q)\left(\mathfrak{B}_T(C) + \sqrt{r}\, v_C\Gamma_T(\beta) + \frac{L\eta\beta}{1-\beta}\right)\right].$$

The bound separates optimization error, clipping bias, stochastic and privacy noise, momentum tracking, and finite-step Newton–Schulz error.

**DP-MuonBC.** Let $H_t = \beta M_{t-1} + \bar{G}_t$ be the adaptive pre-noise momentum from Section 3. Let $\varepsilon_q^{\mathrm{bc}} \in [0, 1)$ denote its corresponding Newton–Schulz approximation error. For strength $\lambda \ge 0$, define

$$b_\lambda := |1 - \lambda|\frac{\nu^2 M_2}{2} + \lambda\frac{\nu^4 M_4}{4},$$

where $M_2$ and $M_4$ are derivative bounds of the smooth Newton–Schulz map. Let $K_\lambda \ge 1$ denote the corresponding update-moment factor defined in Appendix D. In particular, $b_1 = \nu^4 M_4/4$, whereas a general $\lambda \ne 1$ may retain a second-order term.

**Theorem 4.2** (Finite-horizon stationarity of DP-MuonBC). *If $b_\lambda < 1 - \varepsilon_q^{\mathrm{bc}}$, then*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|g_t\|_* \le \frac{\dfrac{D_0}{\eta T} + \dfrac{L\eta K_\lambda^2}{2} + (2-\varepsilon_q^{\mathrm{bc}})\left(\mathfrak{B}_T(C) + \sqrt{r}\, v_C \Gamma_T(\beta) + \dfrac{L\eta K_\lambda \beta}{1-\beta}\right)}{1-\varepsilon_q^{\mathrm{bc}} - b_\lambda}.$$

Theorem 4.2 makes the tradeoff introduced by bias correction explicit. At $\lambda = 1$, the nonlinear output-bias term decreases from second to fourth order in the fresh privacy-noise scale, while the corrected direction incurs an additional update-moment cost through $K_1$.

The comparison becomes sharper in a bias-dominated regime. Consider the same Smooth-NS map and common clipping, noise, and Newton–Schulz quality envelopes for DP-Muon and $\lambda = 1$ DP-MuonBC. Let

$$b_{\mathrm{mu}} = \frac{\nu^2 M_2}{2}, \qquad b_{\mathrm{bc}} = \frac{\nu^4 M_4}{4},$$

and let $\alpha \in (0, 1]$ denote the common clean-direction alignment level.

**Corollary 4.3** (A regime where DP-MuonBC improves the bound). *Suppose $D_0 > 0$ and*

$$b_{\mathrm{bc}} < b_{\mathrm{mu}} < \alpha.$$

*Then there exists an explicit $\eta_\star > 0$ such that, for every $0 < \eta < \eta_\star$, the $\lambda = 1$ DP-MuonBC bias-aware stationarity bound is strictly smaller than the corresponding matched DP-Muon bound. In particular, $b_{\mathrm{bc}} < b_{\mathrm{mu}}$ whenever*

$$\nu^2 M_4 < 2M_2.$$

Thus, when the privacy noise is small enough for our correction to reduce the output-bias envelope and the stepsize is small enough that the additional correction cost does not dominate, DP-MuonBC improves the matched stationarity certificate. The explicit threshold and the exact comparison condition are given in Appendix D.

**Vanishing stationarity.** Assume additionally that there exists $K > 0$, uniform over the horizon-indexed runs, such that

$$\sup_{1\le t\le T}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\exp\left(\frac{\|\mathbf{G}_t^{(i)}\|_2^2}{K^2}\right)\right] \le e,$$

and choose

$$\eta_T = \eta_0 T^{-3/4}, \qquad \beta_T = 1 - T^{-1/2}, \qquad C_T = K\sqrt{2\log T}.$$

**Corollary 4.4** (Vanishing stationarity). *Under the above schedule,*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|g_t^{\mathrm{mu}}\|_* \le \chi_q\, O\left(T^{-1/4}\sqrt{\log T}\right).$$

*If $K_\lambda$ is uniformly bounded and $1 - \varepsilon_q^{\mathrm{bc}} - b_\lambda$ is uniformly bounded away from zero, then*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|g_t^{\mathrm{bc}}\|_* = O\left(T^{-1/4}\sqrt{\log T}\right).$$

These are fixed-noise-multiplier optimization asymptotics: $\sigma$ is held fixed rather than the total privacy budget. For a fixed $(\varepsilon, \delta)$, the accountant-certified noise multiplier must instead be substituted into the finite-horizon bounds.

## 5 Experiments

**Setup.** We evaluate differentially private fine-tuning of GPT-2 on the E2E data-to-text benchmark (Novikova et al., 2017). All optimizers use the same single-global per-example clipping mechanism: the full trainable-parameter gradient is clipped once, followed by one joint Gaussian release per update. We use Poisson sampling, 10 training epochs, maximum sequence length 100, physical batch

| Method | $\varepsilon \approx 1$ | $\varepsilon \approx 2$ | $\varepsilon \approx 4$ | $\varepsilon \approx 8$ |
|---|---|---|---|---|
| DP-SGD$^{\dagger}$ | 0.610520 | 0.635502 | 0.599148 | 0.572277 |
| DP-Adam | 0.570687 | 0.548685 | 0.533640 | 0.519461 |
| DP-AdamBC | 0.571525 | 0.547806 | 0.530832 | 0.515512 |
| DP-Muon | 0.567287 | 0.543061 | 0.526304 | 0.513344 |
| DP-MuonBC | **0.566709** | **0.542657** | **0.525857** | **0.512834** |

Table 1: E2E test NLL under single-global clipping; lower is better. Hyperparameters are selected by validation NLL, and the table reports the common seed-0 comparison across all five methods. DP-MuonBC denotes first-order Smooth-NS correction with $\lambda = 1.25$. $^{\dagger}$The best measured DP-SGD learning rate lies at the upper boundary of its search range.

size 16, gradient accumulation 64, and expected lot size 1,024, with $\delta = 8 \times 10^{-6}$. The target privacy budgets $\varepsilon \in \{1, 2, 4, 8\}$ correspond to recorded values $0.998599$, $1.995747$, $3.988708$, and $7.978774$. Learning rates are selected using validation NLL, and the resulting test NLL is reported.

For the Muon-based methods, hidden matrix parameters are updated by Muon, while excluded vectors, embeddings, and the language-model head are updated by auxiliary Adam. DP-Muon uses the practical five-step Muon map without bias correction. Throughout the experiments, **DP-MuonBC** denotes the selected first-order Smooth-NS inverse-heat variant with correction strength $\lambda = 1.25$, one antithetic probe, $\kappa = 2$, and smooth-normalization parameter $\gamma_{\mathrm{ns}} = 10^{-4}$. The correction is applied only after the private gradient release and therefore incurs no additional privacy cost.

**E2E results.** DP-Muon has lower test NLL than both DP-Adam and DP-AdamBC at every reported privacy budget in the common seed-0 comparison under global clipping. DP-MuonBC achieves the lowest test NLL at all four privacy budgets. Relative to DP-Muon, its NLL improvements are approximately $5.8 \times 10^{-4}$, $4.0 \times 10^{-4}$, $4.5 \times 10^{-4}$, and $5.1 \times 10^{-4}$ at $\varepsilon \approx 1, 2, 4, 8$, respectively. Thus the selected DP-MuonBC configuration gives a modest but consistent test-NLL improvement in the common seed-0 comparison. The baseline and corrected configurations use different matrix maps, so this comparison does not isolate the effect of bias correction alone.

Additional generation metrics, learning-rate selections, bias-correction ablations, and paired multi-seed comparisons are reported in Appendix G.

## 6 Conclusion and Future Directions

We introduced DP-Muon, a differentially private Muon optimizer that applies global per-example clipping, adds Gaussian noise to the released minibatch average, and treats momentum and Newton–Schulz orthogonalization as post-processing. We established its joint-accounting privacy guarantee and finite-horizon and vanishing stationarity guarantees under global clipping, with bounds that separate optimization error, clipping bias, privacy noise, momentum tracking, and finite-step Newton–Schulz error. We further characterized the output-level bias induced by applying the nonlinear Newton–Schulz map to noisy momentum and proposed DP-MuonBC, whose correction removes the leading fresh-noise conditional bias while preserving the same privacy guarantee. Experiments on E2E show lower test NLL for the tested Muon configurations than for the Adam baselines, with small additional gains for the selected DP-MuonBC configuration. Extending the analysis and empirical evaluation to larger models, broader tasks, and practical Muon training regimes is a natural direction for future work.

### AI use statement

We used generative AI tools to assist with drafting and typesetting parts of the mathematical proof sections in LaTeX and to improve the clarity and English writing of the manuscript. All AI-assisted content was reviewed and verified by the authors. We take responsibility for the final content of this work.

## 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*, pp. 308–318, 2016.

John M. Abowd. The US census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2867–2867. ACM, 2018.

Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17455–17466, 2021.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 464–473. IEEE, 2014.

D. P. T. Apple. Learning with privacy at scale. Technical report, Apple, 2017. URL `https://machinelearning.apple.com/research/learning-with-privacy-at-scale`.

Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems 30*, 2017.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284. Springer, 2006.

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://github.com/KellerJordan/Muon`.

Gyuyeol Kim and Minhwan Oh. Convergence of muon with newton–schulz. In *The Fourteenth International Conference on Learning Representations*, 2026.

Antti Koskela and Antti Honkela. Learning rate adaptation for differentially private learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2465–2475. PMLR, 2020.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022.

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

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275. IEEE, 2017.

Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled gaussian mechanism. *CoRR*, abs/1908.10530, 2019.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. The e2e dataset: New challenges for end-to-end generation. In *Proceedings of the 18th annual SIGdial meeting on discourse and dialogue*, pp. 201–206, 2017.

Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X Yu, Sashank J Reddi, and Sanjiv Kumar. Adaclip: Adaptive clipping for private sgd. *arXiv preprint arXiv:1908.07643*, 2019.

Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Analysis of muon's convergence and critical batch size. *arXiv preprint arXiv:2507.01598*, 2025.

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

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 245–248. IEEE, 2013.

Qiaoyue Tang, Frederick Shpilevskiy, and Mathias Lécuyer. Dp-adambc: Your dp-adam is actually dp-sgd (unless you apply bias correction). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15276–15283, 2024.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1054–1067. ACM, 2014.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1226–1235. PMLR, 2019.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Gradient embedding perturbation for differentially private machine learning. In *ICLR 2021-9th International Conference on Learning Representations*, 2021a.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning*, pp. 12208–12218. PMLR, 2021b.

Yingxue Zhou, Zhiwei Steven Wu, and Arindam Banerjee. Bypassing the ambient dimension: Private sgd with gradient subspace identification. In *International Conference on Learning Representations*, 2021.

## A Proof of Theorem 2.1

*Proof.* Fix a step $t$ and condition on the past transcript $\mathcal{R}_{1:t-1}$, public randomness, and hyperparameters. Then $\theta_{t-1}$, the current Muon-managed parameters, and the clipping threshold are fixed. This conditioning does not reveal the privacy-noise draws or unnoised gradients.

Global clipping gives

$$\|H_t(z)\|_2 \le C.$$

Hence, under add/remove adjacency, changing one example changes the unnoised joint averaged query by at most

$$\Delta = \frac{C}{B}$$

in Euclidean norm, with the public normalization $B$ fixed across adjacent datasets. Under replace-one adjacency, the corresponding bound is $2C/B$.

The release is

$$\widetilde{G}_t = \frac{1}{B}\sum_{z\in\mathcal{B}_t} H_t(z) + Z_t, \qquad Z_t \sim \mathcal{N}\left(0, \frac{\sigma^2 C^2}{B^2} I_{d_{\text{all}}}\right).$$

Thus the full release $\mathcal{R}_t = \widetilde{G}_t$ is one sampled Gaussian mechanism, including both the Muon and auxiliary coordinates. The optimizer partition does not introduce additional private queries.

By assumption, the chosen valid accountant for the actual sampling rule, adjacency notion, global clipping threshold, Gaussian noise scale, and number of iterations certifies the adaptive transcript mechanism

$$D \mapsto \mathcal{R}_{1:T}(D)$$

as $(\varepsilon, \delta)$-differentially private. This gives privacy for the entire noisy gradient transcript.

It remains only to check that Muon-specific operations add no privacy loss. Given the initialization, public hyperparameters, and transcript $\mathcal{R}_{1:T}$, the momentum buffers are computed recursively as

$$M_t(W) = \beta M_{t-1}(W) + \widetilde{G}_t(W), \qquad W \in \mathcal{W}.$$

The Newton–Schulz directions

$$O_t(W) = \mathrm{NewtonSchulz}_q(M_t(W))$$

and parameter updates, including optional decoupled weight decay, are functions of the same transcript and public hyperparameters. Any additional randomness used by the implementation is data-independent. Hence the final output is a possibly randomized post-processing of $\mathcal{R}_{1:T}$.

Therefore, by the post-processing property of differential privacy, the full DP-Muon output, including final parameters and any released intermediate Muon quantities that are post-processing of the transcript, is also $(\varepsilon, \delta)$-differentially private. For replace-one adjacency, the same argument applies with joint base sensitivity $2C/B$ and a valid accountant for that sampling and adjacency model. □

For example, under Poisson sampling with inclusion probability $\gamma$, a valid sampled-Gaussian RDP bound $\epsilon_{\mathrm{RDP}}(a; \gamma, \sigma)$ gives

$$\varepsilon = \min_{a \in \mathcal{A}} \left\{ T\epsilon_{\mathrm{RDP}}(a; \gamma, \sigma) + \frac{\log(1/\delta)}{a-1} \right\}, \qquad \varnothing \neq \mathcal{A} \subset (1, \infty),$$

for a finite set of orders $\mathcal{A}$ and $0 < \delta < 1$ (Mironov, 2017; Mironov et al., 2019). A valid tighter conversion may be used instead. These are per-run guarantees; training several candidates on the same private dataset requires accounting for those accesses, even when selection uses public validation data.

# B Details for DP-MuonBC

This appendix gives the technical details for DP-MuonBC, the output-level bias correction method introduced in Section 3. The bias calculation conditions immediately before the current privacy noise and holds along the actual adaptive trajectory. The privacy guarantee follows because DP-MuonBC is a randomized post-processing of the same type of privatized gradient release as DP-Muon.

## B.1 Adaptive clipped-gradient model

Fix one Muon-managed matrix parameter with shape $m \times n$, and write $d = mn$. Let $\bar{G}_t$ be its globally clipped-but-unnoised minibatch mean. The privatized gradient at step $t$ is

$$\widetilde{G}_t = \bar{G}_t + Z_t, \qquad Z_t \sim \mathcal{N}(0, \nu^2 I_d), \qquad \nu^2 := \frac{\sigma^2 C^2}{B^2},$$

where the covariance is understood after vectorizing matrices. The current noise is independent of the complete past and of the current gradient evaluation. Let $\mu = \beta \in [0, 1)$ be the momentum parameter. Define the current clipped pre-noise state and private momentum by

$$H_t = \mu M_{t-1} + \bar{G}_t, \qquad M_t = \mu M_{t-1} + \widetilde{G}_t, \qquad M_0 = 0.$$

Use $\mathcal{G}_t = \sigma(\mathcal{F}_{t-1}, \mathcal{B}_t, \mathcal{R}_t^{\mathrm{grad}})$ as defined in Section 3. In particular, $\mathcal{G}_t$ includes earlier privacy noises, probes and auxiliary updates, but not the current privacy noise or current probes. Unlike a separate noise-free momentum buffer, $H_t$ contains the effects of all past noisy updates.

**Lemma B.1** (Adaptive Gaussian identity)**.** *Along the actual adaptive trajectory,*

$$M_t = H_t + \rho_t E_t, \qquad E_t \mid \mathcal{G}_t \sim \mathcal{N}(0, I_d), \qquad \rho_t^2 = \nu^2.$$

*The state $H_t$ is $\mathcal{G}_t$-measurable.*

*Proof.* The momentum recursion gives $M_t = \mu M_{t-1} + \bar{G}_t + Z_t = H_t + Z_t$. The first two terms are $\mathcal{G}_t$-measurable, whereas $Z_t$ is independent of $\mathcal{G}_t$ and has covariance $\nu^2 I_d$. This proves the identity without any independence assumption between earlier gradients and earlier privacy noises. □

### B.2 The Newton–Schulz output map

We analyze the exact real-arithmetic analogue of the $q$-step Newton–Schulz routine used in DP-Muon, omitting mixed-precision casts. Define the shape-dependent transpose operator

$$\mathcal{T}(G) := \begin{cases} G^\top, & m > n, \\ G, & m \le n. \end{cases}$$

Let $p_\kappa$ be the fixed degree-$\kappa$ Newton–Schulz polynomial in equation 3. For $G \in \mathbb{R}^{m\times n}$, initialize

$$X_0 = \frac{\mathcal{T}(G)}{\alpha(G)}, \qquad \alpha(G) := \sqrt{\gamma_{\mathrm{ns}}^2 + \|\mathcal{T}(G)\|_F^2}, \qquad \gamma_{\mathrm{ns}} > 0.$$

This is the same smooth startup normalization used in Algorithm 1; its fixed regularizer is suppressed in the notation $Q_{\kappa,q}$. For $j = 0, 1, \ldots, q-1$, iterate

$$X_{j+1} = p_\kappa(X_j X_j^\top) X_j.$$

Define the $q$-step Newton–Schulz output map

$$Q_{\kappa,q}(G) := \begin{cases} X_q^\top, & m > n, \\ X_q, & m \le n. \end{cases}$$

The map $Q_{\kappa,q}$ is deterministic and nonlinear. The Newton–Schulz recursion is polynomial in the normalized initializer $X_0$, but the full map is not polynomial because of its smooth normalization.

**Lemma B.2** (Smoothness and bounded derivatives)**.** *For fixed $m, n, \kappa, q$ and $\gamma_{\mathrm{ns}} > 0$, the map $Q_{\kappa,q}$ is globally smooth with bounded derivatives of every fixed order. Moreover, $\|Q_{\kappa,q}(G)\|_{\mathrm{op}} \le 1$, so*

$$\sup_{G\in\mathbb{R}^{m\times n}} \|Q_{\kappa,q}(G)\|_F \le K_Q := \sqrt{\min(m,n)}.$$

*Proof.* The positive denominator makes the normalization smooth, and its derivatives of every fixed order are uniformly bounded. Indeed, each derivative is a finite sum of rational terms with denominator $(\gamma_{\mathrm{ns}}^2 + \|G\|_F^2)^{1/2+j}$ and numerator of degree at most $1 + 2j - s$ at derivative order $s$; such terms are bounded both near zero and at infinity. The normalized initializer has Frobenius norm less than one, so composing it with a fixed finite polynomial recursion retains bounded derivatives. For a singular value $x \in (0, 1]$, the binomial series gives $0 \le x p_\kappa(x^2) \le 1$; the same holds at zero by continuity. Every iterate therefore has operator norm at most one, proving the stated magnitude bounds. □

### B.3 Heat-expansion regularity

For a matrix-valued map $F : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$, let

$$\Delta := \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\partial^2}{\partial X_{ij}^2}$$

denote the entrywise Laplacian, applied componentwise. Lemma B.2 makes the Taylor remainders below uniform in the adaptive expansion point $H_t$, including zero and rank-deficient matrices. The constants may depend on the fixed dimension, polynomial, iteration count and regularizer; they are not asserted to remain uniform as those quantities vary.

For a matrix-valued function $F$, define the Gaussian heat operator

$$(P_\rho F)(X) := \mathbb{E}_{U\sim\mathcal{N}(0,I_d)}\big[F(X + \rho U)\big].$$

Here $P_\rho$ acts componentwise on matrix-valued maps. Bounded derivatives up to order $4$ give

$$(P_\rho Q_{\kappa,q})(H_t) = Q_{\kappa,q}(H_t) + \frac{\rho^2}{2}\Delta Q_{\kappa,q}(H_t) + O(\rho^4).$$

Taylor expansion along $\rho U$ cancels the odd terms in expectation, and Gaussian second moments give the Laplacian term; bounded fourth derivatives and $\mathbb{E}\|U\|_F^4 < \infty$ bound the remainder uniformly. Bounded derivatives up to order 6 similarly give

$$(P_\rho Q_{\kappa,q})(H_t) = Q_{\kappa,q}(H_t) + \frac{\rho^2}{2}\Delta Q_{\kappa,q}(H_t) + \frac{\rho^4}{8}\Delta^2 Q_{\kappa,q}(H_t) + O(\rho^6).$$

### B.4 Adaptive output bias

Let

$$O_t^{(0)} := Q_{\kappa,q}(M_t)$$

be the DP-Muon direction. By Lemma B.1,

$$\mathbb{E}\Big[O_t^{(0)} \,\Big|\, \mathcal{G}_t\Big] = (P_{\rho_t} Q_{\kappa,q})(H_t),$$

where this expectation is over the current privacy noise, conditional on $\mathcal{G}_t$. Thus the conditional Newton–Schulz output bias is

$$\mathcal{B}_t^{\mathrm{NS}} := (P_{\rho_t} Q_{\kappa,q})(H_t) - Q_{\kappa,q}(H_t).$$

By Lemma B.2, uniformly in $H_t$ as $\rho_t \to 0$,

$$(P_{\rho_t} Q_{\kappa,q})(H_t) = Q_{\kappa,q}(H_t) + \frac{\rho_t^2}{2}\Delta Q_{\kappa,q}(H_t) + O(\rho_t^4).$$

Consequently,

$$\mathcal{B}_t^{\mathrm{NS}} = \frac{\rho_t^2}{2}\Delta Q_{\kappa,q}(H_t) + O(\rho_t^4).$$

Thus, conditionally on the current pre-noise state, the leading privacy-induced bias is a matrix-valued output-level curvature term. It is not a scalar additive bias in the momentum buffer.

### B.5 Definition and proof of DP-MuonBC

DP-MuonBC estimates and cancels the leading heat-smoothing term by an additional randomized post-processing step. Let

$$U_{t,1}, \ldots, U_{t,J} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$$

be data-independent Gaussian probe matrices, independent of the history, current lot, and DP noise. Define

$$O_t^{(2)} := \frac{1}{2J}\sum_{j=1}^{J} \left[Q_{\kappa,q}(M_t + \rho_t U_{t,j}) + Q_{\kappa,q}(M_t - \rho_t U_{t,j})\right].$$

The bias-corrected direction is

$$O_t^{\mathrm{bc}} := 2O_t^{(0)} - O_t^{(2)}.$$

The corresponding parameter update is

$$W_t = W_{t-1} - \eta_t O_t^{\mathrm{bc}}.$$

*Proof of Theorem 3.1.* Conditionally on $\mathcal{G}_t, M_t$, the antithetic probe average satisfies

$$\mathbb{E}\Big[O_t^{(2)} \,\Big|\, \mathcal{G}_t, M_t\Big] = (P_{\rho_t} Q_{\kappa,q})(M_t),$$

where this conditional expectation is over the probe variables only. By Lemma B.1,

$$M_t = H_t + \rho_t E_t, \qquad E_t \mid \mathcal{G}_t \sim \mathcal{N}(0, I_d),$$

and the probe noise is independent of $E_t$. Therefore the heat semigroup identity gives

$$\mathbb{E}\Big[O_t^{(2)} \,\Big|\, \mathcal{G}_t\Big] = (P_{\rho_t} P_{\rho_t} Q_{\kappa,q})(H_t) = (P_{\sqrt{2}\rho_t} Q_{\kappa,q})(H_t).$$

Similarly,

$$\mathbb{E}\Big[O_t^{(0)} \,\Big|\, \mathcal{G}_t\Big] = (P_{\rho_t} Q_{\kappa,q})(H_t).$$

Combining the last two identities yields

$$\mathbb{E}\big[O_t^{\mathrm{bc}} \,\big|\, \mathcal{G}_t\big] = \big(2P_{\rho_t} - P_{\sqrt{2}\rho_t}\big)Q_{\kappa,q}(H_t).$$

For the expansion, use the uniform heat expansion justified by Lemma B.2. For a matrix-valued map $F$ with bounded derivatives through order six,

$$P_\rho F = F + \frac{\rho^2}{2}\Delta F + \frac{\rho^4}{8}\Delta^2 F + O(\rho^6).$$

Applying this expansion with $\rho = \rho_t$ and with $\rho = \sqrt{2}\rho_t$ gives

$$P_{\rho_t} F = F + \frac{\rho_t^2}{2}\Delta F + \frac{\rho_t^4}{8}\Delta^2 F + O(\rho_t^6),$$

and

$$P_{\sqrt{2}\rho_t} F = F + \rho_t^2 \Delta F + \frac{\rho_t^4}{2}\Delta^2 F + O(\rho_t^6).$$

Therefore

$$(2P_{\rho_t} - P_{\sqrt{2}\rho_t})F = F - \frac{\rho_t^4}{4}\Delta^2 F + O(\rho_t^6).$$

Applying this identity to $F = Q_{\kappa,q}$ at $X = H_t$ gives

$$\mathbb{E}\big[O_t^{\mathrm{bc}} \,\big|\, \mathcal{G}_t\big] = Q_{\kappa,q}(H_t) - \frac{\rho_t^4}{4}\Delta^2 Q_{\kappa,q}(H_t) + O(\rho_t^6).$$

In particular,

$$\mathbb{E}\big[O_t^{\mathrm{bc}} \,\big|\, \mathcal{G}_t\big] = Q_{\kappa,q}(H_t) + O(\rho_t^4),$$

so the leading $O(\rho_t^2)$ output-level bias is canceled. □

### B.6 Finite-probe Monte Carlo error

The finite-$J$ estimator is unbiased for the probe expectation used in Theorem 3.1. Define

$$\zeta_t := O_t^{(2)} - \mathbb{E}\Big[O_t^{(2)} \,\Big|\, \mathcal{G}_t, M_t\Big],$$

where the conditional expectation is over the probe variables only. Then

$$\mathbb{E}[\zeta_t \mid M_t] = 0.$$

Let

$$Y_j := \frac{1}{2}\left[Q_{\kappa,q}(M_t + \rho_t U_{t,j}) + Q_{\kappa,q}(M_t - \rho_t U_{t,j})\right].$$

Then $O_t^{(2)} = J^{-1}\sum_{j=1}^{J} Y_j$, the $Y_j$'s are conditionally i.i.d. given $M_t$, and the uniform boundedness of $Q_{\kappa,q}$ gives

$$\|Y_j\|_F \le K_Q.$$

Therefore

$$\mathbb{E}\big[\|\zeta_t\|_F^2 \,\big|\, M_t\big] \le \frac{K_Q^2}{J}.$$

Thus $J$ controls only the Monte Carlo variance of the correction; it does not change its expectation.

### B.7 Correction strength and probe calibration

For a fixed strength $\lambda \ge 0$, let $O_t^\lambda = (1+\lambda)O_t^{(0)} - \lambda O_t^{(2)}$ and $D_{\rho_t} = P_{\rho_t} - I$. The same conditional-expectation calculation gives the exact identity

$$\mathbb{E}[O_t^\lambda \mid \mathcal{G}_t] - Q_{\kappa,q}(H_t) = \big[(1-\lambda)D_{\rho_t} - \lambda D_{\rho_t}^2\big]Q_{\kappa,q}(H_t).$$

At $\lambda = 1.25$, the generic leading term is $-\rho_t^2 \Delta Q_{\kappa,q}(H_t)/8$. Thus the bias need not be fourth order. The fourth-order cancellation in Theorem 3.1 is the unit-strength, matched-scale case.

The probe scale must also be measured at the actual map input. If the fresh input-noise standard deviation is $\nu_{\mathrm{in}}$ but the probe standard deviation is a fixed $\tau$, then

$$\mathbb{E}[O_t^{\lambda} \mid \mathcal{G}_t] = \big[(1+\lambda)P_{\nu_{\mathrm{in}}} - \lambda P_{\sqrt{\nu_{\mathrm{in}}^2+\tau^2}}\big]Q_{\kappa,q}(H_{\mathrm{in}}),$$

where $H_{\mathrm{in}}$ is the corresponding pre-noise map input. As $\nu_{\mathrm{in}}, \tau \to 0$ with fixed map parameters and strength, its leading bias coefficient is $(\nu_{\mathrm{in}}^2 - \lambda\tau^2)/2$, multiplying $\Delta Q_{\kappa,q}(H_{\mathrm{in}})$. At startup-normalized momentum $M_t/s_t$, with $s_t = \sum_{j=0}^{t-1}\mu^j$, the fresh scale is $\nu/s_t$, not the accumulated scale $\nu\sqrt{\sum_{j=0}^{t-1}\mu^{2j}}/s_t$. The latter cannot be substituted into the matched-scale theorem along an adaptive trajectory. These identities concern conditional output bias, not convergence of the full Muon–Adam optimizer or elimination of clipping bias.

### B.8 Privacy of DP-MuonBC

In the actual adaptive algorithm, DP-MuonBC uses the same type of privatized gradient release $\widetilde{G}_t$ as DP-Muon. All subsequent operations, including evaluation of $Q_{\kappa,q}$, generation of data-independent Gaussian probes, formation of $O_t^{\mathrm{bc}}$, and the parameter update, are randomized post-processing of already privatized quantities using data-independent randomness.

Therefore, by the post-processing property of differential privacy, DP-MuonBC satisfies exactly the same privacy guarantee as DP-Muon for any number of probes $J$.

## C Proofs for Vanishing Stationarity Bound for DP-Muon with Global Clipping

In this appendix, we give the full proof of the finite-horizon stationarity bound for DP-Muon with global clipping. The analysis is for one fixed Muon-target matrix block, with all other parameters held fixed.

### C.1 Single-block setup

Fix one Muon-target matrix parameter

$$W_t \in \mathbb{R}^{m\times n}, \qquad r := \min\{m,n\}, \qquad d := mn.$$

Let

$$f(W) := \mathbb{E}_{z\sim\mathcal{D}}[\ell(W;z)], \qquad f^\star := \inf_W f(W) > -\infty, \qquad D_0 := f(W_0) - f^\star,$$

and define

$$g_t := \nabla f(W_{t-1}).$$

At iteration $t$, let

$$G_t^{(i)} := \nabla_W \ell(\theta_{t-1}; z_i), \qquad \mathbf{G}_t^{(i)} := \nabla_\theta \ell(\theta_{t-1}; z_i), \qquad i = 1,\dots,N.$$

The expectation over $z\sim\mathcal{D}$ is uniform over these $N$ indexed records. Global clipping uses

$$H_t^{(i)} := \frac{G_t^{(i)}}{\max\{1, \|\mathbf{G}_t^{(i)}\|_2/C\}} = \alpha_t^{(i)} G_t^{(i)}, \qquad \alpha_t^{(i)} := \frac{1}{\max\{1, \|\mathbf{G}_t^{(i)}\|_2/C\}}.$$

The released private matrix gradient is

$$\widetilde{G}_t = \frac{1}{B}\sum_{i=1}^{N} I_{t,i} H_t^{(i)} + Z_t, \qquad Z_t \sim \mathcal{N}\left(0, \frac{\sigma^2 C^2}{B^2} I_d\right), \tag{1}$$

where $I_{t,i}$ are the Poisson inclusion indicators and $B = \gamma N$. The DP-Muon recursion is

$$M_t = \beta M_{t-1} + \widetilde{G}_t, \qquad M_0 = 0, \qquad O_t = \mathrm{NewtonSchulz}_q(M_t), \qquad W_t = W_{t-1} - \eta O_t, \tag{2}$$

where $\beta \in [0,1)$ and $\eta > 0$ may depend on the horizon $T$.

We use the filtration

$$\mathcal{F}_t := \sigma(W_0, M_0, \{I_{s,i}\}_{1\le s\le t,\, 1\le i\le N}, \{Z_s\}_{1\le s\le t}),$$

augmented by all deterministic quantities generated by the algorithm up to time $t$, including $\widetilde{G}_s, M_s, O_s, W_s$ for $s \le t$. Thus $W_{t-1}$ and $M_{t-1}$ are $\mathcal{F}_{t-1}$-measurable, while the minibatch samples and Gaussian privacy noise at time $t$ are drawn after conditioning on $\mathcal{F}_{t-1}$.

Throughout, $\mathrm{Polar}(X)$ denotes the canonical partial polar factor, zero on null singular subspaces, satisfying

$$\|\mathrm{Polar}(X)\|_{\mathrm{op}} \le 1, \qquad \langle X, \mathrm{Polar}(X)\rangle_F = \|X\|_*,$$

with the convention $\mathrm{Polar}(0) = 0$.

### C.2 Assumptions

**Assumption C.1** (Operator–nuclear smoothness)**.** For all $X, Y \in \mathbb{R}^{m\times n}$,

$$\|\nabla f(X) - \nabla f(Y)\|_* \le L\|X - Y\|_{\mathrm{op}}.$$

**Assumption C.2** (Sampling, unbiasedness, and independent Gaussian noise)**.** For every $t$, conditionally on $\mathcal{F}_{t-1}$, the indicators $I_{t,1}, \dots, I_{t,N}$ are independent Bernoulli variables with parameter $\gamma$, and $B = \gamma N$. The per-example loss is deterministic given its parameters and record, and

$$\frac{1}{N}\sum_{i=1}^{N} G_t^{(i)} = g_t.$$

Moreover, $Z_t$ is independent of $\mathcal{F}_{t-1}$ and of $\{I_{t,i}\}_{i=1}^N$, and satisfies

$$\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}[\|Z_t\|_F^2 \mid \mathcal{F}_{t-1}] = d\,\frac{\sigma^2 C^2}{B^2}.$$

The Gaussian noises $\{Z_t\}_{t\ge 1}$ are independent across time.

### C.3 Finite-step Newton–Schulz quality

The proof uses the Newton–Schulz direction only through spectral norm control and its approximation quality to the polar factor of the momentum. We use the finite-step Newton–Schulz analysis of Kim & Oh (2026).

Consider the degree-$\kappa$ Newton–Schulz polynomial

$$p_\kappa(\lambda) = \sum_{s=0}^{\kappa} \frac{(2s)!}{4^s (s!)^2}(1-\lambda)^s, \qquad \kappa \ge 1. \tag{3}$$

For a given $M_t$, initialize

$$X_{t,0} = \frac{\mathcal{T}(M_t)}{\alpha_t}, \qquad \alpha_t := \sqrt{\gamma_{\mathrm{ns}}^2 + \|M_t\|_F^2},$$

using the usual transpose convention when $m > n$, and perform $q$ Newton–Schulz steps

$$X_{t,j+1} = p_\kappa(X_{t,j}X_{t,j}^\top)X_{t,j}, \qquad j = 0, \dots, q-1.$$

Let $O_t := \mathcal{T}^{-1}(X_{t,q})$. If $\Pi_t$ is the orthogonal projector onto $\mathrm{range}(\mathcal{T}(M_t))$ and

$$\delta_{t,0} := \|\Pi_t - X_{t,0}X_{t,0}^\top\|_{\mathrm{op}},$$

with $\delta_{t,0} = 0$ when $M_t = 0$, then the results of Kim & Oh (2026) imply

$$\|O_t\|_{\mathrm{op}} \le 1, \qquad \|O_t - \mathrm{Polar}(M_t)\|_{\mathrm{op}} \le 1 - \sqrt{1 - \delta_{t,0}^{(\kappa+1)^q}}.$$

Set $\varepsilon_{q,t} = 1 - \sqrt{1 - \delta_{t,0}^{(\kappa+1)^q}}$. Then, for all $1 \le t \le T$,

$$\|O_t\|_{\mathrm{op}} \le 1, \qquad \|O_t - \mathrm{Polar}(M_t)\|_{\mathrm{op}} \le \varepsilon_{q,t} < 1. \tag{4}$$

For each fixed $\delta_{t,0} < 1$, this error bound approaches zero doubly exponentially fast in the number of Newton–Schulz steps $q$.

For completeness, write $c_j = \binom{2j}{j}/4^j$ and $P_\kappa(z) = \sum_{j=0}^{\kappa} c_j z^j$. The coefficient recurrence $(j+1)c_{j+1} = (j+\frac{1}{2})c_j$ implies

$$1 - (1-z)P_\kappa(z)^2 = (2\kappa+1)c_\kappa \sum_{j=0}^{\kappa} \frac{c_j}{\kappa+j+1} z^{\kappa+j+1} \le z^{\kappa+1}, \qquad 0 \le z \le 1.$$

Indeed, the derivatives of the two expressions before the inequality are both $(2\kappa+1)c_\kappa z^\kappa P_\kappa(z)$, and both vanish at zero. Their polynomial coefficients are nonnegative and sum to one at $z = 1$. Applying the inequality to the squared singular-value residual proves the exponent $(\kappa+1)^q$ above, including rank-deficient inputs on their nonzero singular subspaces.

For $s_t = \sum_{j=0}^{t-1} \beta^j$ and $\bar{M}_t = M_t/s_t$, define

$$\varepsilon_q := \frac{\sum_{t=1}^{T} \mathbb{E}[\varepsilon_{q,t} \|\bar{M}_t\|_*]}{\sum_{t=1}^{T} \mathbb{E}\|\bar{M}_t\|_*}, \qquad \chi_q := \frac{1}{1-\varepsilon_q}, \tag{5}$$

with $\varepsilon_q = 0$ if the denominator is zero. Thus $\varepsilon_q$ is the average error bound weighted by normalized momentum magnitude, with expectation taken before forming the ratio. Writing $\nu = \sigma C/B$, global clipping gives $\mathbb{E}\|\bar{M}_t\|_* \le \sqrt{r}(C + \nu\sqrt{d})$, so the denominator is finite. If it is positive, $\sum_t \mathbb{E}[(1-\varepsilon_{q,t})\|\bar{M}_t\|_*] > 0$, because $\varepsilon_{q,t} < 1$ whenever $M_t \neq 0$. Consequently $0 \le \varepsilon_q < 1$ and $\chi_q = (1-\varepsilon_q)^{-1}$ is finite. If the denominator is zero, all momenta vanish almost surely and the convention $\varepsilon_q = 0$ is valid. No random path maximum is taken outside an expectation.

### C.4 Clipping residual and private-gradient decomposition

Define the effective clipping residual and effective noise scale by

$$\mathfrak{B}_T(C) := \sqrt{r} \sup_{1 \le t \le T} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \big(\|\mathbf{G}_t^{(i)}\|_2 - C\big)_+\right], \tag{6}$$

where $(x)_+ := \max\{x, 0\}$. Also define

$$v_C^2 := \frac{C^2}{B} + \frac{d\sigma^2 C^2}{B^2}. \tag{7}$$

**Lemma C.3** (Clipped oracle decomposition)**.** *Under Assumption C.2, the private gradient admits the decomposition*

$$\widetilde{G}_t = g_t + b_t + \xi_t,$$

*where $\xi_t$ is $\mathcal{F}_t$-measurable and*

$$\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}\|b_t\|_* \le \mathfrak{B}_T(C), \qquad \mathbb{E}[\|\xi_t\|_F^2 \mid \mathcal{F}_{t-1}] \le v_C^2.$$

*Proof.* Define

$$h_t := \frac{1}{N}\sum_{i=1}^{N} H_t^{(i)}, \qquad b_t := h_t - g_t, \qquad \xi_t := \frac{1}{B}\sum_{i=1}^{N} (I_{t,i} - \gamma) H_t^{(i)} + Z_t.$$

Then $\widetilde{G}_t = g_t + b_t + \xi_t$, since $B = \gamma N$. By conditional independence of the inclusion indicators and the independence and mean-zero property of $Z_t$,

$$\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0.$$

For the bias term,

$$b_t = \frac{1}{N}\sum_{i=1}^{N} (\alpha_t^{(i)} - 1) G_t^{(i)},$$

hence

$$\|b_t\|_* \le \frac{\sqrt{r}}{N}\sum_{i=1}^{N}(1-\alpha_t^{(i)})\|G_t^{(i)}\|_F \le \frac{\sqrt{r}}{N}\sum_{i=1}^{N}\big(\|\mathbf{G}_t^{(i)}\|_2 - C\big)_+.$$

The last inequality uses $\|G_t^{(i)}\|_F \le \|\mathbf{G}_t^{(i)}\|_2$ and the common global clipping multiplier. Taking expectation yields

$$\mathbb{E}\|b_t\|_* \le \mathfrak{B}_T(C).$$

For the fluctuation term, conditional independence gives

$$\mathbb{E}[\|\xi_t\|_F^2 \mid \mathcal{F}_{t-1}] = \frac{\gamma(1-\gamma)}{B^2}\sum_{i=1}^{N}\|H_t^{(i)}\|_F^2 + \mathbb{E}[\|Z_t\|_F^2 \mid \mathcal{F}_{t-1}].$$

Since $\|H_t^{(i)}\|_F \le C$ and $B = \gamma N$,

$$\mathbb{E}[\|\xi_t\|_F^2 \mid \mathcal{F}_{t-1}] \le \frac{(1-\gamma)C^2}{B} + \frac{d\sigma^2C^2}{B^2} \le \frac{C^2}{B} + \frac{d\sigma^2C^2}{B^2} = v_C^2.$$

Thus the original variance envelope remains valid for Poisson sampling; the factor $1-\gamma$ is simply discarded in that envelope. □

### C.5 Normalized momentum tracking

The raw momentum $M_t$ does not track $g_t$ sharply enough because it is not startup-normalized. We therefore use the normalized momentum

$$s_t := \sum_{k=0}^{t-1}\beta^k = \frac{1-\beta^t}{1-\beta}, \qquad \bar{M}_t := \frac{M_t}{s_t}.$$

Since $s_t > 0$, the polar factor is scale-invariant:

$$\mathrm{Polar}(M_t) = \mathrm{Polar}(\bar{M}_t).$$

**Lemma C.4** (Muon inner-product bound with normalized momentum)**.** *Suppose the Newton–Schulz quality bounds in equation 4 hold. Then*

$$\langle g_t, O_t\rangle_F \ge (1-\varepsilon_{q,t})\|g_t\|_* - (2-\varepsilon_{q,t})\|\bar{M}_t - g_t\|_*.$$

*In addition,*

$$\sum_{t=1}^{T}\mathbb{E}\langle g_t, O_t\rangle_F \ge (1-\varepsilon_q)\sum_{t=1}^{T}\mathbb{E}\|g_t\|_* - (2-\varepsilon_q)\sum_{t=1}^{T}\mathbb{E}\|\bar{M}_t - g_t\|_*.$$

*Proof.* Since $\mathrm{Polar}(\bar{M}_t) = \mathrm{Polar}(M_t)$, the bounds in equation 4 imply

$$\begin{aligned}\langle \bar{M}_t, O_t\rangle_F &= \langle \bar{M}_t, \mathrm{Polar}(\bar{M}_t)\rangle_F + \langle \bar{M}_t, O_t - \mathrm{Polar}(M_t)\rangle_F \\ &\ge \|\bar{M}_t\|_* - \|\bar{M}_t\|_*\|O_t - \mathrm{Polar}(M_t)\|_{\mathrm{op}} \\ &\ge (1-\varepsilon_{q,t})\|\bar{M}_t\|_*.\end{aligned}$$

Also,

$$\langle g_t - \bar{M}_t, O_t\rangle_F \ge -\|g_t - \bar{M}_t\|_*\|O_t\|_{\mathrm{op}} \ge -\|g_t - \bar{M}_t\|_*.$$

Therefore

$$\langle g_t, O_t\rangle_F \ge (1-\varepsilon_{q,t})\|\bar{M}_t\|_* - \|g_t - \bar{M}_t\|_*.$$

Using

$$\|\bar{M}_t\|_* \ge \|g_t\|_* - \|\bar{M}_t - g_t\|_*$$

gives the pointwise claim. For the averaged assertion, use the preceding inequality before this triangle bound and the definition of $\varepsilon_q$:

$$\begin{aligned}\sum_t \mathbb{E}\langle g_t, O_t\rangle_F &\ge (1-\varepsilon_q)\sum_t \mathbb{E}\|\bar{M}_t\|_* - \sum_t \mathbb{E}\|\bar{M}_t - g_t\|_* \\ &\ge (1-\varepsilon_q)\sum_t \mathbb{E}\|g_t\|_* - (2-\varepsilon_q)\sum_t \mathbb{E}\|\bar{M}_t - g_t\|_*.\end{aligned}$$

Only the averaged $\varepsilon_q$, not the realized $\varepsilon_{q,t}$, is deterministic. □

Define the exponential weights

$$w_{t,j} := \frac{\beta^{t-j}}{s_t}, \qquad 1 \le j \le t,$$

so that $\sum_{j=1}^{t} w_{t,j} = 1$ and

$$\bar{M}_t = \sum_{j=1}^{t} w_{t,j} \widetilde{G}_j.$$

Also define

$$\Gamma_T(\beta) := \frac{1}{T} \sum_{t=1}^{T} \left( \sum_{j=1}^{t} w_{t,j}^2 \right)^{1/2}. \tag{8}$$

**Lemma C.5** (Normalized momentum tracking)**.** *Under Assumptions C.1 and C.2, and under the Newton–Schulz quality bounds equation 4,*

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \|\bar{M}_t - g_t\|_* \le \mathfrak{B}_T(C) + \sqrt{r}\, v_C \Gamma_T(\beta) + \frac{L\eta\beta}{1-\beta}. \tag{9}$$

*Moreover, there exist universal constants $c_1, c_2 > 0$ such that*

$$\Gamma_T(\beta) \le \frac{c_1}{T\sqrt{1-\beta}} + c_2 \sqrt{1-\beta}. \tag{10}$$

*Proof.* Using $M_t = \sum_{j=1}^{t} \beta^{t-j} \widetilde{G}_j$, we have

$$\bar{M}_t = \sum_{j=1}^{t} w_{t,j} \widetilde{G}_j = \sum_{j=1}^{t} w_{t,j} (g_j + b_j + \xi_j).$$

Thus

$$\bar{M}_t - g_t = \sum_{j=1}^{t} w_{t,j} b_j + \sum_{j=1}^{t} w_{t,j} \xi_j + \sum_{j=1}^{t} w_{t,j} (g_j - g_t).$$

For the bias part,

$$\mathbb{E} \left\| \sum_{j=1}^{t} w_{t,j} b_j \right\|_* \le \sum_{j=1}^{t} w_{t,j} \mathbb{E} \|b_j\|_* \le \mathfrak{B}_T(C).$$

For the stochastic part, $\{\xi_j, \mathcal{F}_j\}_{j \ge 1}$ is a martingale difference sequence by Lemma C.3. Hence, for $i < j$,

$$\mathbb{E} \langle \xi_i, \xi_j \rangle_F = \mathbb{E} \big[ \langle \xi_i, \mathbb{E}[\xi_j \mid \mathcal{F}_{j-1}] \rangle_F \big] = 0,$$

because $\xi_i$ is $\mathcal{F}_{j-1}$-measurable. Therefore

$$\mathbb{E} \left\| \sum_{j=1}^{t} w_{t,j} \xi_j \right\|_F^2 = \sum_{j=1}^{t} w_{t,j}^2 \mathbb{E} \|\xi_j\|_F^2 \le v_C^2 \sum_{j=1}^{t} w_{t,j}^2.$$

Using $\|A\|_* \le \sqrt{r} \|A\|_F$ and Jensen's inequality,

$$\mathbb{E} \left\| \sum_{j=1}^{t} w_{t,j} \xi_j \right\|_* \le \sqrt{r}\, v_C \left( \sum_{j=1}^{t} w_{t,j}^2 \right)^{1/2}.$$

For the drift part, Assumption C.1 and $\|O_s\|_{\mathrm{op}} \le 1$ from equation 4 imply

$$\|g_j - g_t\|_* \le L \|W_{j-1} - W_{t-1}\|_{\mathrm{op}} \le L\eta(t-j).$$

Therefore

$$\sum_{j=1}^{t} w_{t,j}\|g_j - g_t\|_* \le L\eta \sum_{j=1}^{t} w_{t,j}(t-j).$$

Writing $k = t - j$,

$$\sum_{j=1}^{t} w_{t,j}(t-j) = \frac{\sum_{k=0}^{t-1} k\beta^k}{\sum_{k=0}^{t-1} \beta^k} \le \frac{\beta}{1-\beta},$$

where the last inequality follows from

$$\sum_{k=0}^{t-1} k\beta^k = \sum_{m=1}^{t-1}\sum_{k=m}^{t-1} \beta^k \le \frac{1}{1-\beta}\sum_{m=1}^{t-1}\beta^m \le \frac{\beta}{1-\beta}\sum_{k=0}^{t-1}\beta^k.$$

Combining the three bounds and averaging over $t$ yields equation 9.

It remains to bound $\Gamma_T(\beta)$. A direct calculation gives

$$\sum_{j=1}^{t} w_{t,j}^2 = \frac{\sum_{k=0}^{t-1}\beta^{2k}}{\left(\sum_{k=0}^{t-1}\beta^k\right)^2} = \frac{1-\beta}{1+\beta}\cdot\frac{1+\beta^t}{1-\beta^t}.$$

Let $t_\beta := \lceil (1-\beta)^{-1} \rceil$. For $t \le t_\beta$, there is a universal constant $c > 0$ such that

$$1 - \beta^t \ge c\, t(1-\beta),$$

and hence

$$\left(\sum_{j=1}^{t} w_{t,j}^2\right)^{1/2} \le \frac{C'}{\sqrt{t}}$$

for a universal constant $C' > 0$. For $t > t_\beta$, there is a universal constant $C'' > 0$ such that

$$\left(\sum_{j=1}^{t} w_{t,j}^2\right)^{1/2} \le C''\sqrt{1-\beta}.$$

Averaging the two regimes gives

$$\Gamma_T(\beta) \le \frac{C'}{T}\sum_{t=1}^{t_\beta} t^{-1/2} + C''\sqrt{1-\beta} \le \frac{c_1}{T\sqrt{1-\beta}} + c_2\sqrt{1-\beta}.$$

□

### C.6 Proof of the finite-horizon theorem

*Proof of Theorem 4.1.* By Assumption C.1 and the update $W_t = W_{t-1} - \eta O_t$,

$$f(W_t) \le f(W_{t-1}) - \eta\langle g_t, O_t\rangle_F + \frac{L\eta^2}{2}\|O_t\|_{\mathrm{op}}^2 \le f(W_{t-1}) - \eta\langle g_t, O_t\rangle_F + \frac{L\eta^2}{2},$$

where the last inequality uses $\|O_t\|_{\mathrm{op}} \le 1$ from equation 4. Taking expectation, summing from $t = 1$ to $T$, and then applying the averaged assertion of Lemma C.4,

$$\eta(1-\varepsilon_q)\sum_{t=1}^{T}\mathbb{E}\|g_t\|_* \le D_0 + \eta(2-\varepsilon_q)\sum_{t=1}^{T}\mathbb{E}\|\bar{M}_t - g_t\|_* + \frac{L\eta^2 T}{2}.$$

Dividing by $\eta T(1-\varepsilon_q)$ and invoking Lemma C.5 gives

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|g_t\|_* \le \chi_q\left[\frac{D_0}{\eta T} + \frac{L\eta}{2} + (2-\varepsilon_q)\left(\mathfrak{B}_T(C) + \sqrt{r}\, v_C\Gamma_T(\beta) + \frac{L\eta\beta}{1-\beta}\right)\right],$$

which proves the claim. □

### C.7 VANISHING-RATE COROLLARIES

We now specialize Theorem 4.1 to two concrete regimes. In the following $O(\cdot)$ statements, $N, B, \gamma, d, r, \sigma, L, D_0$, the map parameters $\gamma_{\mathrm{ns}}, \kappa, q$ and the displayed moment/tail constants are horizon-independent. The expected quality factor $\chi_q = \chi_{q,T}$ may depend on $T$; its uniform boundedness is an additional condition for convergence, not a consequence of a finite realized trajectory. These are fixed-noise-multiplier optimization rates, not fixed-total-privacy rates. For a fixed target $(\varepsilon, \delta)$, replace $\sigma$ by the value certified by the actual accountant in the finite-horizon theorem; increasing $T$ generally changes that value.

**Corollary C.6** (Vanishing bound under a uniform second-moment condition)**.** *Assume in addition that there exists $G_2 > 0$ such that*

$$\sup_{1\le t\le T} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{G}_t^{(i)}\|_2^2\right] \le G_2^2. \tag{11}$$

*Choose the horizon-dependent parameters*

$$\eta_T = \eta_0 T^{-5/6}, \qquad \beta_T = 1 - T^{-2/3}, \qquad C_T = C_0 T^{1/6}, \tag{12}$$

*where $\eta_0, C_0 > 0$ are constants independent of $T$. Then*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|g_t\|_* \le \chi_q \cdot O(T^{-1/6}). \tag{13}$$

*In particular, if $\chi_q$ is bounded uniformly in $T$, then*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|g_t\|_* \to 0 \qquad \text{as } T \to \infty.$$

*Proof.* For any nonnegative $x$, $(x - C)_+ \le x^2/C$. Hence equation 11 implies

$$\mathfrak{B}_T(C) \le \frac{\sqrt{r}\, G_2^2}{C}.$$

With $C = C_T$, this gives

$$\mathfrak{B}_T(C_T) = O(T^{-1/6}).$$

Next, by Lemma C.5,

$$\Gamma_T(\beta_T) \le \frac{c_1}{T\sqrt{1-\beta_T}} + c_2\sqrt{1-\beta_T} = O(T^{-1/3}).$$

Therefore

$$v_{C_T}\Gamma_T(\beta_T) = C_T\sqrt{\frac{1}{B} + \frac{d\sigma^2}{B^2}}\,\Gamma_T(\beta_T) = O(T^{1/6}) \cdot O(T^{-1/3}) = O(T^{-1/6}).$$

Also,

$$\frac{D_0}{\eta_T T} = O(T^{-1/6}), \qquad \eta_T = O(T^{-5/6}), \qquad \frac{\eta_T \beta_T}{1-\beta_T} = O(T^{-1/6}).$$

Substituting these estimates into Theorem 4.1 proves equation 13. □

**Corollary C.7** (Improved vanishing rate under a light-tail condition)**.** *Assume in addition that there exists $K > 0$ such that*

$$\sup_{1\le t\le T} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \exp\left(\frac{\|\mathbf{G}_t^{(i)}\|_2^2}{K^2}\right)\right] \le e. \tag{14}$$

*For $T \ge 2$, choose*

$$\eta_T = \eta_0 T^{-3/4}, \qquad \beta_T = 1 - T^{-1/2}, \qquad C_T = K\sqrt{2\log T}. \tag{15}$$

*Then*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|g_t\|_* \le \chi_q \cdot O\left(T^{-1/4}\sqrt{\log T}\right). \tag{16}$$

*In particular, if $\chi_q$ is bounded uniformly in $T$, then*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|g_t\|_* \to 0 \qquad \text{as } T\to\infty.$$

*Proof.* By Markov's inequality, equation 14 implies

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}(\|\mathbf{G}_t^{(i)}\|_2 \ge u) \le e\exp(-u^2/K^2), \qquad u\ge 0.$$

Hence

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\|\mathbf{G}_t^{(i)}\|_2 - C)_+\right] = \int_C^\infty \frac{1}{N}\sum_{i=1}^{N}\mathbb{P}(\|\mathbf{G}_t^{(i)}\|_2 \ge u)\,du$$
$$\le e\int_C^\infty e^{-u^2/K^2}\,du \le \frac{eK^2}{2C}e^{-C^2/K^2}.$$

Therefore

$$\mathfrak{B}_T(C) \le \frac{e\sqrt{r}\,K^2}{2C}e^{-C^2/K^2}.$$

With $C = C_T = K\sqrt{2\log T}$,

$$\mathfrak{B}_T(C_T) \le \frac{e\sqrt{r}\,K}{2\sqrt{2\log T}}T^{-2} = o(T^{-1/4}).$$

Next, by Lemma C.5,

$$\Gamma_T(\beta_T) \le \frac{c_1}{T\sqrt{1-\beta_T}} + c_2\sqrt{1-\beta_T} = O(T^{-1/4}).$$

Thus

$$v_{C_T}\Gamma_T(\beta_T) = C_T\sqrt{\frac{1}{B}+\frac{d\sigma^2}{B^2}}\,\Gamma_T(\beta_T) = O(\sqrt{\log T})\cdot O(T^{-1/4}) = O(T^{-1/4}\sqrt{\log T}).$$

Finally,

$$\frac{D_0}{\eta_T T} = O(T^{-1/4}), \qquad \eta_T = O(T^{-3/4}), \qquad \frac{\eta_T\beta_T}{1-\beta_T} = O(T^{-1/4}).$$

Substituting these estimates into Theorem 4.1 yields equation 16. □

# D DP-MuonBC Stationarity and a Regime of Strict Bound Improvement

This appendix derives the DP-MuonBC stationarity bound and identifies a simple regime in which its matched bias-aware certificate is strictly smaller than that of DP-Muon.

## D.1 Bias and update moments

Write $Q = Q_{\kappa,q}$, let

$$M_2 := \sup_X \|\Delta Q(X)\|_{\mathrm{op}}, \qquad M_4 := \sup_X \|\Delta^2 Q(X)\|_{\mathrm{op}},$$

and define the Hessian envelope

$$\mathcal{H}_2 := \sup_X \sup_{\|U\|_F=\|V\|_F=1} \|D^2Q(X)[U,V]\|_F.$$

These constants are finite for fixed matrix shape, Newton–Schulz parameters, and $\gamma_{\mathrm{ns}} > 0$ by Lemma B.2.

For strength $\lambda \geq 0$, the exact identity in Appendix B.7 gives

$$\mathbb{E}[O_t^\lambda \mid \mathcal{G}_t] - Q(H_t) = \big[(1-\lambda)D_\nu - \lambda D_\nu^2\big]Q(H_t), \qquad D_\nu := P_\nu - I.$$

Writing $S_s = P_{\sqrt{2s}}$ and $a = \nu^2/2$, the heat equation gives

$$D_\nu Q = \int_0^a S_s \Delta Q \, ds, \qquad D_\nu^2 Q = \int_0^a \int_0^a S_{s+u} \Delta^2 Q \, ds \, du.$$

Gaussian averaging is a contraction in the uniform operator norm, so

$$\|D_\nu Q\|_{\infty,\mathrm{op}} \leq \frac{\nu^2 M_2}{2}, \qquad \|D_\nu^2 Q\|_{\infty,\mathrm{op}} \leq \frac{\nu^4 M_4}{4}.$$

Therefore

$$\big\|\mathbb{E}[O_t^\lambda \mid \mathcal{G}_t] - Q(H_t)\big\|_{\mathrm{op}} \leq b_\lambda := |1-\lambda|\frac{\nu^2 M_2}{2} + \lambda \frac{\nu^4 M_4}{4}. \tag{17}$$

To control the corrected direction, define

$$C_{\nu,J}(X) := Q(X) - \frac{1}{2J}\sum_{j=1}^{J}\big[Q(X+\nu U_j) + Q(X-\nu U_j)\big].$$

Taylor's theorem with integral remainder yields

$$\left\| Q(X) - \frac{1}{2}\big[Q(X+\nu U) + Q(X-\nu U)\big] \right\|_F \leq \frac{\mathcal{H}_2 \nu^2}{2}\|U\|_F^2.$$

Hence

$$\mathbb{E}\|C_{\nu,J}(X)\|_F^2 \leq \Xi_J^2, \qquad \Xi_J := \frac{\mathcal{H}_2 \nu^2}{2}\sqrt{d^2 + \frac{2d}{J}}.$$

Since $O_t^\lambda = Q(M_t) + \lambda C_{\nu,J}(M_t)$,

$$\mathbb{E}\|O_t^\lambda\|_{\mathrm{op}} \leq 1 + \lambda \Xi_J, \qquad \mathbb{E}\|O_t^\lambda\|_{\mathrm{op}}^2 \leq (1+\lambda\Xi_J)^2.$$

Also, $\|Q(X)\|_{\mathrm{op}} \leq 1$ implies the deterministic bound $\|O_t^\lambda\|_{\mathrm{op}} \leq 1 + 2\lambda$. We therefore set

$$K_\lambda := \min\{1+2\lambda,\ 1+\lambda\Xi_J\}. \tag{18}$$

Then

$$\mathbb{E}\|O_t^\lambda\|_{\mathrm{op}} \leq K_\lambda, \qquad \mathbb{E}\|O_t^\lambda\|_{\mathrm{op}}^2 \leq K_\lambda^2.$$

### D.2 Tracking the adaptive pre-noise momentum

Let

$$s_t := \sum_{k=0}^{t-1} \beta^k, \qquad \bar{H}_t := \frac{H_t}{s_t}, \qquad w_{t,j} := \frac{\beta^{t-j}}{s_t}.$$

The current pre-noise momentum satisfies

$$\bar{H}_t = \sum_{j=1}^{t-1} w_{t,j}\widetilde{G}_j + w_{t,t}\bar{G}_t.$$

Using the clipped-oracle decomposition from Lemma C.3, the previous releases contain the usual martingale-difference fluctuations, while the current term contains the same minibatch fluctuation without the fresh Gaussian noise.

**Lemma D.1** (Pre-noise momentum tracking)**.** *If* $\sup_{t\leq T}\mathbb{E}\|O_t\|_{\mathrm{op}} \leq K$*, then*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\bar{H}_t - g_t\|_* \leq \mathfrak{B}_T(C) + \sqrt{r}\, v_C \Gamma_T(\beta) + \frac{L\eta K\beta}{1-\beta}.$$

*Proof.* The clipping-bias contribution is bounded by $\mathfrak{B}_T(C)$ exactly as in Lemma C.5. The stochastic fluctuations form a martingale-difference sequence, and each relevant second moment is bounded by $v_C^2$, so nuclear–Frobenius comparison and Jensen's inequality give the term $\sqrt{r}\, v_C \Gamma_T(\beta)$. Finally,

$$\mathbb{E}\|g_j - g_t\|_* \le L\,\mathbb{E}\|W_{j-1} - W_{t-1}\|_{\mathrm{op}} \le L\eta K(t-j),$$

and the normalized momentum weights satisfy

$$\sum_{j=1}^{t} w_{t,j}(t-j) \le \frac{\beta}{1-\beta}.$$

Combining the three terms proves the claim. □

### D.3 DP-MuonBC stationarity

For the pre-noise states, define the pointwise Newton–Schulz error

$$\varepsilon_{q,t}^{\mathrm{bc}} := \varepsilon_q(H_t),$$

where $\varepsilon_q(\cdot)$ is the finite-step error function used in Appendix C. Define the deterministic expected weighted average

$$\varepsilon_q^{\mathrm{bc}} := \frac{\sum_{t=1}^{T} \mathbb{E}[\varepsilon_{q,t}^{\mathrm{bc}} \|\bar{H}_t\|_*]}{\sum_{t=1}^{T} \mathbb{E}\|\bar{H}_t\|_*},$$

with $\varepsilon_q^{\mathrm{bc}} = 0$ if the denominator is zero.

The same argument as Lemma C.4 gives

$$\sum_{t=1}^{T} \mathbb{E}\langle g_t, Q(H_t)\rangle_F \ge (1-\varepsilon_q^{\mathrm{bc}}) \sum_{t=1}^{T} \mathbb{E}\|g_t\|_* - (2-\varepsilon_q^{\mathrm{bc}}) \sum_{t=1}^{T} \mathbb{E}\|\bar{H}_t - g_t\|_*.$$

Combining this inequality with equation 17 yields

$$\sum_{t=1}^{T} \mathbb{E}\langle g_t, O_t^{\lambda}\rangle_F \ge (1-\varepsilon_q^{\mathrm{bc}} - b_\lambda) \sum_{t=1}^{T} \mathbb{E}\|g_t\|_* - (2-\varepsilon_q^{\mathrm{bc}}) \sum_{t=1}^{T} \mathbb{E}\|\bar{H}_t - g_t\|_*.$$

*Proof of Theorem 4.2.* Operator–nuclear smoothness gives

$$f(W_t) \le f(W_{t-1}) - \eta\langle g_t, O_t^{\lambda}\rangle_F + \frac{L\eta^2}{2}\|O_t^{\lambda}\|_{\mathrm{op}}^2.$$

Take expectations, sum over $t$, use $f(W_T) \ge \inf_W f(W)$, apply the preceding alignment inequality, and use equation 18. This gives

$$\eta(1-\varepsilon_q^{\mathrm{bc}} - b_\lambda) \sum_{t=1}^{T} \mathbb{E}\|g_t\|_* \le D_0 + \eta(2-\varepsilon_q^{\mathrm{bc}}) \sum_{t=1}^{T} \mathbb{E}\|\bar{H}_t - g_t\|_* + \frac{L\eta^2 T K_\lambda^2}{2}.$$

Apply Lemma D.1 with $K = K_\lambda$ and divide by $\eta T(1-\varepsilon_q^{\mathrm{bc}} - b_\lambda)$. □

### D.4 When does bias correction strictly improve the bound?

A direct DP-Muon proof may work with the realized noisy momentum and therefore does not need an explicit output-bias denominator. To isolate the effect of bias correction itself, we compare parallel pre-noise-reference certificates for the same Smooth-NS map.

Assume that the two trajectories share common deterministic clipping and tracking envelopes, the same initialization and public hyperparameters, and a common Newton–Schulz quality envelope $\bar{\varepsilon}_q < 1$. Set

$$\alpha := 1 - \bar{\varepsilon}_q, \qquad b_{\mathrm{mu}} := \frac{\nu^2 M_2}{2}, \qquad b_{\mathrm{bc}} := \frac{\nu^4 M_4}{4}.$$

Let $K := K_1$, and define

$$\mathcal{N}_T(K;\alpha) := \frac{D_0}{\eta T} + (1+\alpha)\left[\mathfrak{B}_T(C) + \sqrt{r}\, v_C \Gamma_T(\beta) + \frac{L\eta K\beta}{1-\beta}\right] + \frac{L\eta K^2}{2}. \tag{19}$$

The corresponding bias-aware DP-Muon and unit-strength DP-MuonBC certificates are

$$\mathcal{U}_{\mathrm{mu}} := \frac{\mathcal{N}_T(1;\alpha)}{\alpha - b_{\mathrm{mu}}}, \qquad \mathcal{U}_{\mathrm{bc}} := \frac{\mathcal{N}_T(K;\alpha)}{\alpha - b_{\mathrm{bc}}}.$$

**Proposition D.2** (Exact comparison of the matched certificates)**.** *Suppose*

$$b_{\mathrm{bc}} < b_{\mathrm{mu}} < \alpha.$$

*Then*

$$\mathcal{U}_{\mathrm{bc}} < \mathcal{U}_{\mathrm{mu}}$$

*if and only if*

$$L\eta(K-1)\left[(1+\alpha)\frac{\beta}{1-\beta} + \frac{K+1}{2}\right] < \frac{b_{\mathrm{mu}} - b_{\mathrm{bc}}}{\alpha - b_{\mathrm{mu}}}\mathcal{N}_T(1;\alpha). \tag{20}$$

*Proof.* All denominators are positive. Cross-multiplying

$$\frac{\mathcal{N}_T(K;\alpha)}{\alpha - b_{\mathrm{bc}}} < \frac{\mathcal{N}_T(1;\alpha)}{\alpha - b_{\mathrm{mu}}}$$

and rearranging gives

$$\mathcal{N}_T(K;\alpha) - \mathcal{N}_T(1;\alpha) < \frac{b_{\mathrm{mu}} - b_{\mathrm{bc}}}{\alpha - b_{\mathrm{mu}}}\mathcal{N}_T(1;\alpha).$$

Direct subtraction of equation 19 gives

$$\mathcal{N}_T(K;\alpha) - \mathcal{N}_T(1;\alpha) = L\eta(K-1)\left[(1+\alpha)\frac{\beta}{1-\beta} + \frac{K+1}{2}\right].$$

This proves the equivalence. □

The exact criterion has a simple sufficient form. Define

$$A_K(\alpha,\beta) := (1+\alpha)\frac{\beta}{1-\beta} + \frac{K+1}{2}.$$

**Corollary D.3** (Moderate-noise, small-stepsize improvement)**.** *Suppose $D_0 > 0$, $K > 1$, and*

$$b_{\mathrm{bc}} < b_{\mathrm{mu}} < \alpha.$$

*Define*

$$\eta_\star := \left[\frac{D_0(b_{\mathrm{mu}} - b_{\mathrm{bc}})}{LT(\alpha - b_{\mathrm{mu}})(K-1)A_K(\alpha,\beta)}\right]^{1/2}. \tag{21}$$

*Then every $0 < \eta < \eta_\star$ satisfies*

$$\mathcal{U}_{\mathrm{bc}} < \mathcal{U}_{\mathrm{mu}}.$$

*If $K = 1$, the same strict inequality holds whenever $b_{\mathrm{bc}} < b_{\mathrm{mu}} < \alpha$.*

*Proof.* Since every term in $\mathcal{N}_T(1;\alpha)$ is nonnegative,

$$\mathcal{N}_T(1;\alpha) \geq \frac{D_0}{\eta T}.$$

Therefore equation 20 is guaranteed whenever

$$L\eta(K-1)A_K(\alpha,\beta) < \frac{b_{\mathrm{mu}} - b_{\mathrm{bc}}}{\alpha - b_{\mathrm{mu}}}\frac{D_0}{\eta T},$$

which is equivalent to $\eta < \eta_\star$. If $K = 1$, the two numerators are equal and the smaller bias gives the larger denominator, so the DP-MuonBC certificate is strictly smaller. □

For unit-strength correction,

$$b_{\mathrm{bc}} < b_{\mathrm{mu}}$$

whenever

$$\nu^2 M_4 < 2M_2.$$

Hence, provided also $b_{\mathrm{mu}} < \alpha$, Corollary D.3 gives a concrete moderate-noise, sufficiently small-stepsize regime in which DP-MuonBC provably improves the matched bias-aware stationarity bound.

This comparison does not assert that DP-MuonBC is uniformly better than every valid DP-Muon certificate. In particular, Theorem 4.1 uses the realized noisy momentum directly and may give a sharper DP-Muon bound because it does not pay an explicit output-bias denominator. The result above isolates the theoretical benefit of correcting the nonlinear Gaussian output bias while accounting for the correction's additional update-moment cost.

# E Coupled Multi-Block Muon Updates and Auxiliary Adam

This section extends the matrix-block analysis to several coupled Muon-managed matrices and moving auxiliary coordinates. The privacy guarantee already applies to the complete model because all coordinates are obtained from the same globally clipped joint Gaussian release. The purpose here is only to track how simultaneous matrix updates and the auxiliary optimizer affect the optimization bound.

## E.1 Product geometry and the global private oracle

Let $W = (W^{(1)}, \dots, W^{(L_m)})$ denote the Muon-managed matrix blocks, with $W^{(\ell)} \in \mathbb{R}^{m_\ell \times n_\ell}$, $r_\ell := \min\{m_\ell, n_\ell\}$, $d_\ell := m_\ell n_\ell$, $R := \sum_{\ell=1}^{L_m} r_\ell$, and $D := \sum_{\ell=1}^{L_m} d_\ell$. Let $u \in \mathbb{R}^{d_u}$ collect the remaining trainable parameters and write $\theta = (W, u)$. For matrix collections define

$$\|H\|_{\max,\mathrm{op}} := \max_{1 \le \ell \le L_m} \|H^{(\ell)}\|_{\mathrm{op}}, \qquad \|G\|_{\Sigma,*} := \sum_{\ell=1}^{L_m} \|G^{(\ell)}\|_*.$$

For the full parameter space define

$$\|(H, v)\|_\times := \max\{\|H\|_{\max,\mathrm{op}}, \|v\|_2\}, \qquad \|(G, h)\|_{\times,*} := \|G\|_{\Sigma,*} + \|h\|_2. \tag{22}$$

These norms are dual because blockwise nuclear–operator duality and Euclidean duality give

$$|\langle G, H\rangle_F + \langle h, v\rangle| \le \|(G, h)\|_{\times,*} \|(H, v)\|_\times,$$

and equality in the dual supremum is attained by choosing each matrix block as a corresponding partial polar factor and the auxiliary vector in the direction of $h$.

Let

$$f(\theta) := \frac{1}{N} \sum_{i=1}^{N} \ell(\theta; z_i), \qquad f_* := \inf_\theta f(\theta) > -\infty,$$

and define

$$g_t := \nabla_W f(\theta_{t-1}), \qquad h_t := \nabla_u f(\theta_{t-1}).$$

Assume the full gradient is Lipschitz in the product geometry:

$$\|\nabla f(\theta) - \nabla f(\theta')\|_{\times,*} \le L \|\theta - \theta'\|_\times \qquad \text{for all } \theta, \theta'. \tag{23}$$

Then the usual line-segment argument gives

$$f(\theta + \Delta) \le f(\theta) + \langle \nabla f(\theta), \Delta\rangle + \frac{L}{2} \|\Delta\|_\times^2. \tag{24}$$

At iteration $t$, let the full per-example gradient be

$$\mathbf{G}_{t,i} := \nabla_\theta \ell(\theta_{t-1}; z_i),$$

let

$$a_{t,i} := \min\left\{1, \frac{C}{\|\mathbf{G}_{t,i}\|_2}\right\},$$

with $a_{t,i} = 1$ when $\mathbf{G}_{t,i} = 0$, and let $\mathbf{H}_{t,i} := a_{t,i}\mathbf{G}_{t,i}$. Let $I_{t,i}$ be conditionally independent Bernoulli variables with parameter $\gamma$, let $B = \gamma N$, and define the joint release

$$\widetilde{\mathbf{G}}_t = \frac{1}{B}\sum_{i=1}^{N} I_{t,i}\mathbf{H}_{t,i} + Z_t, \qquad Z_t \sim \mathcal{N}(0, \nu^2 I), \qquad \nu := \frac{\sigma C}{B}.$$

Write $\widetilde{G}_t^W$ and $\widetilde{g}_t^u$ for its matrix and auxiliary coordinates, respectively, and let

$$\bar{G}_t^W := \frac{1}{B}\sum_{i=1}^{N} I_{t,i} H_{t,i}^W$$

be the matrix coordinates before the current Gaussian noise is added. The Muon matrix momenta satisfy

$$M_t^{(\ell)} = \beta M_{t-1}^{(\ell)} + \widetilde{G}_t^{W,(\ell)}, \qquad M_0^{(\ell)} = 0,$$

and the corresponding current pre-noise states are

$$A_t^{(\ell)} := \beta M_{t-1}^{(\ell)} + \bar{G}_t^{W,(\ell)}.$$

Let

$$s_t := \sum_{k=0}^{t-1}\beta^k, \qquad \bar{A}_t := A_t/s_t, \qquad w_{t,j} := \beta^{t-j}/s_t, \qquad \Gamma_T(\beta) := \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{j=1}^{t} w_{t,j}^2\right)^{1/2}.$$

The parameter updates are written as

$$W_t = W_{t-1} - \eta O_t, \qquad u_t = u_{t-1} - \eta V_t, \tag{25}$$

Any auxiliary-to-Muon learning-rate ratio is absorbed into $V_t$.

Define the global-clipping residual along the actual hybrid trajectory by

$$\mathfrak{B}_T^{\mathrm{hyb}}(C) := \sqrt{R}\sup_{1\le t\le T}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\|\mathbf{G}_{t,i}\|_2 - C)_+\right], \tag{26}$$

and define the matrix-coordinate fluctuation envelope

$$(v_C^W)^2 := \frac{(1-\gamma)C^2}{B} + D\nu^2. \tag{27}$$

**Lemma E.1** (Multi-block clipped-oracle bounds)**.** *Conditionally on the past, the matrix coordinates of the private release can be written as*

$$\widetilde{G}_t^W = g_t + b_t + \xi_t,$$

*where* $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0$ *and*

$$\mathbb{E}\|b_t\|_{\Sigma,*} \le \mathfrak{B}_T^{\mathrm{hyb}}(C), \qquad \mathbb{E}[\|\xi_t\|_F^2 \mid \mathcal{F}_{t-1}] \le (v_C^W)^2. \tag{28}$$

*Proof.* Let $H_{t,i}^{W,(\ell)}$ denote the matrix-block coordinates of $\mathbf{H}_{t,i}$, set $h_t^W = N^{-1}\sum_i H_{t,i}^W$, and define $b_t = h_t^W - g_t$. Then

$$\|b_t\|_{\Sigma,*} \le \frac{1}{N}\sum_i (1-a_{t,i})\sum_{\ell=1}^{L_m}\sqrt{r_\ell}\|G_{t,i}^{(\ell)}\|_F \le \frac{\sqrt{R}}{N}\sum_i (1-a_{t,i})\|\mathbf{G}_{t,i}\|_2 = \frac{\sqrt{R}}{N}\sum_i (\|\mathbf{G}_{t,i}\|_2 - C)_+,$$

which proves the first bound after taking expectations. Define

$$\xi_t := \frac{1}{B}\sum_i (I_{t,i} - \gamma)H_{t,i}^W + Z_t^W.$$

Conditional centering is immediate. Conditional independence and $\|H_{t,i}^W\|_F \le \|\mathbf{H}_{t,i}\|_2 \le C$ give

$$\mathbb{E}\left[\left\|\frac{1}{B}\sum_i (I_{t,i}-\gamma)H_{t,i}^W\right\|_F^2 \middle| \mathcal{F}_{t-1}\right] = \frac{\gamma(1-\gamma)}{B^2}\sum_i \|H_{t,i}^W\|_F^2 \le \frac{(1-\gamma)C^2}{B}.$$

The independent Gaussian matrix coordinates contribute $D\nu^2$, proving equation 28. □

### E.2 SIMULTANEOUS MATRIX-BLOCK STATIONARITY ON A HYBRID TRAJECTORY

For each block let $Q_\ell$ be a matrix map satisfying $\|Q_\ell(X)\|_{\mathrm{op}} \le 1$. Let $Q(A_t) = (Q_1(A_t^{(1)}), \dots, Q_{L_m}(A_t^{(L_m)}))$. Let $\mathcal{G}_t$ contain the past and all current information before the fresh Gaussian matrix noise and any independent probe randomness used to form $O_t$. Assume that for deterministic $b \ge 0$,

$$\|\mathbb{E}[O_t \mid \mathcal{G}_t] - Q(A_t)\|_{\max,\mathrm{op}} \le b \qquad \text{for all } t. \tag{29}$$

For $\alpha \in (0, 1]$, define the finite-step alignment defect

$$\delta_t(\alpha) := \frac{1}{s_t} \sum_{\ell=1}^{L_m} \left[\alpha \|A_t^{(\ell)}\|_* - \langle A_t^{(\ell)}, Q_\ell(A_t^{(\ell)})\rangle_F\right]_+, \qquad \mathcal{E}_T(\alpha) := \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\delta_t(\alpha). \tag{30}$$

The definition permits different matrix shapes and different finite-step Newton–Schulz quality across blocks, and it does not require the objective to separate across blocks.

Let deterministic constants $K_{\times,1}, K_{\times,2} < \infty$ satisfy

$$\sup_{t\le T} \mathbb{E}\|(O_t, V_t)\|_\times \le K_{\times,1}, \qquad \sup_{t\le T} \mathbb{E}\|(O_t, V_t)\|_\times^2 \le K_{\times,2}^2. \tag{31}$$

Finally, define the nonnegative auxiliary adverse-descent term

$$\mathcal{U}_T := \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[-\langle h_t, V_t\rangle\right]_+. \tag{32}$$

**Lemma E.2** (Multi-block pre-noise momentum tracking)**.** *Under equation 23, equation 25, and equation 31,*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\bar{A}_t - g_t\|_{\Sigma,*} \le \mathfrak{B}_T^{\mathrm{hyb}}(C) + \sqrt{R}\, v_C^W \Gamma_T(\beta) + \frac{L\eta K_{\times,1}\beta}{1-\beta}. \tag{33}$$

*Proof.* The normalized pre-noise momentum has the expansion

$$\bar{A}_t = \sum_{j=1}^{t-1} w_{t,j}\widetilde{G}_j^W + w_{t,t}\bar{G}_t^W.$$

Applying Lemma E.1 to the previous releases and using the same clipped mean decomposition for the current term gives a weighted sum of clipping biases, martingale-difference fluctuations, and gradient drift. The clipping contribution is at most $\mathfrak{B}_T^{\mathrm{hyb}}(C)$. Orthogonality of martingale differences, Jensen's inequality, and $\|G\|_{\Sigma,*} \le \sqrt{R}\|G\|_F$ give the stochastic term $\sqrt{R}\, v_C^W (\sum_j w_{t,j}^2)^{1/2}$. For the drift, full product-space smoothness implies

$$\mathbb{E}\|g_j - g_t\|_{\Sigma,*} \le L\mathbb{E}\|\theta_{j-1} - \theta_{t-1}\|_\times \le L\eta K_{\times,1}(t-j).$$

Since $\sum_{j=1}^{t} w_{t,j}(t-j) \le \beta/(1-\beta)$, averaging over $t$ proves equation 33. □

**Proposition E.3** (Matrix stationarity with coupled blocks and auxiliary updates)**.** *Assume equation 23, equation 29, and equation 31. If $b < \alpha \le 1$, then*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|g_t\|_{\Sigma,*} \le \frac{1}{\alpha - b}\Bigg\{\frac{f(\theta_0) - f_*}{\eta T} + (\alpha+1)\left[\mathfrak{B}_T^{\mathrm{hyb}}(C) + \sqrt{R}\, v_C^W \Gamma_T(\beta) + \frac{L\eta K_{\times,1}\beta}{1-\beta}\right] + \mathcal{E}_T(\alpha) + \frac{L\eta K_{\times,2}^2}{2} + \mathcal{U}_T\Bigg\}. \tag{34}$$

*Proof.* By the definition of $\delta_t(\alpha)$ and the unit operator-norm bound on every $Q_\ell$,

$$\langle g_t, Q(A_t)\rangle_F \ge \alpha\|g_t\|_{\Sigma,*} - (\alpha+1)\|\bar{A}_t - g_t\|_{\Sigma,*} - \delta_t(\alpha).$$

Since $g_t$ is $\mathcal{G}_t$-measurable, equation 29 and duality give

$$\mathbb{E}[\langle g_t, O_t\rangle_F \mid \mathcal{G}_t] \geq (\alpha - b)\|g_t\|_{\Sigma,*} - (\alpha + 1)\|\bar{A}_t - g_t\|_{\Sigma,*} - \delta_t(\alpha).$$

Applying equation 24 to the update $-\eta(O_t, V_t)$ yields

$$f(\theta_t) \leq f(\theta_{t-1}) - \eta\langle g_t, O_t\rangle_F - \eta\langle h_t, V_t\rangle + \frac{L\eta^2}{2}\|(O_t, V_t)\|_\times^2.$$

Use $-\langle h_t, V_t\rangle \leq [-\langle h_t, V_t\rangle]_+$, take expectations, sum over $t$, use $f(\theta_T) \geq f_*$, and apply Lemma E.2 and equation 31. Division by $\alpha - b > 0$ proves equation 34. □

If the auxiliary coordinates are fixed, then $V_t = 0$, $\mathcal{U}_T = 0$, and Proposition E.3 becomes a simultaneous stationarity bound for all Muon-managed matrix blocks. Thus coupling among matrix blocks does not require the objective to decompose; it enters through the product-space smoothness constant and the shared global-clipping residual.

For the smooth Newton–Schulz maps used in the bias analysis, let $\Delta$ denote the entrywise Laplacian applied componentwise to a matrix-valued map, and define

$$M_{2,\ell} := \sup_X \|\Delta Q_\ell(X)\|_{\mathrm{op}}, \qquad M_{4,\ell} := \sup_X \|\Delta^2 Q_\ell(X)\|_{\mathrm{op}}.$$

With matched fresh-noise scale $\nu$, the uncorrected pre-noise-reference analysis may take

$$b_{\mathrm{mu}} := \max_\ell \frac{\nu^2 M_{2,\ell}}{2},$$

while strength-$\lambda$ DP-MuonBC may take

$$b_\lambda := \max_\ell \left\{ |1-\lambda| \frac{\nu^2 M_{2,\ell}}{2} + \lambda \frac{\nu^4 M_{4,\ell}}{4} \right\}. \tag{35}$$

Moreover, ordinary Smooth-NS DP-Muon has $\|O_t\|_{\max,\mathrm{op}} \leq 1$, while strength-$\lambda$ DP-MuonBC satisfies the deterministic bound

$$\|O_t\|_{\max,\mathrm{op}} \leq 1 + 2\lambda, \tag{36}$$

because both the base Newton–Schulz output and every antithetic probe average have operator norm at most one. These substitutions give explicit multi-block specializations of Proposition E.3. An ordinary DP-Muon analysis based directly on the realized noisy momentum can again avoid the explicit output-bias term; the proposition above is the common pre-noise-reference formulation that also covers DP-MuonBC.

### E.3 A UNIFORM MOVEMENT BOUND FOR AUXILIARY ADAM

The auxiliary optimizer affects the matrix theory in two distinct ways: its movement changes later matrix gradients, and its own objective contribution appears through $\langle h_t, V_t\rangle$. The first effect can be bounded for ordinary Adam without assuming anything about gradient signs.

Consider one scalar auxiliary coordinate with private gradient sequence $x_1, x_2, \ldots$, and define

$$m_t := (1-\beta_1)\sum_{j=1}^t \beta_1^{t-j} x_j, \qquad v_t := (1-\beta_2)\sum_{j=1}^t \beta_2^{t-j} x_j^2,$$

$$\widehat{m}_t := \frac{m_t}{1-\beta_1^t}, \qquad \widehat{v}_t := \frac{v_t}{1-\beta_2^t}.$$

Let $\varepsilon_A > 0$ be the Adam denominator stabilizer, and let $r_\eta \geq 0$ denote the auxiliary-to-Muon learning-rate ratio, so the corresponding coordinate of $V_t$ is

$$r_\eta \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \varepsilon_A}.$$

**Lemma E.4** (Uniform ordinary-Adam movement bound)**.** *Assume $0 \le \beta_1 < 1$, $0 < \beta_2 < 1$, and $\beta_1^2 < \beta_2$. Then for every $t$ and every real gradient sequence,*

$$\left| \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \varepsilon_A} \right| \le A_* := \frac{1}{\sqrt{(1-\beta_2)(1-\beta_1^2/\beta_2)}}. \tag{37}$$

*Consequently, for an auxiliary block of dimension $d_u$,*

$$\|V_t\|_2 \le K_A := r_\eta \sqrt{d_u}\, A_*. \tag{38}$$

*Proof.* Weighted Cauchy–Schwarz gives

$$\left( \sum_{k=0}^{t-1} \beta_1^k x_{t-k} \right)^2 \le \left( \sum_{k=0}^{t-1} (\beta_1^2/\beta_2)^k \right) \left( \sum_{k=0}^{t-1} \beta_2^k x_{t-k}^2 \right).$$

If $\widehat{v}_t > 0$, substitution into the bias-corrected moments yields

$$\frac{\widehat{m}_t^2}{\widehat{v}_t} \le \frac{(1-\beta_1)^2}{1-\beta_2} \frac{1-\beta_2^t}{(1-\beta_1^t)^2} \sum_{k=0}^{t-1} (\beta_1^2/\beta_2)^k \le \frac{1}{(1-\beta_2)(1-\beta_1^2/\beta_2)}.$$

The last inequality uses $1-\beta_1^t \ge 1-\beta_1$ and $1-\beta_2^t \le 1$. Adding the positive denominator stabilizer can only decrease the ratio. If $\widehat{v}_t = 0$, then every preceding $x_j$ is zero and the Adam update is zero. Summing the coordinatewise bounds proves equation 38. □

Combining Lemma E.4 with the deterministic matrix-direction bounds gives explicit movement constants. For ordinary Smooth-NS DP-Muon one may take

$$K_{\times,1} = K_{\times,2} = \max\{1, K_A\}, \tag{39}$$

and for strength-$\lambda$ DP-MuonBC one may take

$$K_{\times,1} = K_{\times,2} = \max\{1 + 2\lambda, K_A\}. \tag{40}$$

These bounds are conservative but deterministic. If decoupled weight decay or another auxiliary rule is added, its contribution must also be included in $V_t$ and in the corresponding movement bound.

### E.4 What is required for full hybrid stationarity

A magnitude bound on Adam controls how strongly the auxiliary trajectory perturbs the matrix gradients, but it does not imply that the Adam direction is a descent direction for the auxiliary gradient. The residual $\mathcal{U}_T$ in Proposition E.3 keeps this distinction explicit. A full-gradient stationarity statement follows if an auxiliary alignment condition is available.

**Proposition E.5** (Full hybrid stationarity under auxiliary alignment)**.** *Assume the conditions of Proposition E.3. Suppose in addition that there exist $a_u > 0$ and deterministic $e_t \ge 0$ such that*

$$\mathbb{E}\langle h_t, V_t \rangle \ge a_u \mathbb{E}\|h_t\|_2 - e_t \qquad \text{for all } t. \tag{41}$$

*Let $\bar{e}_T := T^{-1} \sum_{t=1}^T e_t$ and $c := \min\{\alpha - b, a_u\} > 0$. Then*

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}\left[ \|g_t\|_{\Sigma,*} + \|h_t\|_2 \right] \le \frac{1}{c} \Bigg\{ \frac{f(\theta_0) - f_*}{\eta T} + (\alpha+1) \left[ \mathfrak{B}_T^{\mathrm{hyb}}(C) + \sqrt{R}\, v_C^W \Gamma_T(\beta) + \frac{L\eta K_{\times,1}\beta}{1-\beta} \right] + \mathcal{E}_T(\alpha) + \frac{L\eta K_{\times,2}^2}{2} + \bar{e}_T \Bigg\}. \tag{42}$$

*Proof.* Repeat the proof of Proposition E.3, but use equation 41 instead of replacing the auxiliary contribution by its adverse part. After summation, the left-hand side contains

$$(\alpha - b) \frac{1}{T} \sum_t \mathbb{E}\|g_t\|_{\Sigma,*} + a_u \frac{1}{T} \sum_t \mathbb{E}\|h_t\|_2.$$

This quantity is at least $cT^{-1} \sum_t \mathbb{E}[\|g_t\|_{\Sigma,*} + \|h_t\|_2]$, which gives equation 42. □

The auxiliary-alignment assumption is not implied by Lemma E.4; the lemma controls movement only. Therefore the unconditional conclusion of Proposition E.3 is matrix stationarity along the actual coupled Muon–Adam trajectory, with the auxiliary effect exposed through $\mathcal{U}_T$. A full-model stationarity guarantee requires additional structure ensuring descent in the auxiliary coordinates.

# F Interaction of Global Clipping, Privacy Noise, Momentum, and Newton–Schulz Orthogonalization

This section makes explicit how the four ingredients of DP-Muon interact. Global clipping changes the mean gradient and sets the sensitivity scale, Gaussian perturbation adds stochastic error at that scale, momentum averages past private gradients while introducing temporal lag, and the nonlinear Newton–Schulz map converts the current Gaussian perturbation into an additional output-level bias. We analyze one Muon-managed matrix block; the global clipping norm is always computed over all trainable coordinates.

## F.1 A decomposition of the normalized private momentum

Fix a matrix parameter $W_t \in \mathbb{R}^{m\times n}$, and let $r = \min\{m, n\}$ and $d = mn$. Let $\mathbf{G}_t^{(i)} = \nabla_\theta \ell(\theta_{t-1}; z_i)$ be the full per-example gradient and let $G_t^{(i)} \in \mathbb{R}^{m\times n}$ be its coordinates in the matrix block under consideration. We assume the per-example losses are deterministic given the current parameters and records. Let $\mathcal{F}_{t-1}$ contain the complete history before the current Poisson sample is drawn, so $G_t^{(i)}$ is $\mathcal{F}_{t-1}$-measurable. For Poisson inclusion indicators $I_{t,i} \sim \mathrm{Bernoulli}(\gamma)$, independent conditionally on $\mathcal{F}_{t-1}$, let $B = \gamma N$ and define the global clipping multiplier and clipped matrix contribution by

$$\alpha_t^{(i)} := \frac{1}{\max\{1, \|\mathbf{G}_t^{(i)}\|_2 / C\}}, \qquad H_t^{(i)} := \alpha_t^{(i)} G_t^{(i)}.$$

Define

$$g_t := \frac{1}{N}\sum_{i=1}^{N} G_t^{(i)}, \qquad h_t := \frac{1}{N}\sum_{i=1}^{N} H_t^{(i)}, \qquad b_t := h_t - g_t,$$

and write the Poisson sampling fluctuation as

$$\xi_t^{\mathrm{sam}} := \frac{1}{B}\sum_{i=1}^{N} (I_{t,i} - \gamma) H_t^{(i)}.$$

The clipped but unnoised minibatch average and the private release are therefore

$$\bar{G}_t = h_t + \xi_t^{\mathrm{sam}}, \qquad \widetilde{G}_t = \bar{G}_t + Z_t = g_t + b_t + \xi_t^{\mathrm{sam}} + Z_t, \qquad Z_t \sim \mathcal{N}(0, \nu^2 I_d), \qquad \nu := \frac{\sigma C}{B}.$$

Let

$$M_t = \beta M_{t-1} + \widetilde{G}_t, \qquad M_0 = 0, \qquad s_t := \sum_{k=0}^{t-1} \beta^k, \qquad w_{t,j} := \frac{\beta^{t-j}}{s_t}, \qquad \widehat{M}_t := \frac{M_t}{s_t}.$$

Then $\sum_{j=1}^{t} w_{t,j} = 1$.

**Proposition F.1** (Four-term decomposition)**.** *For every $t \geq 1$,*

$$\widehat{M}_t - g_t = \underbrace{\sum_{j=1}^{t} w_{t,j} b_j}_{\textit{clipping bias}} + \underbrace{\sum_{j=1}^{t} w_{t,j} \xi_j^{\mathrm{sam}}}_{\textit{sampling fluctuation}} + \underbrace{\sum_{j=1}^{t} w_{t,j} Z_j}_{\textit{privacy noise}} + \underbrace{\sum_{j=1}^{t} w_{t,j}(g_j - g_t)}_{\textit{momentum lag}}. \tag{43}$$

*Moreover, if*

$$\mathfrak{B}_T(C) := \sqrt{r} \sup_{1\leq t\leq T} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} (\|\mathbf{G}_t^{(i)}\|_2 - C)_+\right],$$

*then*

$$\mathbb{E}\|b_t\|_* \leq \mathfrak{B}_T(C), \qquad \mathbb{E}[\|\xi_t^{\mathrm{sam}}\|_F^2 \mid \mathcal{F}_{t-1}] \leq \frac{(1-\gamma)C^2}{B}, \tag{44}$$

*and*

$$\mathbb{E}\left\|\sum_{j=1}^{t} w_{t,j} Z_j\right\|_F^2 = d\nu^2 \sum_{j=1}^{t} w_{t,j}^2. \tag{45}$$

*If $f$ satisfies $\|\nabla f(X) - \nabla f(Y)\|_* \le L\|X - Y\|_{\mathrm{op}}$, the update is $W_t = W_{t-1} - \eta O_t$, and $\sup_t \mathbb{E}\|O_t\|_{\mathrm{op}} \le K_O$, then*

$$\mathbb{E}\left\|\sum_{j=1}^{t} w_{t,j}(g_j - g_t)\right\|_* \le L\eta K_O\left(\frac{\beta}{1-\beta} - \frac{t\beta^t}{1-\beta^t}\right) \le \frac{L\eta K_O \beta}{1-\beta}. \tag{46}$$

*Proof.* Expanding the momentum recursion gives $\widehat{M}_t = \sum_{j=1}^{t} w_{t,j}\widetilde{G}_j$, and substituting $\widetilde{G}_j = g_j + b_j + \xi_j^{\mathrm{sam}} + Z_j$ proves equation 43. For the clipping term, nuclear–Frobenius comparison and the common global clipping multiplier give

$$\|b_t\|_* \le \frac{\sqrt{r}}{N}\sum_{i=1}^{N}(1-\alpha_t^{(i)})\|G_t^{(i)}\|_F \le \frac{\sqrt{r}}{N}\sum_{i=1}^{N}(\|\mathbf{G}_t^{(i)}\|_2 - C)_+.$$

Taking expectations proves the first bound in equation 44. Conditional independence of the Bernoulli indicators gives

$$\mathbb{E}[\|\xi_t^{\mathrm{sam}}\|_F^2 \mid \mathcal{F}_{t-1}] = \frac{\gamma(1-\gamma)}{B^2}\sum_{i=1}^{N}\|H_t^{(i)}\|_F^2 \le \frac{(1-\gamma)C^2}{B},$$

because $\|H_t^{(i)}\|_F \le C$ and $B = \gamma N$. Equation equation 45 follows from independence and isotropy of the Gaussian noises. Finally, smoothness and the update recursion imply $\mathbb{E}\|g_j - g_t\|_* \le L\eta K_O(t-j)$. Since

$$\sum_{j=1}^{t} w_{t,j}(t-j) = \frac{\beta}{1-\beta} - \frac{t\beta^t}{1-\beta^t},$$

the lag bound follows. □

### F.2 Momentum filters privacy noise but increases lag

The exact normalized-momentum weights make the noise–lag tradeoff explicit.

**Proposition F.2** (Noise filtering by momentum)**.** *For $0 \le \beta < 1$,*

$$q_t^2(\beta) := \sum_{j=1}^{t} w_{t,j}^2 = \frac{1-\beta}{1+\beta}\frac{1+\beta^t}{1-\beta^t}. \tag{47}$$

*Consequently, the privacy-noise component of $\widehat{M}_t$ has root-mean-square Frobenius norm*

$$\left(\mathbb{E}\left\|\sum_{j=1}^{t} w_{t,j} Z_j\right\|_F^2\right)^{1/2} = \sqrt{d}\,\nu\, q_t(\beta), \tag{48}$$

*and, for fixed $\beta < 1$,*

$$q_t^2(\beta) \longrightarrow \frac{1-\beta}{1+\beta} \qquad \text{as } t \to \infty. \tag{49}$$

*By contrast, the weighted gradient age satisfies*

$$\sum_{j=1}^{t} w_{t,j}(t-j) = \frac{\beta}{1-\beta} - \frac{t\beta^t}{1-\beta^t} \longrightarrow \frac{\beta}{1-\beta}. \tag{50}$$

*Proof.* Both identities follow by summing finite geometric series. In particular,

$$\sum_{j=1}^{t} w_{t,j}^2 = \frac{\sum_{k=0}^{t-1}\beta^{2k}}{\left(\sum_{k=0}^{t-1}\beta^k\right)^2} = \frac{1-\beta}{1+\beta}\frac{1+\beta^t}{1-\beta^t}.$$

The remaining claims follow from Proposition F.1 and the limit $\beta^t \to 0$. □

Thus larger momentum has two opposing effects. It suppresses the variance of the normalized Gaussian component, asymptotically by the factor $(1-\beta)/(1+\beta)$, but it simultaneously increases the effective age of the gradients in the momentum buffer on the scale $\beta/(1-\beta)$. This is the origin of the separate stochastic-noise and momentum-tracking terms in the stationarity bound.

### F.3 Clipping and privacy noise are coupled through the momentum filter

Because the Gaussian standard deviation is $\nu = \sigma C/B$, the clipping threshold simultaneously controls clipping bias and privacy-noise magnitude. Define

$$\widetilde{v}_C^2 := \frac{(1-\gamma)C^2}{B} + d\nu^2 = C^2\left(\frac{1-\gamma}{B} + \frac{d\sigma^2}{B^2}\right), \qquad \Gamma_T(\beta) := \frac{1}{T}\sum_{t=1}^{T} q_t(\beta).$$

The exact Poisson factor $1-\gamma$ is retained here; dropping it recovers the slightly looser variance envelope used in the main stationarity theorem.

**Proposition F.3** (Clipping–noise balance under momentum)**.** *The clipping-bias and stochastic terms in equation 43 satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\sum_{j=1}^{t} w_{t,j} b_j + \sum_{j=1}^{t} w_{t,j}(\xi_j^{\mathrm{sam}} + Z_j)\right\|_* \le \Phi_T(C,\beta), \tag{51}$$

*where*

$$\Phi_T(C,\beta) := \mathfrak{B}_T(C) + \sqrt{r}\, C\sqrt{\frac{1-\gamma}{B} + \frac{d\sigma^2}{B^2}}\,\Gamma_T(\beta). \tag{52}$$

*Suppose in addition that there exists $K > 0$ such that*

$$\sup_{1\le t\le T}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\exp\left(\frac{\|\mathbf{G}_t^{(i)}\|_2^2}{K^2}\right)\right] \le e. \tag{53}$$

*Then, for every $C > 0$,*

$$\Phi_T(C,\beta) \le \sqrt{r}\left[\frac{eK^2}{2C}e^{-C^2/K^2} + C\, a_T(\beta)\right], \qquad a_T(\beta) := \sqrt{\frac{1-\gamma}{B} + \frac{d\sigma^2}{B^2}}\,\Gamma_T(\beta). \tag{54}$$

*If $a_T(\beta) > 0$, the right-hand side of equation 54 has a unique minimizer $C_* > 0$, characterized by*

$$a_T(\beta) = \frac{eK^2}{2}e^{-C_*^2/K^2}\left(\frac{1}{C_*^2} + \frac{2}{K^2}\right). \tag{55}$$

*Proof.* The combined sampling-plus-Gaussian fluctuation is a martingale-difference sequence with conditional squared-Frobenius moment at most $\widetilde{v}_C^2$. Applying nuclear–Frobenius comparison, Jensen's inequality, and the normalized momentum weights gives the second term in equation 52. Under equation 53, Markov's inequality and tail integration yield

$$\mathfrak{B}_T(C) \le \sqrt{r}\, e\int_C^\infty e^{-u^2/K^2}\, du \le \frac{e\sqrt{r}\,K^2}{2C}e^{-C^2/K^2},$$

which proves equation 54. Differentiating the bracketed expression in equation 54 shows that its derivative is

$$a_T(\beta) - \frac{eK^2}{2}e^{-C^2/K^2}\left(\frac{1}{C^2} + \frac{2}{K^2}\right).$$

The second term decreases continuously from $+\infty$ to 0 as $C$ increases, so there is a unique zero, which gives equation 55. □

For fixed $\beta < 1$, Cesàro convergence and Proposition F.2 give $\Gamma_T(\beta) \to \sqrt{(1-\beta)/(1+\beta)}$ as $T \to \infty$. In this long-horizon limit, stronger momentum reduces the stochastic coefficient in equation 54; because the right-hand side of equation 55 is strictly decreasing in $C_*$, the minimizer of this upper bound shifts toward a larger clipping threshold. Thus the clipping threshold and momentum parameter should not be interpreted as independent controls of privacy noise.

### F.4 Gaussian noise through the nonlinear Newton–Schulz map

The preceding analysis concerns the input to the matrix map. We next isolate what happens when the current Gaussian perturbation passes through the nonlinear Newton–Schulz transformation. Let

$$A_t := \beta M_{t-1} + \bar{G}_t, \qquad M_t = A_t + Z_t, \qquad Z_t \sim \mathcal{N}(0, \nu^2 I_d),$$

where $\bar{G}_t = B^{-1} \sum_i I_{t,i} H_t^{(i)}$ is the current clipped but unnoised Poisson average. Let $\mathcal{G}_t$ contain the complete past and the current Poisson sample, but not $Z_t$. Then $A_t$ is $\mathcal{G}_t$-measurable and the current noise is independent of $\mathcal{G}_t$. Let $Q$ denote the smooth finite-step Newton–Schulz map, let $\Delta = \sum_{a=1}^{m} \sum_{b=1}^{n} \partial^2 / \partial A_{ab}^2$ act componentwise on matrix-valued maps, and define the Gaussian averaging operator

$$(P_\rho Q)(A) := \mathbb{E}_{U \sim \mathcal{N}(0, I_d)}[Q(A + \rho U)].$$

For $J \geq 1$, let

$$\widehat{P}_{\nu,J} Q(M_t) := \frac{1}{2J} \sum_{j=1}^{J} \left[ Q(M_t + \nu U_{t,j}) + Q(M_t - \nu U_{t,j}) \right], \qquad O_t^{\mathrm{bc}} := 2Q(M_t) - \widehat{P}_{\nu,J} Q(M_t),$$

where the probes $U_{t,j}$ are independent standard Gaussian matrices, independent of the training history and privacy noise.

**Proposition F.4** (Adaptive clipping–momentum–noise–orthogonalization identity)**.** *Conditionally on $\mathcal{G}_t$,*

$$\mathbb{E}[Q(M_t) \mid \mathcal{G}_t] = (P_\nu Q)(A_t). \tag{56}$$

*If $Q$ has uniformly bounded derivatives through order four, then*

$$\mathbb{E}[Q(M_t) \mid \mathcal{G}_t] = Q(A_t) + \frac{\nu^2}{2} \Delta Q(A_t) + O(\nu^4), \tag{57}$$

*uniformly in the adaptive state $A_t$. If $Q$ has uniformly bounded derivatives through order six and the DP-MuonBC correction uses matched probe scale $\nu$ and strength $\lambda = 1$, then*

$$\mathbb{E}[O_t^{\mathrm{bc}} \mid \mathcal{G}_t] = Q(A_t) - \frac{\nu^4}{4} \Delta^2 Q(A_t) + O(\nu^6). \tag{58}$$

*Proof.* Equation equation 56 follows immediately from $M_t = A_t + Z_t$ and independence of the current Gaussian noise. The Gaussian heat expansion gives $P_\nu Q = Q + (\nu^2/2)\Delta Q + O(\nu^4)$, proving equation 57. For matched unit-strength DP-MuonBC, the conditional mean is $(2P_\nu - P_{\sqrt{2}\nu})Q(A_t)$. Expanding both heat operators through sixth order gives $(2P_\nu - P_{\sqrt{2}\nu})Q = Q - (\nu^4/4)\Delta^2 Q + O(\nu^6)$, proving equation 58. □

This identity couples all four ingredients directly. Global clipping determines $\bar{G}_t$ and the sensitivity scale $\nu = \sigma C / B$, momentum places the nonlinear map at the adaptive state $A_t = \beta M_{t-1} + \bar{G}_t$, and the curvature of the Newton–Schulz map at that state determines the leading Gaussian output bias. Holding the conditional state fixed, the leading DP-Muon output bias scales quadratically with the clipping threshold through $\nu^2 \propto C^2$, whereas the residual bias after matched $\lambda = 1$ correction is fourth order in $\nu$. DP-MuonBC therefore targets the nonlinear output-bias component; it does not remove clipping bias, sampling fluctuation, or the variance of the underlying private gradient release.

### F.5 Pathwise singular-subspace stability

The heat expansion describes a mean effect. Privacy noise can also rotate the singular directions of the matrix before Newton–Schulz is applied. The following deterministic statement gives a sufficient stability condition.

**Proposition F.5** (Stability of leading singular subspaces)**.** *Let $A, Z \in \mathbb{R}^{m \times n}$, let $\widetilde{A} = A + Z$, and fix $1 \leq k < \min\{m, n\}$. Define*

$$g_k := \sigma_k(A)^2 - \sigma_{k+1}(A)^2, \qquad e_Z := 2\|A\|_{\mathrm{op}} \|Z\|_{\mathrm{op}} + \|Z\|_{\mathrm{op}}^2.$$

*Suppose $g_k > 0$ and $e_Z < g_k/2$. If $U_k$ and $\widetilde{U}_k$ are orthonormal bases of the leading $k$ left singular subspaces of $A$ and $\widetilde{A}$, respectively, then*

$$\|(I - U_k U_k^\top)\widetilde{U}_k\|_{\mathrm{op}} \le \frac{e_Z}{g_k - e_Z}. \tag{59}$$

*The same bound holds for the corresponding right singular subspaces.*

*Proof.* Set $S = AA^\top$ and $\widetilde{S} = \widetilde{A}\widetilde{A}^\top = S + E$, where $E = AZ^\top + ZA^\top + ZZ^\top$. Then $\|E\|_{\mathrm{op}} \le e_Z$. Weyl's inequality implies that the leading $k$ eigenvalues of $\widetilde{S}$ remain separated from the remaining eigenvalues because $g_k - 2e_Z > 0$. Let $U_\perp$ complete $U_k$ to an orthonormal eigenbasis of $S$, let $S_\perp = U_\perp^\top S U_\perp$, and let $\widetilde{\Lambda}_k$ contain the leading $k$ eigenvalues of $\widetilde{S}$. The matrix $X = U_\perp^\top \widetilde{U}_k$ satisfies the Sylvester equation

$$X\widetilde{\Lambda}_k - S_\perp X = U_\perp^\top E \widetilde{U}_k.$$

The spectral separation between $\widetilde{\Lambda}_k$ and $S_\perp$ is at least $g_k - e_Z$, so the standard integral representation of the separated Sylvester equation yields $\|X\|_{\mathrm{op}} \le e_Z/(g_k - e_Z)$. Since $\|X\|_{\mathrm{op}} = \|(I - U_k U_k^\top)\widetilde{U}_k\|_{\mathrm{op}}$, this proves equation 59. Applying the same argument to $A^\top A$ and $\widetilde{A}^\top \widetilde{A}$ proves the right-subspace statement. □

For a standard Gaussian matrix $Z \sim \mathcal{N}(0, \nu^2 I_d)$, a simple net argument gives a fully explicit sufficient condition. For $0 < \omega < 1$, define

$$z_\omega := 2\nu\sqrt{2\big[(m+n)\log 9 + \log(2/\omega)\big]}.$$

Then $\mathbb{P}\{\|Z\|_{\mathrm{op}} > z_\omega\} \le \omega$. Indeed, choose $1/4$-nets of the two unit spheres with cardinalities at most $9^m$ and $9^n$; the operator norm is at most twice the maximum bilinear form over the two nets, and a Gaussian tail bound followed by a union bound gives the claim. Consequently,

$$2\|A\|_{\mathrm{op}} z_\omega + z_\omega^2 < \frac{g_k}{2}$$

is sufficient for the subspace bound in Proposition F.5 to hold with probability at least $1 - \omega$.

In exact arithmetic, if $X = U\Sigma V^\top$, then one Newton–Schulz step satisfies $p(XX^\top)X = Up(\Sigma^2)\Sigma V^\top$. Thus every step acts diagonally in the singular-vector basis of its input; it transforms singular values but does not reconstruct a clean singular subspace that has already been rotated by privacy noise. Momentum and clipping affect the reference state $A_t$ and its spectral gaps, while the Gaussian scale $\nu = \sigma C/B$ determines the size of the fresh perturbation. This separates the pathwise subspace effect of privacy noise from the mean output-bias effect corrected by DP-MuonBC.

**Summary.** The interaction can therefore be read in sequence. Global clipping changes the mean gradient and fixes the sensitivity scale; Gaussian noise enters at a magnitude proportional to the clipping threshold; normalized momentum suppresses the stochastic component but increases gradient lag; the finite-step Newton–Schulz map acts on the resulting noisy matrix and converts the current Gaussian perturbation into a curvature-dependent mean bias; and DP-MuonBC removes the leading term of this final nonlinear bias without altering the preceding clipping, sampling, or privacy-noise terms.

# G Additional Experimental Details

This appendix provides additional implementation details, hyperparameter selection, and multi-seed results for the experiments in Section 5.

## G.1 Experimental protocol

We fine-tune GPT-2 on E2E for 10 epochs with maximum sequence length 100. The packaged data contain 42,061 training references, 4,672 validation references, and 4,693 test references; generation metrics are evaluated on 547 validation and 630 test prompts. The physical batch size is 16 and

gradient accumulation is 64, giving an expected lot size of 1,024. Sampling is Poisson, and privacy is accounted using RDP with $\delta = 8 \times 10^{-6}$.

All methods use a single global clipping operation. For each example, the gradients of all trainable parameters are viewed as one vector, clipped once in Euclidean/Frobenius norm, and included in one joint Gaussian release. Thus each optimizer step contributes one private Gaussian mechanism; all subsequent optimizer operations are post-processing.

For Muon-based methods, hidden matrix parameters use matrix momentum $0.95$ and five matrix-map iterations. Vectors, token and positional embeddings, and the language-model head are excluded from Muon and use auxiliary Adam with learning rate $0.002$. DP-Muon uses the practical five-step Jordan Newton–Schulz map without bias correction.

The DP-MuonBC configuration reported in the main text uses first-order Smooth-NS inverse-heat correction with strength $\lambda = 1.25$, one antithetic probe, $\kappa = 2$, and $\gamma_{\mathrm{ns}} = 10^{-4}$. The strength $\lambda = 1.25$ is a validation-selected extrapolation parameter. At a matched fresh-noise scale, the leading second-order heat term cancels at $\lambda = 1$; Appendix B.7 distinguishes strength and probe calibration for other configurations.

**Additional matrix-map variants.** For the practical baseline, transpose the momentum to the wide convention and initialize $X_0 = \mathcal{T}(M_t)/(\|M_t\|_F + 10^{-7})$. The five Jordan iterations are

$$X_{j+1} = a_J X_j + (b_J S_j + c_J S_j^2) X_j, \qquad S_j = X_j X_j^\top,$$
$$(a_J, b_J, c_J) = (3.4445, -4.7750, 2.0315).$$

The result is transposed back and multiplied by $\sqrt{\max(1, m/n)}$ in the original matrix shape. Smooth-Jordan uses the same polynomial with the positive smooth startup normalization instead. These maps differ from the Smooth-NS map in the theory; in particular the unit operator-norm bound used in Section 4 is not asserted for the practical Jordan map or its shape-rescaled output. For a smooth map $F$, the second-order variant uses $F + \lambda(2F - 3\widehat{P}_\tau F + \widehat{P}_{\sqrt{2}\tau} F)$, where $\widehat{P}$ is the antithetic probe average. The exploratory Jordan1BCx5 variant applies a one-step Gaussian inverse-heat correction to the quintic Jordan polynomial at each of five iterations. Its intermediate outputs are no longer fresh Gaussian observations, so this repetition is not an exact unbiased correction of the full five-step map.

### G.2 Learning-rate selection

Table 2 gives the learning rates selected by validation NLL for the configurations reported in Table 1. Except for DP-SGD, the selected learning rates are bracketed by worse validation-NLL configurations. For DP-SGD, performance continued to improve up to the largest tested learning rate at each privacy budget, so these runs are reported as the best observed DP-SGD configurations rather than as bracketed optima.

| Method | $\varepsilon \approx 1$ | $\varepsilon \approx 2$ | $\varepsilon \approx 4$ | $\varepsilon \approx 8$ |
|---|---|---|---|---|
| DP-SGD$^\dagger$ | 0.0880 | 0.0520 | 0.1040 | 0.1920 |
| DP-Adam | 0.0010 | 0.0010 | 0.0015 | 0.0015 |
| DP-AdamBC | 0.0005 | 0.0010 | 0.0015 | 0.0020 |
| DP-Muon | 0.0050 | 0.0060 | 0.0060 | 0.0060 |
| DP-MuonBC | 0.0070 | 0.0070 | 0.0080 | 0.0080 |

Table 2: Validation-selected learning rates for the main E2E comparison. For Muon-based methods, the listed value is the matrix learning rate. $^\dagger$The best measured DP-SGD point is at the upper search boundary.

### G.3 Generation metrics

Tables 3 and 4 report BLEU and ROUGE-L for the same validation-NLL-selected seed-0 configurations used in Table 1. These metrics are not used for hyperparameter selection. Their optimizer ordering does not always coincide with the NLL ordering.

| Method | $\varepsilon \approx 1$ | $\varepsilon \approx 2$ | $\varepsilon \approx 4$ | $\varepsilon \approx 8$ |
|---|---|---|---|---|
| DP-SGD | 32.275 | 30.456 | 32.908 | 41.100 |
| DP-Adam | 51.458 | 55.893 | 60.661 | 62.607 |
| DP-AdamBC | 52.307 | 57.810 | 62.159 | 62.915 |
| DP-Muon | **56.088** | **59.707** | 62.106 | **63.783** |
| DP-MuonBC | 55.913 | 59.602 | **62.591** | 63.664 |

Table 3: E2E test BLEU for the validation-NLL-selected seed-0 configurations; higher is better.

| Method | $\varepsilon \approx 1$ | $\varepsilon \approx 2$ | $\varepsilon \approx 4$ | $\varepsilon \approx 8$ |
|---|---|---|---|---|
| DP-SGD | 60.224 | 59.703 | 59.683 | 57.438 |
| DP-Adam | 60.982 | 63.110 | 65.968 | 66.659 |
| DP-AdamBC | 61.298 | 64.814 | **66.762** | **66.976** |
| DP-Muon | 63.612 | 64.808 | 66.578 | 66.957 |
| DP-MuonBC | **63.917** | **65.050** | 66.717 | 66.705 |

Table 4: E2E test ROUGE-L for the validation-NLL-selected seed-0 configurations; higher is better.

### G.4 Bias-correction strength ablation

We additionally vary the strength of the first-order Smooth-NS correction at $\varepsilon \approx 2$. The learning rate $0.007$ for the corrected configuration is locally bracketed by learning rates $0.006$ and $0.008$. Within the tested strength range, validation NLL improves gradually as $\lambda$ increases from $0.5$ to $1.25$.

| Configuration | Matrix LR | Val NLL | Test NLL |
|---|---|---|---|
| DP-Muon | 0.006 | 0.534040 | 0.543061 |
| Jordan1BCx5 | 0.006 | 0.533739 | 0.542853 |
| Smooth-NS, first order, $\lambda = 0.50$ | 0.007 | 0.533460 | 0.542709 |
| Smooth-NS, first order, $\lambda = 0.75$ | 0.007 | 0.533438 | 0.542685 |
| Smooth-NS, first order, $\lambda = 1.00$ | 0.007 | 0.533423 | 0.542670 |
| Smooth-NS, first order, $\lambda = 1.25$ | 0.007 | **0.533411** | **0.542657** |

Table 5: First-order bias-correction strength ablation at $\varepsilon = 1.995747$ and seed 0. The selected DP-MuonBC configuration uses $\lambda = 1.25$.

The best tested strength is therefore $\lambda = 1.25$. The difference between $\lambda = 1$ and $\lambda = 1.25$ is small: their test NLLs are $0.542670$ and $0.542657$, respectively. Thus the best tested strength above one should be interpreted as a tuned extrapolation strength rather than as evidence that stronger correction has a better formal bias order.

### G.5 Correction-order ablation

We also evaluated the second-order Smooth-Jordan correction in the systematic global-clipping sweep. As shown in Table 6, it remains close to DP-Muon but is slightly worse in test NLL at each privacy budget. The tested second-order configuration does not improve test NLL over the practical baseline. This is a comparison of configurations, not an isolated change in correction order with every other setting held fixed.

| Method | $\varepsilon \approx 1$ | $\varepsilon \approx 2$ | $\varepsilon \approx 4$ | $\varepsilon \approx 8$ |
|---|---|---|---|---|
| DP-Muon | 0.567287 | 0.543061 | 0.526304 | 0.513344 |
| Smooth-Jordan, second order | 0.567414 | 0.543427 | 0.526572 | 0.513459 |

Table 6: E2E test NLL for the systematic second-order Smooth-Jordan comparison. The higher-order correction is close to, but does not improve upon, the practical DP-Muon baseline in this sweep.

### G.6 PAIRED MULTI-SEED COMPARISON

For $\varepsilon \approx 1, 4, 8$, we additionally run DP-Muon and the selected first-order DP-MuonBC configuration with seeds $0, 1, 2$. Table 7 reports the mean and sample standard deviation. DP-MuonBC obtains lower mean validation and test NLL at all three privacy budgets.

| $\varepsilon$ | Method | Val NLL | Test NLL |
|---|---|---|---|
| 1 | DP-MuonBC | $0.559190 \pm 0.000990$ | $0.564605 \pm 0.001837$ |
| 1 | DP-Muon | $0.559705 \pm 0.000882$ | $0.565070 \pm 0.001956$ |
| 4 | DP-MuonBC | $0.515643 \pm 0.001074$ | $0.523921 \pm 0.001907$ |
| 4 | DP-Muon | $0.516220 \pm 0.001210$ | $0.524436 \pm 0.001903$ |
| 8 | DP-MuonBC | $0.502334 \pm 0.000931$ | $0.511026 \pm 0.001806$ |
| 8 | DP-Muon | $0.503007 \pm 0.001010$ | $0.511677 \pm 0.001708$ |

Table 7: Paired E2E seed-$0/1/2$ comparison between DP-Muon and DP-MuonBC. Values are mean $\pm$ sample standard deviation.

Across these three privacy budgets, DP-MuonBC has lower mean validation and test NLL. Its mean test-NLL improvements over DP-Muon are approximately $4.65 \times 10^{-4}$, $5.15 \times 10^{-4}$, and $6.51 \times 10^{-4}$ at $\varepsilon \approx 1, 4, 8$, respectively. For $\varepsilon \approx 2$, the selected $\lambda = 1.25$ configuration is currently reported only for seed 0, so it is not included in this multi-seed table.