An Optimization View of DP LLM Fine-tuning: When Does Bias Correction Help, and Can the Optimizer Be Improved?

Zixuan Wang

Optimization Methods for Artificial Intelligence Course Project
(compiled May 31, 2026)

Abstract

Differentially private (DP) fine-tuning of large language models runs on DP-SGD/DP-Adam, and the field's reflex when DP noise hurts utility is to *fix the adaptive preconditioner*. The reference for this report, DP-AdamBC [1], does this analytically: it subtracts the closed-form DP-noise variance bias $\Phi = (\sigma_{\mathrm{DP}} C/B)^2$ from Adam's second moment $\hat{v}$. We take an *optimization view* and ask when that correction actually helps in LLM DP-LoRA fine-tuning, and whether any principled optimizer change can beat a well-tuned, privacy-amplified DP-Adam at the same (ε, δ). Our organizing quantity is the *noise share* $\rho := \Phi/\hat{v} = \Phi/(v_{\mathrm{true}} + \Phi) \in [0, 1]$, a one-line, a-priori diagnostic for whether bias correction can do anything. We prove and measure that in the standard recipe ρ **saturates at** ≈ 1: the second moment is an inert noise floor carrying no per-coordinate signal, and learning instead rides the *first* moment $\hat{m}$ (a gradient prefix-sum that averages the zero-mean DP noise toward a below-floor signal). Consequently bias correction **collapses to momentum-SGD** at $\rho \approx 1$ and is only an **effective-learning-rate knob** at $\rho < 1$; four controls (a floor sweep, a floor-only control, a learning-rate control, and a Muon-geometry probe), two models, and three seeds show it never robustly beats a tuned DP-Adam, whose learning-rate plateau (≈ 56.7 proxy BLEU) matches the best DP-AdamBC (56.73). We then *attempt a positive method*, DP-CorrMom: anti-correlated DP noise $w_t = z_t - \lambda z_{t-1}$ injected on the first-moment / prefix-sum path (a DP-FTRL / matrix-factorization mechanism), and find a **unified negative**—across Adam-momentum, $\beta_1 = 0$, and the matched plain-SGD workload, $\lambda > 0$ never beats $\lambda = 0$, and a privacy-*amplified* DP-Adam wins at less budget. We explain this with a *signal-ceiling* theorem: at $\rho \approx 1$ the model is **signal-limited, not variance-limited**, so once basic averaging has extracted the recoverable signal, further noise reduction cannot help—which is precisely why correlated-noise mechanisms that provably help *from-scratch* SGD do not transfer to LLM DP-LoRA fine-tuning. Beyond the reference, we contribute (i) the ρ diagnostic, (ii) the $\hat{m}/\hat{v}$ mechanism, (iii) a four-control attribution of DP-AdamBC, (iv) a verified privacy-sensitivity correction $\kappa = \sqrt{(1-\lambda^{2T})/(1-\lambda^2)}$ for $w_t = z_t - \lambda z_{t-1}$ (the naive $\sqrt{1+\lambda^2}$ *breaks* privacy), and (v) the signal-ceiling theorem for first-moment denoising—none of which appear in DP-AdamBC. *Honesty markers:* the E2E utility is a teacher-forced proxy BLEU, and the positive-method sweeps are single-seed (consistent across five variants).

1 Introduction and Motivation

Differential privacy [2, 3] has become the default formal guarantee for fine-tuning language models on sensitive text. The workhorse is DP-SGD: clip each per-sample gradient to ℓ_2-norm C, add isotropic Gaussian noise of scale $\sigma_{\mathrm{DP}} C$, and average over a batch of size B. The Gaussian mechanism

therefore injects, *per coordinate*, a variance

$$\Phi := \left(\frac{\sigma_{\mathrm{DP}}\, C}{B}\right)^2, \tag{1}$$

that is *independent of the parameters* and does not vanish at the optimum—it is the irreducible price of privacy. When DP-Adam is used instead of DP-SGD, this same Φ silently contaminates Adam's second-moment estimate $\hat{v}$, biasing the adaptive preconditioner.

Two-fold motivation

This report is driven by two questions that the literature poses but, for LLM fine-tuning, leaves open.

Motivation A — Does the field's go-to fix actually hold here? The dominant response to "DP noise corrupts the optimizer" is to repair the adaptive preconditioner. The reference paper, **DP-AdamBC** [1], gives the cleanest instance: because $\mathbb{E}[\hat{v}] = v_{\mathrm{true}} + \Phi$, it simply subtracts the analytic bias, $\hat{v}^{\mathrm{BC}} = \max(\hat{v} - \Phi, \xi)$, restoring Adam's intended geometry. This is principled and reports gains. But *when*, in the operating regime of LLM DP-LoRA fine-tuning, does removing Φ translate into utility—and is the gain anything other than a step-size effect that a practitioner could obtain by simply re-tuning the learning rate? An optimization-view answer requires (i) an a-priori test for whether the correction is even active and (ii) controls that separate genuine per-coordinate de-biasing from a scalar learning-rate change.

Motivation B — If the second moment is a dead lever, can a *different* optimizer change win? If bias correction turns out inert, the natural next move is to ask whether *any* principled change to the private optimizer—changing what we denoise, the update geometry, or the *temporal structure* of the injected noise—can beat a well-tuned, privacy-amplified DP-Adam at the same (ε, δ). This connects to a deep and currently fashionable line of work: correlated-noise / matrix-factorization mechanisms (DP-FTRL and successors) that provably reduce the error accumulated along the gradient *prefix-sum*. The question is whether that provable advantage, established largely for *from-scratch* training, transfers to gentle LLM fine-tuning.

How the literature attacks DP-optimizer utility (four lines)

We situate both motivations against the four families of optimizer-side remedies in the DP literature.

1. **Fix the second moment.** Subtract the analytic noise bias Φ from $\hat{v}$ so the adaptive preconditioner is no longer poisoned by DP noise—*DP-AdamBC* [1], our reference. This is the line Motivation A interrogates.
2. **Reduce the noised dimension.** Privatize gradients in a lower-dimensional subspace (random projection or learned low-rank), so less Gaussian noise is needed for the same ε—e.g. *DP-GRAPE* [15] and gradient-embedding perturbation. The floor's linear dependence on dimension (our Corollary in §3) is exactly what this exploits; LoRA is itself a coarse instance.
3. **Re-allocate the clip/noise budget.** Spend the clipping and noise budget unevenly across parameter groups so high-signal layers are noised less—group-wise clipping [16].
4. **Correlate the injected noise across time** so it cancels along the gradient prefix-sum: *DP-FTRL* [8], matrix factorization [9], banded matrix factorization *BANDMF* [10], the proof

that "correlated noise provably beats independent noise" [12], the single-scalar *DP-λCGD* [13], and buffered linear toeplitz *BLT* mechanisms [14]. This is the line Motivation B tests, and the source of our positive-method attempt.

What we find: numbered prior findings vs. our results

Against these four lines, our optimization-view study produces five results. We number them so the empirical sections (§4) can refer back to each; every number below is grounded in the project's recorded experiments.

(R1) **The noise share ρ saturates at ≈ 1.** Measuring $\rho = \Phi/\hat{v}$ live during training, we find $\rho = 1.00$–1.07 across RoBERTa-large/MNLI (all $\varepsilon \in \{1, 3, 8\}$, all $B \in [16, 2048]$) and Qwen2.5-1.5B/E2E ($B \leq 512$); ρ drops below 1 only when the batch is pushed to B=4096 at ε=8 (ρ=0.955). **The DP-LoRA gradient signal lies below the noise floor**, so the bias-correction sweet spot $\rho \approx \frac{1}{2}$ is structurally out of reach.

(R2) **Learning rides the first moment, not the second.** Models still learn at $\rho \approx 1$ (proxy BLEU ≈ 56 vs. a DP-SGD floor ≈ 21) because Adam's update is $\hat{m}/\sqrt{\hat{v}}$: the *first* moment $\hat{m}$ (a prefix-sum) averages the zero-mean DP noise toward the below-floor signal over many steps, while $\hat{v}$ is a near-constant noise level—**adaptivity is effectively switched off**. This resolves the apparent paradox "$\hat{v}$ is all noise yet the model learns."

(R3) **Against line 1: bias correction collapses to momentum-SGD ($\rho \approx 1$).** With $\hat{v} - \Phi \approx 0$, every DP-AdamBC floor is fully clamped, so the update is $\hat{m}/\sqrt{\xi}$—momentum-SGD with a *fixed* step scale. Utility tracks the floor (a learning-rate effect, three seeds), DP-AdamBC at a binding floor is statistically indistinguishable from a floor-only control with *no* Φ-subtraction, and a Muon-style geometry probe (line 2's discard-the-second-moment cousin) recovers no signal ($+8 \times 10^{-5}$ cosine). The saturation is fundamental to the noise level, not a preconditioner artifact.

(R4) **Where it is active, bias correction is only an effective learning rate.** At ρ=0.955 (large batch) DP-AdamBC genuinely de-biases ($\approx 3.8\times$ larger step) and leads at a short budget (step-40 +1.2 BLEU), *but DP-Adam catches up and passes it by convergence.* A learning-rate control settles it: tuned DP-Adam traces a clean inverted-U with a plateau $\approx$ 56.7 over $\text{lr} \in [2, 5] \times 10^{-3}$ that matches the best DP-AdamBC (56.73). An earlier step-40 "+1.57 win" was correctly flagged as a convergence-speed transient and overturned—**not overclaimed**.

(R5) **Against line 4: the positive-method attempt is a unified negative.** Motivated by R2 (learning rides the prefix-sum), we built *DP-CorrMom*, anti-correlated noise $w_t = z_t - \lambda z_{t-1}$ on the first-moment path, with verified privacy ($\kappa = \sqrt{(1 - \lambda^{2T})/(1 - \lambda^2)}$; λ=0 reproduces DP-SGD-momentum bit-for-bit). Across Adam-momentum, β_1=0, and the *matched* plain-SGD prefix-sum workload (DP-CorrSGD), $\lambda > 0$ never beats λ=0, and a privacy-*amplified* DP-Adam beats all of them at less budget. We explain this with a **signal-ceiling**: at $\rho \approx 1$ the model is signal-limited, so further noise reduction is wasted—**which is exactly why DP matrix factorization, proven to help from-scratch SGD, does not transfer to LLM DP-LoRA fine-tuning.**

Contributions beyond the reference

DP-AdamBC contributes the analytic Φ-subtraction and demonstrates gains in its tested settings. Relative to it, this report contributes, and clearly highlights for the bonus, five new results (§5):

(1) the $\rho = \Phi/\hat{v}$ *noise-share diagnostic*; (2) the $\hat{m}/\hat{v}$ *mechanism*; (3) the *four-control attribution* showing DP-AdamBC = effective LR / momentum-SGD; (4) the *verified* κ *sensitivity* for $w_t = z_t - \lambda z_{t-1}$, correcting the wrong $\sqrt{1+\lambda^2}$ that under-noises and breaks privacy; and (5) the *signal-ceiling theorem* for first-moment denoising. The report is organized as: background and setup (§2), the mathematical theory with proofs (§3), the numerical simulation in a question-driven voice (§4), the highlighted new results (§5), limitations (§6), the contribution declaration (§8), and a self-evaluation (§9).

The spine question

> *The reference DP-AdamBC [1] subtracts the analytic DP-noise variance bias Φ from Adam's second moment $\hat{v}$. From an optimization view:* ***when*** *does this bias correction actually help in LLM DP-LoRA fine-tuning, and* ***can*** *a principled optimizer change—second-moment de-biasing, update geometry, or first-moment noise correlation—beat a well-tuned, privacy-amplified DP-Adam at the same (ε, δ)?*

Our answer. The noise share $\rho = \Phi/\hat{v} = \Phi/(v_{\text{true}} + \Phi)$ is an a-priori diagnostic. In the standard recipe ρ saturates at ≈ 1, so $\hat{v}$ is an inert noise floor and learning rides the first moment $\hat{m}$ (a prefix-sum); bias correction then collapses to momentum-SGD and is, at most, an effective-learning-rate knob; and because at $\rho \approx 1$ the model is *signal-limited, not variance-limited*, even a correctly-accounted anti-correlated-noise mechanism on the first-moment path (DP-CorrMom) cannot beat a tuned, amplified DP-Adam—a unified negative explained by a **signal ceiling**. The practical levers are privacy amplification and momentum, not optimizer cleverness.

2 Background and Setup

DP-SGD and DP-Adam. Given per-example losses $\{\ell_i\}_{i=1}^N$, DP-SGD samples a Poisson minibatch $\mathcal{B}$ at rate $q = B/N$, clips each per-sample gradient to ℓ_2-norm C, adds isotropic Gaussian noise of scale $\sigma_{\text{DP}}C$ (the noise multiplier σ_{DP} calibrates (ε, δ)), and mean-reduces: $g_{\text{DP}} = \frac{1}{B}\big[\sum_{i\in\mathcal{B}} \text{clip}_C(\nabla\ell_i) + z\big]$ with $z \sim \mathcal{N}(0, (\sigma_{\text{DP}}C)^2 I)$. DP-Adam feeds this g_{DP} into the standard bias-corrected Adam update $\Delta = -\text{lr}\,\hat{m}/(\sqrt{\hat{v}} + \epsilon)$, with first/second EMA moments $\hat{m}, \hat{v}$. The per-coordinate Gaussian mechanism injects variance $\Phi = (\sigma_{\text{DP}}C/B_{\text{eff}})^2$ of (1).

The DP-LoRA recipe. All experiments fine-tune with LoRA (rank r=16), clip C=0.1, and Opacus Poisson sampling with PRV accounting at δ=10^{-5}. Two model/task pairs: *RoBERTa-large / MNLI* (accuracy, an unambiguous metric) and *Qwen2.5-1.5B / E2E-NLG* (a teacher-forced argmax *proxy* BLEU). We log, per step, the noise share ρ, the analytic Φ, the measured median second moment $\hat{v}$, the clamp fraction (the share of coordinates where $(\hat{v} - \Phi)$ hits the floor ξ), and the median effective step.

The optimizer zoo (method codes). The empirical study compares eight optimizer variants, all sharing the DP-Adam learning rate tuned to the baseline (the conservative choice for surfacing a bias-correction gain):

`dp-adam` i.i.d.-Gaussian DP-SGD (Opacus clip+noise) feeding the standard bias-corrected Adam update $\hat{m}/(\sqrt{\hat{v}} + \epsilon)$; the well-tuned, amplified reference.

`dp-adambc` DP-AdamBC [1]: DP-Adam with $\hat{v}$ replaced by $\max(\hat{v} - \Phi, \xi)$, $\Phi = (\sigma_{\mathrm{DP}} C / B_{\mathrm{eff}})^2$; subtracts the second-moment bias.

`dp-adam-`ξ floor-only control: $\hat{v} \leftarrow \max(\hat{v}, \xi)$ with *no* Φ-subtraction; isolates the ξ-floor (step-size) effect from genuine de-biasing.

`muon-probe` a privacy-neutral diagnostic comparing cosine-to-clean-gradient of the noisy momentum M vs. its orthogonalization $\mathrm{msign}(M) = UV^\top$ (never applied, accountant untouched).

`dp-adam-lp` STEP-0 first-moment lever validator: a causal low-pass EMA on the already-private $\hat{m}$ (pure post-processing, zero extra ε).

`dp-corrmom` DP-CorrMom: anti-correlated DP noise $w_t = z_t - \lambda z_{t-1}$ on Adam's first-moment / prefix-sum path; per-step std inflated by $\kappa = \sqrt{(1 - \lambda^{2T})/(1 - \lambda^2)}$; λ=0 reproduces DP-SGD-momentum bit-for-bit.

`dp-corrsgd` DP-CorrSGD: the same anti-correlated noise on plain SGD (momentum 0), the textbook DP-FTRL/MF matched workload where the parameter trajectory *is* the gradient prefix-sum (no momentum-window mismatch, no $\sqrt{\hat{v}}$ denominator to poison).

Honesty notes (carried throughout). The E2E utility is a teacher-forced argmax proxy BLEU, not autoregressive generation; the positive-method sweeps (`dp-corrmom`, `dp-corrsgd`) are single-seed (consistent across five variants); and correlated noise voids Poisson amplification, so DP-CorrMom uses an *unamplified* single-participation (route-A) accountant. We calibrate each claim's strength to its evidence.

3 Mathematical Theory

This section develops, with full proofs, the optimization theory behind the spine question. The thread is a single scalar, $\Phi = (\sigma_{\mathrm{DP}} C / B_{\mathrm{eff}})^2$, shown to play three roles: it inflates Adam's second moment (Theorem 1), it governs the diagnostic ρ that decides whether bias correction is active (Theorem 3), and it is the irreducible DP term in the DP-SGD convergence floor (Theorem 5). We then turn to the first moment: it carries the learning at $\rho \approx 1$ (Theorem 6), the correlated-noise mechanism that denoises it admits a verified privacy sensitivity κ (Theorems 7–9), and yet a signal ceiling (Theorem 10) forbids first-moment denoising from helping at $\rho \approx 1$.

3.1 Second-moment inflation: $\mathbb{E}[\hat{v}_t] = v_t^{\mathrm{true}} + \Phi$

Theorem 1 (Second-moment inflation by the DP noise variance Φ)**.** *Consider DP-Adam fine-tuning with the per-step privatized gradient produced by an Opacus* `DPOptimizer` *(per-sample ℓ_2 clipping at threshold C, Gaussian noise with multiplier σ, mean reduction over the effective batch B_{eff}), fed into Adam's exponential-moving-average (EMA) second-moment estimator with bias correction. Fix one coordinate $j \in \{1, \ldots, d\}$ and let $\hat{v}_t$ denote the bias-corrected second moment at step t for that coordinate. Define the per-coordinate DP-noise variance*

$$\Phi \;:=\; \Big(\frac{\sigma\, C}{B_{\mathrm{eff}}}\Big)^2, \qquad B_{\mathrm{eff}} \;:=\; (\textit{expected batch size}) \times (\textit{accumulated iterations}).$$

Then, under Assumptions (A1)–(A4) *below,*

$$\boxed{\mathbb{E}[\hat{v}_t] \;=\; v_t^{\mathrm{true}} \;+\; \Phi} \qquad \textit{for every } t \geq 1,$$

where $v_t^{\text{true}} := \dfrac{(1-\beta_2)\sum_{k=0}^{t-1}\beta_2^k\,\mathbb{E}[\bar{g}_{t-k}^2]}{1-\beta_2^t}$ *is the bias-corrected EMA of the squared* clipped-mean *gradient (the noise-free target). That is, the DP Gaussian noise inflates Adam's per-coordinate second-moment estimate by* exactly *the injected per-coordinate noise variance* Φ*, independently of the step index* t*, of the EMA decay* β_2*, and of the gradient signal. Consequently the DP-AdamBC correction* $\hat{v}_t^{\text{BC}} = \hat{v}_t - \Phi$ *is, in expectation, an exact debiasing of the second moment, and* Φ *equals the closed form computed in* `dp_adaptive.DPAdaptive.phi`*.*

Proof. We make explicit the modelling assumptions, identify the exact per-coordinate noise variance Φ baked into the gradient that reaches the optimizer, and then propagate it through the bias-corrected EMA recursion. Throughout, all quantities are for a single fixed coordinate j; we suppress j to lighten notation. Expectations are taken over the DP Gaussian noise (and, where stated, over the Poisson minibatch sampling), conditioning the clipped signal as described.

Assumptions.

(A1) Opacus clip-and-noise pipeline. At step t the optimizer receives, in `p.grad`, the gradient produced by `DPOptimizer.pre_step()`, namely

$$g_t \;=\; \frac{1}{B_{\text{eff}}}\Big(\underbrace{\textstyle\sum_{i\in\mathcal{B}_t}\operatorname{clip}_C(\nabla\ell_i)}_{=:S_t} + z_t\Big), \qquad z_t \sim \mathcal{N}\big(0,\,(\sigma C)^2 I_d\big),$$

where $\operatorname{clip}_C(v) = v\cdot\min\{1, C/\|v\|_2\}$ is the per-sample ℓ_2 clip, σ is the noise multiplier (`noise_multiplier`), C is the clipping norm (`max_grad_norm`), and B_{eff} = `expected_batch_size` × `accumulated_iterations` is the mean-reduction denominator (`loss_reduction="mean"`). This is exactly the Opacus mechanism: `clip_and_accumulate` forms S_t, `add_noise` adds z_t, and `scale_grad` divides by B_{eff}. The Gaussian standard deviation σC is the gradient ℓ_2-sensitivity C times the multiplier σ, which is what calibrates the (ε,δ) guarantee. Crucially, the noise z_t is injected onto the *sum* S_t and then the whole quantity is divided by B_{eff}, so the noise inherits the $1/B_{\text{eff}}$ factor (not $1/B_{\text{eff}}^{1/2}$); this is what makes Φ scale as $1/B_{\text{eff}}^2$.

(A2) Noise model. The DP noise z_t is drawn from an isotropic Gaussian with independent coordinates, and is independent of the data and of the clipped sum S_t (it is freshly sampled inside `add_noise`, after clipping). In particular, for the fixed coordinate, the contribution to g_t is

$$\xi_t \;:=\; \frac{z_t}{B_{\text{eff}}} \;\sim\; \mathcal{N}(0,\,\Phi), \qquad \Phi := \operatorname{Var}(\xi_t) = \frac{(\sigma C)^2}{B_{\text{eff}}^2} = \Big(\frac{\sigma C}{B_{\text{eff}}}\Big)^2.$$

Write also $\bar{g}_t := S_t/B_{\text{eff}}$ for the (data-dependent) clipped-mean signal in that coordinate, so that

$$g_t \;=\; \bar{g}_t + \xi_t, \qquad \xi_t \perp \bar{g}_t, \qquad \mathbb{E}[\xi_t \mid \bar{g}_t] = 0, \qquad \mathbb{E}[\xi_t^2 \mid \bar{g}_t] = \Phi.$$

Scope note (what is and is not used). Only two facts about ξ_t enter the proof: for each fixed t, ξ_t is centered with conditional variance Φ given the *same-step* signal $\bar{g}_t$. Independence of ξ_t *across* distinct steps is *not* required for the mean identity, because (as Step 3 shows) no product $\xi_{t-k}\xi_{t-k'}$ with $k \neq k'$ ever appears; only the marginal second moment of each squared term is used. We retain "independent across steps" merely because it holds in the code (fresh draws) and is needed for any *variance*/concentration statement, not for the expectation here.

(A3) EMA second moment with bias correction. Adam maintains, per coordinate,

$$v_t \;=\; \beta_2\, v_{t-1} + (1-\beta_2)\, g_t^2, \qquad v_0 = 0, \qquad \hat{v}_t \;=\; \frac{v_t}{1-\beta_2^t},$$

with decay $\beta_2 \in (0,1)$. This matches `v.mul_(b2).addcmul_(g,g,value=1-b2)` followed by `vhat = v / (1 - b2**t)` in `_adam_update`, with the standard zero initialization $v_0 = 0$ (`state["exp_avg_sq"] = torch.zeros_like(p)`).

(A4) Stationarity of the mechanism within a step. σ, C and B_{eff} are fixed for the run (equivalently, fixed at each step), so Φ is a constant across the t steps that enter v_t. (If they vary, the statement holds per step with the obvious step-indexed Φ_k in place of Φ in (4); we take them constant, as in the code where `phi` is a single scalar property.)

Step 1: each squared gradient carries an additive bias Φ in expectation. Fix any step index s and condition on the clipped-mean signal $\bar{g}_s$. Using $g_s = \bar{g}_s + \xi_s$ and expanding the square,

$$\mathbb{E}\Big[g_s^2 \,\Big|\, \bar{g}_s\Big] \;=\; \mathbb{E}\Big[\bar{g}_s^2 + 2\,\bar{g}_s\,\xi_s + \xi_s^2 \,\Big|\, \bar{g}_s\Big] \;=\; \bar{g}_s^2 + 2\,\bar{g}_s\,\underbrace{\mathbb{E}[\xi_s \mid \bar{g}_s]}_{=0} + \underbrace{\mathbb{E}[\xi_s^2 \mid \bar{g}_s]}_{=\Phi} \;=\; \bar{g}_s^2 + \Phi,$$

where the first two summands are pulled out because $\bar{g}_s$ is $\sigma(\bar{g}_s)$-measurable, the cross term vanishes because $\mathbb{E}[\xi_s \mid \bar{g}_s] = 0$ (A2), and the last term equals Φ by the conditional-variance identity of (A2). Taking the outer expectation over $\bar{g}_s$ and using the tower property $\mathbb{E}[\cdot] = \mathbb{E}[\mathbb{E}[\cdot \mid \bar{g}_s]]$,

$$\mathbb{E}[g_s^2] \;=\; \mathbb{E}[\bar{g}_s^2] + \Phi \qquad \text{for every } s \geq 1. \tag{2}$$

This is a statement about the *marginal* second moment of g_s alone; it involves no other step. Thus each squared gradient that enters the EMA carries the *same* additive bias Φ on top of the noise-free target $\mathbb{E}[\bar{g}_s^2]$. (Note $\mathbb{E}[\bar{g}_s^2]$ itself contains the minibatch-sampling variance of $\bar{g}_s$; this is part of v^{true} and is precisely what BC must *not* remove.)

Step 2: solve the EMA recursion in closed form. With $v_0 = 0$, unrolling $v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$ gives the standard geometric sum

$$v_t \;=\; (1-\beta_2) \sum_{k=0}^{t-1} \beta_2^k\, g_{t-k}^2. \tag{3}$$

We verify (3) by induction on t. Base case $t = 1$: $v_1 = \beta_2 v_0 + (1-\beta_2) g_1^2 = (1-\beta_2) g_1^2$, matching the sum with a single term $k = 0$. Inductive step: assume (3) at t; then

$$v_{t+1} = \beta_2 v_t + (1-\beta_2) g_{t+1}^2 = (1-\beta_2) \sum_{k=0}^{t-1} \beta_2^{k+1} g_{t-k}^2 + (1-\beta_2) g_{t+1}^2 = (1-\beta_2) \sum_{k=0}^{t} \beta_2^k g_{(t+1)-k}^2,$$

where the last equality re-indexes $k \mapsto k+1$ on the first sum (covering $k = 1, \ldots, t$, with $g_{t-(k-1)} = g_{(t+1)-k}$) and absorbs the g_{t+1}^2 term as the $k = 0$ summand. This is (3) at $t+1$, completing the induction.

Step 3: take expectations and apply the per-term bias. The unrolled v_t in (3) is a *linear combination of single squared terms* g_{t-k}^2; in particular it contains no cross products $g_{t-k} g_{t-k'}$

($k \neq k'$). Hence, by linearity of expectation, only the marginal second moments $\mathbb{E}[g_{t-k}^2]$ are needed, and we may substitute (2) term by term with $s = t - k$:

$$\mathbb{E}[v_t] = (1-\beta_2)\sum_{k=0}^{t-1}\beta_2^k\,\mathbb{E}[g_{t-k}^2] = (1-\beta_2)\sum_{k=0}^{t-1}\beta_2^k\Big(\mathbb{E}[\bar{g}_{t-k}^2] + \Phi\Big).$$

(No assumption about dependence between ξ_{t-k} and $\xi_{t-k'}$ is invoked, by the scope note in A2.) Split the sum into the signal part and the bias part:

$$\mathbb{E}[v_t] = \underbrace{(1-\beta_2)\sum_{k=0}^{t-1}\beta_2^k\,\mathbb{E}[\bar{g}_{t-k}^2]}_{=:\,v_t^{\text{true, raw}}} + \Phi\,(1-\beta_2)\sum_{k=0}^{t-1}\beta_2^k. \tag{4}$$

The bias multiplier is a finite geometric series; since $\beta_2 \neq 1$,

$$(1-\beta_2)\sum_{k=0}^{t-1}\beta_2^k = (1-\beta_2)\cdot\frac{1-\beta_2^t}{1-\beta_2} = 1-\beta_2^t. \tag{5}$$

Substituting (5) into (4),

$$\mathbb{E}[v_t] = v_t^{\text{true, raw}} + (1-\beta_2^t)\,\Phi. \tag{6}$$

Step 4: apply the bias correction $1/(1-\beta_2^t)$. Since $0 < \beta_2 < 1$ we have $1-\beta_2^t > 0$ for all $t \geq 1$, so dividing (6) by $1-\beta_2^t$ is well defined. Using $\hat{v}_t = v_t/(1-\beta_2^t)$ (A3) and linearity,

$$\mathbb{E}[\hat{v}_t] = \frac{\mathbb{E}[v_t]}{1-\beta_2^t} = \frac{v_t^{\text{true, raw}}}{1-\beta_2^t} + \frac{(1-\beta_2^t)\,\Phi}{1-\beta_2^t} = \underbrace{\frac{v_t^{\text{true, raw}}}{1-\beta_2^t}}_{=:\,v_t^{\text{true}}} + \Phi.$$

By the definition of $v_t^{\text{true, raw}}$ in (4), the first term is exactly

$$v_t^{\text{true}} = \frac{(1-\beta_2)\sum_{k=0}^{t-1}\beta_2^k\,\mathbb{E}[\bar{g}_{t-k}^2]}{1-\beta_2^t},$$

the bias-corrected EMA of the squared clipped-mean (noise-free) gradient, as claimed. Hence

$$\mathbb{E}[\hat{v}_t] \;=\; v_t^{\text{true}} + \Phi, \qquad \text{for every } t \geq 1.$$

Step 5: identification of Φ with the code and consequences. The bias is the *constant* Φ, with no residual t- or β_2-dependence: the bias-correction factor $1/(1-\beta_2^t)$ cancels exactly the factor $(1-\beta_2^t)$ that the EMA places on Φ in (6). This cancellation is the whole point of subtracting the *constant* Φ (not $(1-\beta_2^t)\Phi$) in DP-AdamBC. By (A2),

$$\Phi = \Big(\frac{\sigma C}{B_{\text{eff}}}\Big)^2 = \Big(\frac{\texttt{noise\_multiplier}\cdot\texttt{max\_grad\_norm}}{\texttt{expected\_batch\_size}\cdot\texttt{accumulated\_iterations}}\Big)^2,$$

which is line-for-line the scalar returned by `DPAdaptive.phi` (this identification is a consequence of A1–A2 matching the code, not an additional probabilistic hypothesis). Two immediate corollaries:

1. *Exact debiasing in expectation.* $\mathbb{E}[\hat{v}_t - \Phi] = v_t^{\text{true}}$, so the DP-AdamBC update $\hat{v}_t^{\text{BC}} = \hat{v}_t - \Phi$ (before any numerical floor ξ) removes precisely the injected-noise inflation and restores Adam's intended preconditioner geometry in expectation.

2. *Degenerate case* $\sigma = 0$. If `noise_multiplier`$= 0$ then $\Phi = 0$ and $\mathbb{E}[\hat{v}_t] = v_t^{\text{true}}$; subtracting Φ is a no-op, consistent with the code comment that DP-AdamBC then reduces to stock Adam.

Remarks (scope and tightness). (i) The result is an identity in expectation; it does not require unbiasedness of $\bar{g}_t$ as an estimator of ∇F. Per-sample clipping may bias $\bar{g}_t$, but that bias lives *inside* v_t^{true} (through $\mathbb{E}[\bar{g}_t^2]$) and is exactly what BC must preserve; the theorem isolates only the additive DP-*noise* inflation Φ. (ii) The only probabilistic inputs are the same-step conditional-mean and conditional-variance of ξ_t given $\bar{g}_t$ (A2); they hold because the Gaussian is drawn after, and independently of, clipping. The proof does *not* use cross-step noise independence for the mean identity (it would be needed only to control the variance of $\hat{v}_t$). (iii) Distributed (DDP) training replaces B_{eff} by the global batch `expected_batch_size` $\cdot$ `accumulated_iterations` $\cdot$ `world_size` because `reduce_gradients` applies an extra $1/$`world_size` mean reduction to the rank-0 noise; substituting this B_{eff} leaves the statement and proof verbatim, matching `DistributedDPAdaptive.phi`. □

3.2 The ρ-diagnostic for DP-AdamBC

Definition 2 (Per-coordinate DP-noise variance)**.** Fix a trainable coordinate of the DP-LoRA model. Following the DP-Adam update populated by Opacus (see `src/dp_optim/dp_adaptive.py`, the `phi` property), the privatized minibatch gradient seen by the optimizer is

$$g_t \;=\; \bar{g}_t + \xi_t, \qquad \xi_t \sim \mathcal{N}(0, \Phi), \qquad \Phi \;:=\; \Big(\frac{\sigma_{\text{DP}}\, C}{B_{\text{eff}}}\Big)^2,$$

where $\bar{g}_t = \frac{1}{B_{\text{eff}}} \sum_{i \in \mathcal{B}_t} \text{clip}_C(\nabla \ell_i)$ is the clipped, batch-averaged (conditionally deterministic) gradient mean, σ_{DP} is the noise multiplier, C the clipping norm, and $B_{\text{eff}} =$ `expected_batch_size` $\cdot$ `accumulated_iterations` ($\cdot$ `world_size`) the effective batch size. The DP noise ξ_t is drawn independently of $\bar{g}_t$ and is zero-mean with the same variance Φ on every coordinate.

Theorem 3 (The ρ-diagnostic for DP-AdamBC)**.** *Let $\hat{v}$ denote Adam's bias-corrected second-moment estimate at the coordinate, and let*

$$v_{\text{true}} \;:=\; \mathbb{E}[\hat{v}] - \Phi$$

be its noise-free counterpart, with $v_{\text{true}} \geq 0$ and $\Phi > 0$. Define the noise share

$$\rho \;:=\; \frac{\Phi}{\mathbb{E}[\hat{v}]} \;=\; \frac{\Phi}{v_{\text{true}} + \Phi} \;\in\; (0, 1].$$

Then:

(i) ***(Inflation / debiasing)*** *$\mathbb{E}[\hat{v}] = v_{\text{true}} + \Phi$, so the DP-AdamBC corrected denominator recovers the signal second moment in expectation: $\mathbb{E}[\hat{v} - \Phi] = v_{\text{true}}$.*

(ii) ***(Step-correction factor)*** *In expectation (in the ratio-of-expected-denominators sense made precise below) the multiplicative change DP-AdamBC applies to the per-coordinate Adam step is*

$$\kappa_{\text{BC}}(\rho) \;:=\; \frac{\sqrt{\mathbb{E}[\hat{v}]}}{\sqrt{v_{\text{true}}}} \;=\; \frac{1}{\sqrt{1-\rho}},$$

with relative change $g(\rho) := \kappa_{\text{BC}}(\rho) - 1 = (1-\rho)^{-1/2} - 1$.

(iii) ***(Materiality)*** *Writing* $r := \Phi/v_{\text{true}} = \rho/(1-\rho) \geq 0$*, the correction is an* order-one *relative change iff* $r = \Theta(1)$*, i.e. iff* $v_{\text{true}} = \Theta(\Phi)$*, equivalently iff* ρ *is bounded away from both* 0 *and* 1*. Quantitatively* $g(\rho)$ *is sandwiched,* for all $r \geq 0$*, as*

$$\tfrac{1}{2}\, r \;-\; \tfrac{1}{2} r^2 \;\leq\; g(\rho) \;\leq\; \tfrac{1}{2}\, r, \qquad r = \frac{\rho}{1-\rho},$$

so $g(\rho) \to 0$ *as* $\rho \to 0$ *(BC inert) and* $g(\rho) \to \infty$ *as* $\rho \to 1$*.*

(iv) ***(Maximal sensitivity / sweet spot)*** *On the natural scale* $u := \log r = \log(\Phi/v_{\text{true}})$ *the noise share* $\rho(u) = \dfrac{e^u}{1+e^u}$ *is the logistic sigmoid; its sensitivity* $\dfrac{d\rho}{du} = \rho(1-\rho)$ *attains its unique maximum* $\frac{1}{4}$ *at*

$$\rho = \tfrac{1}{2} \iff \Phi = v_{\text{true}} \iff r = 1.$$

Thus $\rho = \frac{1}{2}$ *is the diagnostic's most sensitive operating point: the crossover where the DP-noise variance exactly equals the true second moment.*

(v) ***(Saturation)*** $\rho \to 1$ *monotonically as* $v_{\text{true}} \to 0^+$*; equivalently* ρ *is strictly decreasing in* v_{true} *for fixed* Φ*, with* $\rho = 1$ *attained only in the limit* $v_{\text{true}} = 0$*.*

Proof. Throughout, fix one trainable coordinate; all quantities are scalar. We use Definition 2: $g_t = \bar{g}_t + \xi_t$ with $\xi_t \sim \mathcal{N}(0, \Phi)$ independent of the clipped mean $\bar{g}_t$, and $\Phi = (\sigma_{\text{DP}} C / B_{\text{eff}})^2 > 0$. Adam maintains the exponential moving average $v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$ ($v_0 = 0$) with bias correction $\hat{v}_t = v_t/(1-\beta_2^t)$, exactly as implemented in `DPAdaptive._adam_update`. We suppress the time index t where it is inessential and write $\mathbb{E}[\hat{v}]$ for the conditional expectation over the DP noise $\{\xi_s\}_{s \leq t}$ given the (deterministic) clipped means $\{\bar{g}_s\}_{s \leq t}$; the unconditional statement then follows by the tower property, since Φ is a fixed constant and the quantities below are affine in $\{\bar{g}_s^2\}$.

(i) Inflation and exact debiasing in expectation. We first establish the inflation identity $\mathbb{E}[\hat{v}] = v_{\text{true}} + \Phi$ with

$$v_{\text{true}} \;:=\; \frac{(1-\beta_2)\sum_{k=0}^{t-1} \beta_2^k\, \bar{g}_{t-k}^2}{1-\beta_2^t},$$

the bias-corrected EWMA of the *clipped, noise-free* squared gradients (note $v_{\text{true}} \geq 0$ since every summand is nonnegative). Conditioning on $\bar{g}_s$ and using $\mathbb{E}[\xi_s] = 0$, $\mathbb{E}[\xi_s^2] = \Phi$, and the independence of ξ_s from $\bar{g}_s$,

$$\mathbb{E}\Big[g_s^2 \mid \bar{g}_s\Big] = \mathbb{E}\Big[(\bar{g}_s + \xi_s)^2 \mid \bar{g}_s\Big] = \bar{g}_s^2 + 2\bar{g}_s\, \mathbb{E}[\xi_s] + \mathbb{E}[\xi_s^2] = \bar{g}_s^2 + \Phi,$$

because the cross term $2\bar{g}_s\, \mathbb{E}[\xi_s] = 0$ vanishes. Unrolling the linear EWMA recursion from $v_0 = 0$ gives $v_t = (1-\beta_2)\sum_{k=0}^{t-1} \beta_2^k\, g_{t-k}^2$, hence by linearity of expectation

$$\mathbb{E}[v_t] = (1-\beta_2)\sum_{k=0}^{t-1} \beta_2^k\, (\bar{g}_{t-k}^2 + \Phi) = (1-\beta_2)\sum_{k=0}^{t-1} \beta_2^k\, \bar{g}_{t-k}^2 \;+\; \Phi\,(1-\beta_2)\sum_{k=0}^{t-1} \beta_2^k.$$

The geometric sum evaluates to $(1-\beta_2)\sum_{k=0}^{t-1} \beta_2^k = (1-\beta_2)\cdot \frac{1-\beta_2^t}{1-\beta_2} = 1-\beta_2^t$. Dividing by the bias-correction factor $1-\beta_2^t$ gives

$$\mathbb{E}[\hat{v}_t] = \frac{\mathbb{E}[v_t]}{1-\beta_2^t} = v_{\text{true}} + \Phi.$$

The crucial point is that after bias correction the additive contamination is the *constant* Φ (the factor $1-\beta_2^t$ has cancelled), which is exactly what makes a constant subtraction the correct debiasing. DP-AdamBC (1; `bias_correction=True` in `_adam_update`, where $\hat{v} \leftarrow \max(\hat{v}-\Phi, \xi)$) therefore satisfies, whenever the floor clamp is inactive,

$$\mathbb{E}\big[\hat{v}-\Phi\big] = \mathbb{E}[\hat{v}] - \Phi = (v_{\text{true}}+\Phi) - \Phi = v_{\text{true}},$$

i.e. the corrected denominator recovers the signal second moment v_{true} in expectation. This proves (i). In particular $\rho = \Phi/\mathbb{E}[\hat{v}] = \Phi/(v_{\text{true}}+\Phi)$ is well defined and, since $v_{\text{true}} \geq 0$ and $\Phi > 0$, lies in $(0,1]$, with $\rho = 1$ iff $v_{\text{true}} = 0$.

(ii) The step-correction factor $\kappa_{\text{BC}}(\rho) = (1-\rho)^{-1/2}$. Adam's per-coordinate update has magnitude proportional to $\hat{m}/(\sqrt{\hat{v}}+\epsilon)$ (`p.addcdiv_(mhat, vhat.sqrt().add_(eps), -lr)`). With $\hat{m}$ held fixed (BC changes only the denominator) and the regularizer ϵ negligible against $\sqrt{\hat{v}}$, we compare the *deterministic* surrogate steps obtained by replacing the random denominator $\sqrt{\hat{v}}$ by the square root of its expectation, $\sqrt{\mathbb{E}[\hat{v}]}$ (the inflated denominator) versus $\sqrt{\mathbb{E}[\hat{v}]-\Phi} = \sqrt{v_{\text{true}}}$ (the debiased denominator). Their ratio is

$$\kappa_{\text{BC}} := \frac{\text{step}_{\text{BC}}}{\text{step}_{\text{no BC}}} = \frac{\sqrt{\mathbb{E}[\hat{v}]}}{\sqrt{\mathbb{E}[\hat{v}]-\Phi}} = \sqrt{\frac{v_{\text{true}}+\Phi}{v_{\text{true}}}} = \sqrt{\frac{1}{1-\rho}},$$

where the last equality uses $v_{\text{true}} = (1-\rho)\,\mathbb{E}[\hat{v}]$, which follows from $\rho = \Phi/\mathbb{E}[\hat{v}]$ and $v_{\text{true}} = \mathbb{E}[\hat{v}] - \Phi$. Hence $\kappa_{\text{BC}}(\rho) = (1-\rho)^{-1/2}$ and $g(\rho) = \kappa_{\text{BC}}(\rho) - 1 = (1-\rho)^{-1/2} - 1$. This is the amount by which DP-AdamBC enlarges the noise-shrunk Adam step at the level of expected denominators. (At the empirically reachable point $\rho = 0.955$ this gives $\kappa_{\text{BC}} = 1/\sqrt{1-0.955} = 4.71$, consistent with the measured ≈ 3.8–$4.7\times$ effective-step increase; as $\rho \to 1$ it diverges, reflecting the unstable over-subtraction regime.)

Scope of the claim. κ_{BC} is defined as the ratio of the square roots of the *expected* denominators, i.e. a plug-in (first-moment / delta-method) approximation to the random per-step ratio $\sqrt{\hat{v}}/\sqrt{\hat{v}-\Phi}$. This is precisely the quantity the project's telemetry measures (medians of $\sqrt{\hat{v}}$). For the strictly stochastic per-step ratio one would invoke Jensen: since $x \mapsto 1/\sqrt{x}$ is convex on $(0,\infty)$, $\mathbb{E}[1/\sqrt{\hat{v}-\Phi}] \geq 1/\sqrt{\mathbb{E}[\hat{v}-\Phi]} = 1/\sqrt{v_{\text{true}}}$, so the true expected enlargement is *at least* κ_{BC}; the plug-in value is a lower bound and the Jensen gap only sharpens the $\rho \to 1$ instability conclusion. This proves (ii) in the stated sense.

(iii) Materiality: order-one change iff $v_{\text{true}} = \Theta(\Phi)$. Substitute $\rho = r/(1+r)$, so $1-\rho = 1/(1+r)$ and $\kappa_{\text{BC}} = \sqrt{1+r}$, hence

$$g(\rho) = \sqrt{1+r} - 1, \qquad r = \frac{\Phi}{v_{\text{true}}} = \frac{\rho}{1-\rho} \;\geq 0.$$

We bound $h(r) := \sqrt{1+r} - 1$ on $[0,\infty)$.

Upper bound. The map $r \mapsto \sqrt{1+r}$ is concave on $[0,\infty)$, so it lies below its tangent line at $r=0$: $\sqrt{1+r} \leq 1 + \frac{1}{2}r$. Subtracting 1,

$$g(\rho) = \sqrt{1+r} - 1 \;\leq\; \tfrac{1}{2}\,r \qquad (\forall\, r \geq 0).$$

Lower bound. We claim $g(\rho) = \sqrt{1+r} - 1 \geq \frac{1}{2}r - \frac{1}{2}r^2$ for *all* $r \geq 0$, and we split into two cases. *Case* $r > 1$. Then $\frac{1}{2}r - \frac{1}{2}r^2 = \frac{1}{2}r(1-r) < 0 \leq g(\rho)$, so the bound holds trivially.

Case $0 \le r \le 1$. Set $a := 1 + \frac{1}{2}r - \frac{1}{2}r^2$. On $[0,1]$ we have $a = 1 + \frac{1}{2}r(1-r) \ge 1 > 0$. It therefore suffices to show $a^2 \le 1 + r$ (then $a \le \sqrt{1+r}$, i.e. $g(\rho) = \sqrt{1+r} - 1 \ge a - 1 = \frac{1}{2}r - \frac{1}{2}r^2$). Expanding,

$$a^2 = \left(1 + \tfrac{1}{2}r - \tfrac{1}{2}r^2\right)^2 = 1 + r - \tfrac{3}{4}r^2 - \tfrac{1}{2}r^3 + \tfrac{1}{4}r^4,$$

so

$$(1+r) - a^2 = \tfrac{3}{4}r^2 + \tfrac{1}{2}r^3 - \tfrac{1}{4}r^4 = \frac{r^2}{4}\,(3 + 2r - r^2) = \frac{r^2}{4}\,(3-r)(1+r) \;\ge\; 0$$

for all $r \in [0,3]$, in particular on $[0,1]$. (This identity makes precise where the squared inequality is valid: exactly for $r \le 3$; we only need $r \le 1$.) Hence $a^2 \le 1 + r$ on $[0,1]$, giving the lower bound there.

Combining the two cases with the upper bound,

$$\tfrac{1}{2}r - \tfrac{1}{2}r^2 \;\le\; g(\rho) \;\le\; \tfrac{1}{2}r \qquad (\forall\, r \ge 0).$$

In particular, for bounded r the two bounds pinch $g(\rho) = \frac{1}{2}r + O(r^2)$, so $g(\rho) = \Theta(r)$, and:

- $g(\rho) = \Theta(1)$ (a genuine order-one change to the step) *iff* $r = \Theta(1)$, i.e. $v_{\text{true}} = \Theta(\Phi)$, i.e. ρ bounded away from 0 and 1;
- $g(\rho) \le \frac{1}{2}r \to 0$ as $\rho \to 0$ ($\Phi \ll v_{\text{true}}$): DP-AdamBC and DP-Adam coincide to first order, so bias correction is a no-op;
- $g(\rho) = (1-\rho)^{-1/2} - 1 \to \infty$ as $\rho \to 1$ ($v_{\text{true}} \ll \Phi$): the correction is unboundedly large, the regime where $\hat{v} - \Phi \to 0$ and the floor ξ takes over (the "collapse to momentum-SGD" regime documented in the experiments).

This proves (iii): bias correction is material exactly when $v_{\text{true}} = \Theta(\Phi)$. (The earlier project bound $|\kappa_{\text{BC}} - 1| \le r/2$ in `theory.tex` is exactly the upper half established here.)

(iv) Maximal sensitivity at $\rho = \frac{1}{2}$. The materiality threshold $r = \Theta(1)$ singles out a multiplicative scale; the canonical coordinate is $u := \log r = \log(\Phi / v_{\text{true}})$, the log noise-to-signal ratio (well defined for $v_{\text{true}} > 0$, i.e. $\rho \in (0,1)$). From $\rho = r/(1+r)$ with $r = e^u$,

$$\rho(u) = \frac{e^u}{1+e^u} = \frac{1}{1+e^{-u}},$$

the logistic sigmoid, which is smooth and strictly increasing from $\rho(-\infty) = 0$ to $\rho(+\infty) = 1$. Its sensitivity is

$$\frac{d\rho}{du} = \frac{e^u(1+e^u) - e^u \cdot e^u}{(1+e^u)^2} = \frac{e^u}{(1+e^u)^2} = \rho(u)\,(1-\rho(u)).$$

Maximizing $s(\rho) := \rho(1-\rho)$ over $\rho \in (0,1)$: $s'(\rho) = 1 - 2\rho = 0 \iff \rho = \frac{1}{2}$, and $s''(\rho) = -2 < 0$ confirms a strict interior maximum, with value $s(\frac{1}{2}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Equivalently, $\frac{d^2\rho}{du^2} = \frac{d}{du}\left[\rho(1-\rho)\right] = (1-2\rho)\frac{d\rho}{du} = \rho(1-\rho)(1-2\rho) = 0 \iff \rho = \frac{1}{2}$, the inflection point of the sigmoid. At this point

$$\rho = \tfrac{1}{2} \iff r = e^u = 1 \iff \Phi = v_{\text{true}}.$$

Hence $\rho = \frac{1}{2}$ is the unique operating point of maximal diagnostic sensitivity: a small change in the log noise/signal ratio produces the largest change in ρ, and it is exactly the crossover where the DP-noise variance equals the true second moment $\Phi = v_{\text{true}}$. This is the "sweet spot": for $\rho < \frac{1}{2}$ the

noise is sub-dominant and BC is a small correction ($g \le \frac{1}{2}r < \frac{1}{2}$); for $\rho > \frac{1}{2}$ the noise dominates and the correction g grows past 1 and then diverges as $\rho \to 1$, so the well-behaved order-one regime is centered at $\rho = \frac{1}{2}$. This proves (iv).

(v) Saturation $\rho \to 1$ as $v_{\text{true}} \to 0$. With $\Phi > 0$ fixed, view $\rho = \Phi/(v_{\text{true}} + \Phi)$ as a function of $v_{\text{true}} \ge 0$. Then

$$\frac{\partial \rho}{\partial v_{\text{true}}} = -\frac{\Phi}{(v_{\text{true}} + \Phi)^2} < 0,$$

so ρ is strictly decreasing in v_{true}. As $v_{\text{true}} \to 0^+$ we get $\rho \to \Phi/\Phi = 1$; as $v_{\text{true}} \to \infty$, $\rho \to 0$. Thus ρ saturates monotonically at its ceiling 1 precisely when the signal second moment falls to zero, $\rho = 1$ being attained only in the limit $v_{\text{true}} = 0$. This is the regime measured throughout the experiments: $\rho = 1.00$–1.07 across RoBERTa-large/MNLI ($\varepsilon \in \{1, 3, 8\}$, $B \in [16, 2048]$) and Qwen2.5-1.5B/E2E ($B \le 512$), i.e. v_{true} lies below the DP-noise floor, placing the optimizer at $\rho \approx 1$ where, by (iii), $g(\rho)$ is large but the corrected denominator $\hat{v} - \Phi \to 0$ and the floor ξ governs the step. The logged ratio saturating at 1 (rather than at the sweet spot $\frac{1}{2}$) is therefore the diagnostic signature that bias correction cannot act as intended in this recipe.

(Measured ratios slightly exceeding 1. The reported $\rho \in [1.00, 1.07]$ takes values just above the theoretical ceiling 1; this is consistent with the theory because the logged quantity is $\Phi/\operatorname{median}(\hat{v})$ over a finite sample of coordinates rather than $\Phi/\mathbb{E}[\hat{v}]$, and a downward sampling/median fluctuation of $\hat{v}$ below its mean $v_{\text{true}} + \Phi \approx \Phi$ pushes the ratio marginally above 1. The population quantity $\rho = \Phi/\mathbb{E}[\hat{v}]$ is bounded by 1 as proved.) This proves (v) and completes the proof. □

Remark (consistency of the floored estimator). The above treats the regime where DP-AdamBC's floor clamp $\hat{v} \leftarrow \max(\hat{v} - \Phi, \xi)$ is inactive, which is exactly where the debiasing identity $\mathbb{E}[\hat{v} - \Phi] = v_{\text{true}}$ holds. When $\rho \to 1$ the residual $\hat{v} - \Phi$ falls below ξ on a fraction of coordinates approaching 1 (the measured clamp fraction $\to 1$), and the clamp introduces a Jensen gap $\mathbb{E}[\max(\hat{v} - \Phi, \xi)] \ge \max(v_{\text{true}}, \xi)$; there the estimator is biased upward toward the constant ξ and the update degenerates to $\hat{m}/\sqrt{\xi}$, i.e. momentum-SGD with fixed step scale $1/\sqrt{\xi}$. This is consistent with (iii): the unbounded $g(\rho)$ as $\rho \to 1$ is not realized as a useful step but is absorbed by the floor, which is why the maximal-sensitivity sweet spot $\rho = \frac{1}{2}$ — not the saturation point $\rho = 1$ — is the regime in which bias correction can be both material *and* well-conditioned.

Remark (privacy-side cross-check of the companion κ). The diagnostic above governs the *second*-moment lever (DP-AdamBC). The project's positive-method attempt (DP-CorrMom) instead acts on the *first*-moment / prefix-sum path by injecting anti-correlated noise $w_t = z_t - \lambda z_{t-1}$, whose privacy sensitivity is the distinct scalar $\kappa(\lambda, T) = \text{corr\_sensitivity}(\lambda, T)$ in `dp_adaptive.py`. We independently re-derived it to confirm the implemented value. Writing the injection as $n = Lz$ with L lower-bidiagonal (1 on the diagonal, $-\lambda$ on the sub-diagonal), the matrix mechanism for the prefix-sum workload A uses strategy $\mathbf{C} = L^{-1}$, the lower-triangular Toeplitz matrix with entries $(\mathbf{C})_{ij} = \lambda^{i-j}$ for $i \ge j$ (and 0 otherwise). The single-participation ℓ_2 sensitivity is the maximum column norm of $\mathbf{C}$. Column j has entries $\lambda^0, \lambda^1, \ldots, \lambda^{T-1-j}$ down the column, so

$$\|\mathbf{C}e_j\|_2^2 = \sum_{k=0}^{T-1-j} \lambda^{2k},$$

which is maximized at $j = 0$ (the first column, the longest run), giving

$$\kappa(\lambda, T) = \max_j \|\mathbf{C}e_j\|_2 = \Big(\sum_{k=0}^{T-1} \lambda^{2k}\Big)^{1/2} = \sqrt{\frac{1 - \lambda^{2T}}{1 - \lambda^2}} \xrightarrow[T\to\infty]{} \frac{1}{\sqrt{1 - \lambda^2}}.$$

A Mahalanobis cross-check confirms this: the privatized output is $\hat{g} \sim \mathcal{N}(g, \nu^2 LL^\top)$, and a unit change in the data at step t gives Mahalanobis separation $\propto (LL^\top)^{-1}_{tt} = \|(L^{-1})e_t\|_2^2$, again the squared column norm of L^{-1}. We verified numerically (e.g. $\lambda = 0.9$, $T = 8$) that the exact maximum column norm of L^{-1} equals $\sqrt{(1-\lambda^{2T})/(1-\lambda^2)} = 2.0707$ and that the large-T limit is $1/\sqrt{1-\lambda^2} = 2.2942$. The naive alternative $\sqrt{1+\lambda^2} = 1.345$ is the column norm of L *itself* (not L^{-1}); using it would *under*-noise the mechanism and *break* the privacy guarantee. Thus the implemented $\kappa = \sqrt{(1-\lambda^{2T})/(1-\lambda^2)}$ is correct, and the alternative $\sqrt{1+\lambda^2}$ is refuted; $\lambda = 0$ recovers $\kappa = 1$ (i.i.d. DP-SGD), as required. □

3.3 DP-SGD converges to a noise floor with an irreducible $d\Phi$ term

Assumption 1 (Smoothness and strong convexity)**.** The fine-tuning objective $F : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and μ-strongly convex on $\mathbb{R}^d$, with $0 < \mu \le L < \infty$. It has a unique minimizer $x^\star := \arg\min_x F(x)$, $F^\star := F(x^\star)$, and $\nabla F(x^\star) = 0$. Equivalently, the Bregman divergence $D_F(x, y) := F(x) - F(y) - \langle \nabla F(y), x - y\rangle$ obeys $\frac{\mu}{2}\|x-y\|^2 \le D_F(x,y) \le \frac{L}{2}\|x-y\|^2$ for all x, y.

Definition 4 (DP-SGD update)**.** Let N be the dataset size and $\mathcal{D} = \{\ell_1, \dots, \ell_N\}$ the per-example losses. Fix a clip norm $C > 0$, expected batch size B (Poisson rate $q = B/N$), and noise multiplier $\sigma_{\mathrm{DP}} \ge 0$. Writing $\mathrm{clip}_C(v) := v \cdot \min\{1, C/\|v\|_2\}$, the privatized gradient at iterate x is

$$g_{\mathrm{DP}}(x) \;=\; \frac{1}{B}\Big[\sum_{i\in\mathcal{B}} \mathrm{clip}_C(\nabla \ell_i(x)) \;+\; z\Big], \qquad z \sim \mathcal{N}\big(0, (\sigma_{\mathrm{DP}} C)^2 I_d\big),$$

where $\mathcal{B}$ is a Poisson(q) subsample and z is independent of $\mathcal{B}$ and of the data. DP-SGD runs $x^{k+1} = x^k - \gamma\, g_{\mathrm{DP}}(x^k)$ with constant step $\gamma > 0$. The per-coordinate DP-noise variance baked into g_{DP} is

$$\boxed{\Phi \;:=\; \Big(\frac{\sigma_{\mathrm{DP}}\, C}{B}\Big)^2} \qquad \Longrightarrow \qquad \mathbb{E}\|z/B\|^2 = d\,\Phi,$$

which matches the optimizer code exactly (`phi = (noise_multiplier*max_grad_norm/B_eff)**2`, with B_{eff} = expected batch × accumulation × world size).

Assumption 2 (Unbiasedness / small-clip regime)**.** On the trajectory the per-sample gradients satisfy $\|\nabla \ell_i(x)\| \le C$, so clipping acts as the identity, and Poisson subsampling is unbiased. Hence $\mathbb{E}[g_{\mathrm{DP}}(x) \mid x] = \nabla F(x)$. (When clipping bites, an additive bias $b(x)$ appears; this is treated in the clipping-bias remark and only inflates the floor, so the statement below is a lower bound on the true floor.)

Assumption 3 (Expected smoothness / ABC inequality)**.** There is a finite constant $A > 0$ (the *expected-smoothness* constant) and the optimum-variance $\sigma_\star^2 := \mathbb{E}\|g_{\mathrm{DP}}(x^\star)\|^2$ such that, for all x,

$$\mathbb{E}\big\|g_{\mathrm{DP}}(x) - g_{\mathrm{DP}}(x^\star)\big\|^2 \;\le\; 2A\,\big(F(x) - F^\star\big).$$

This is the standard unified-SGD / expected-smoothness hypothesis (Gower et al. 2019; Khaled & Richtárik 2020). The optimum variance splits, by independence of the Gaussian noise from the clipped signal, as

$$\sigma_\star^2 \;=\; \sigma_{\mathrm{sub}}^2 \;+\; d\,\Phi, \qquad \text{where } \sigma_{\mathrm{sub}}^2 := \mathbb{E}\|\bar{g}_{\mathrm{clip}}(x^\star)\|^2 \text{ is the subsampling variance.}$$

Theorem 5 (DP-SGD converges to a noise floor with an irreducible $d\Phi$ DP term)**.** *Under Assumptions 1–3, DP-SGD with any constant step*

$$0 < \gamma \le \frac{1}{2A}$$

satisfies, for every $k \ge 0$*,*

$$\mathbb{E}\|x^k - x^\star\|^2 \;\le\; (1-\gamma\mu)^k \, \|x^0 - x^\star\|^2 \;+\; \frac{2\gamma\,\sigma_\star^2}{\mu}.$$

Consequently $\limsup_{k\to\infty} \mathbb{E}\|x^k - x^\star\|^2 \le \dfrac{2\gamma\,\sigma_\star^2}{\mu}$*, and substituting the variance split of Assumption 3 isolates the differential-privacy contribution to the floor,*

$$\boxed{\limsup_{k\to\infty} \mathbb{E}\|x^k - x^\star\|^2 \;\le\; \underbrace{\frac{2\gamma\,\sigma_{\text{sub}}^2}{\mu}}_{\textit{subsampling}} \;+\; \underbrace{\frac{2\gamma\, d\,\Phi}{\mu}}_{=:\ \text{FLOOR}_{\text{DP}}}, \qquad \text{FLOOR}_{\text{DP}} = \frac{2\gamma\, d\,\sigma_{\text{DP}}^2 C^2}{\mu\, B^2}}.$$

The DP floor is linear in the step γ*, linear in the trainable dimension* d *(so DP-LoRA, which replaces* d *by the LoRA parameter count, lowers it), grows with the noise multiplier* σ_{DP} *(stricter* ε*) and clip* C*, shrinks like* $1/B^2$ *in the batch, and is irreducible: it is independent of* x *and does not vanish at* $x^\star$*, so no number of iterations can anneal it.*

Proof. We prove Theorem 5 in four steps: (i) a second-moment ("AC") bound on g_{DP}; (ii) a one-step contraction of $\mathbb{E}\|x^k - x^\star\|^2$; (iii) unrolling the recursion to the stated bound; (iv) substituting the DP-variance split to expose $\text{FLOOR}_{\text{DP}} \propto d\Phi$. Throughout, $\mathbb{E}_k[\,\cdot\,] := \mathbb{E}[\,\cdot \mid x^k]$ denotes expectation over the fresh Poisson subsample and Gaussian noise of step k, conditioned on x^k; $\mathbb{E}[\cdot]$ is the total expectation. All steps use only Assumptions 1–3.

Preliminary fact ($A \ge \mu$). We record one inequality used in Step 2 to control the contraction factor. Apply Assumption 3 to the special case where the stochastic gradient is the full (noise-free, full-batch) gradient $g = \nabla F$, for which $g(x^\star) = \nabla F(x^\star) = 0$; this gives $\|\nabla F(x)\|^2 \le 2A\,(F(x) - F^\star)$. On the other hand, μ-strong convexity implies the Polyak–Łojasiewicz inequality $\|\nabla F(x)\|^2 \ge 2\mu\,(F(x) - F^\star)$ for all x (apply (9) below at x with Cauchy–Schwarz and $\|\nabla F(x)\|\,\|x - x^\star\| \ge \langle \nabla F(x), x - x^\star\rangle \ge (F(x) - F^\star) + \frac{\mu}{2}\|x - x^\star\|^2 \ge 2\sqrt{\frac{\mu}{2}(F - F^\star)}\,\|x - x^\star\| \cdot \frac{1}{\sqrt{2}}$, the standard derivation). Chaining,

$$2\mu\,(F(x) - F^\star) \;\le\; \|\nabla F(x)\|^2 \;\le\; 2A\,(F(x) - F^\star) \qquad \Longrightarrow \qquad A \ge \mu,$$

taking any x with $F(x) > F^\star$. Hence the prescribed step obeys $\gamma \le \frac{1}{2A} \le \frac{1}{2\mu}$, so $\gamma\mu \le \frac{1}{2}$ and the contraction factor $r := 1 - \gamma\mu \in [\frac{1}{2}, 1)$ is in particular nonnegative and strictly less than 1. (If one prefers to take Assumption 3 as a black box, note that the unified-SGD literature always has $A \ge \mu$; either way $r \in [\frac{1}{2}, 1)$.)

Step 1 (Second-moment / AC bound). We first bound $\mathbb{E}\|g_{\text{DP}}(x)\|^2$ by the suboptimality gap. For any vectors a, b, the parallelogram-type inequality $\|a + b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ (from $\|a - b\|^2 \ge 0$) gives, with $a = g_{\text{DP}}(x) - g_{\text{DP}}(x^\star)$ and $b = g_{\text{DP}}(x^\star)$ (so $a + b = g_{\text{DP}}(x)$, and the inequality holds pointwise for each shared realization of subsample and noise),

$$\big\|g_{\text{DP}}(x)\big\|^2 \;\le\; 2\big\|g_{\text{DP}}(x) - g_{\text{DP}}(x^\star)\big\|^2 \;+\; 2\big\|g_{\text{DP}}(x^\star)\big\|^2.$$

Taking expectations and applying the expected-smoothness inequality (Assumption 3) to the first term and the definition $\sigma_\star^2 = \mathbb{E}\|g_{\mathrm{DP}}(x^\star)\|^2$ to the second,

$$\mathbb{E}\|g_{\mathrm{DP}}(x)\|^2 \;\le\; \underbrace{2(2A)}_{=:\mathcal{A}}\,(F(x)-F^\star) \;+\; \underbrace{2\sigma_\star^2}_{=:\mathcal{C}}, \qquad \mathcal{A}=4A,\quad \mathcal{C}=2\sigma_\star^2. \tag{7}$$

This is the standard $\mathbb{E}\|g\|^2 \le \mathcal{A}\,(F-F^\star)+\mathcal{C}$ inequality: the stochastic-gradient second moment is controlled by the *suboptimality gap* plus an additive constant $\mathcal{C}$ that does *not* vanish as $x \to x^\star$.

Step 2 (One-step contraction). Fix k and expand the squared distance after the update $x^{k+1} = x^k - \gamma\, g_{\mathrm{DP}}(x^k)$:

$$\|x^{k+1}-x^\star\|^2 = \|x^k-x^\star\|^2 - 2\gamma\langle g_{\mathrm{DP}}(x^k),\, x^k-x^\star\rangle + \gamma^2\|g_{\mathrm{DP}}(x^k)\|^2.$$

Apply $\mathbb{E}_k[\cdot]$. By unbiasedness (Assumption 2), $\mathbb{E}_k[g_{\mathrm{DP}}(x^k)] = \nabla F(x^k)$, and since $x^k - x^\star$ is x^k-measurable, $\mathbb{E}_k\langle g_{\mathrm{DP}}(x^k), x^k-x^\star\rangle = \langle \nabla F(x^k), x^k-x^\star\rangle$. Hence

$$\mathbb{E}_k\|x^{k+1}-x^\star\|^2 = \|x^k-x^\star\|^2 - 2\gamma\langle \nabla F(x^k), x^k-x^\star\rangle + \gamma^2\,\mathbb{E}_k\|g_{\mathrm{DP}}(x^k)\|^2. \tag{8}$$

Lower bound on the inner product. By μ-strong convexity, for all x,

$$F(x^\star) \ge F(x) + \langle \nabla F(x), x^\star - x\rangle + \tfrac{\mu}{2}\|x^\star - x\|^2,$$

which rearranges (negate the inner product, move $F(x)-F(x^\star)$ across) to the standard "three-point" inequality

$$\langle \nabla F(x), x-x^\star\rangle \;\ge\; (F(x)-F^\star) + \tfrac{\mu}{2}\|x-x^\star\|^2. \tag{9}$$

Upper bound on the second moment. By (7) with $x = x^k$,

$$\mathbb{E}_k\|g_{\mathrm{DP}}(x^k)\|^2 \le 4A(F(x^k)-F^\star) + 2\sigma_\star^2.$$

Substituting both bounds into (8),

$$\begin{aligned}\mathbb{E}_k\|x^{k+1}-x^\star\|^2 &\le \|x^k-x^\star\|^2 - 2\gamma\Big[(F(x^k)-F^\star) + \tfrac{\mu}{2}\|x^k-x^\star\|^2\Big] + \gamma^2\Big[4A(F(x^k)-F^\star)+2\sigma_\star^2\Big]\\ &= (1-\gamma\mu)\|x^k-x^\star\|^2 - \underbrace{(2\gamma-4A\gamma^2)}_{=2\gamma(1-2A\gamma)}(F(x^k)-F^\star) + 2\gamma^2\sigma_\star^2.\end{aligned}$$

The step-size cap $\gamma \le 1/(2A)$ gives $1-2A\gamma \ge 0$, and $F(x^k)-F^\star \ge 0$, so the middle term is ≤ 0 and may be dropped. This yields the contraction

$$\mathbb{E}_k\|x^{k+1}-x^\star\|^2 \;\le\; (1-\gamma\mu)\,\|x^k-x^\star\|^2 \;+\; 2\gamma^2\sigma_\star^2. \tag{10}$$

By the Preliminary fact the contraction factor satisfies $r := 1-\gamma\mu \in [\tfrac{1}{2}, 1)$. Taking total expectation $\mathbb{E}[\cdot]$ of (10) via the tower property,

$$\mathbb{E}\|x^{k+1}-x^\star\|^2 \;\le\; (1-\gamma\mu)\,\mathbb{E}\|x^k-x^\star\|^2 \;+\; 2\gamma^2\sigma_\star^2. \tag{11}$$

Step 3 (Unrolling the recursion). Write $r := 1-\gamma\mu$, $a_k := \mathbb{E}\|x^k-x^\star\|^2$, and $c := 2\gamma^2\sigma_\star^2 \ge 0$, so (11) reads $a_{k+1} \le r\,a_k + c$, with $r \in [\tfrac{1}{2}, 1)$ established above. We claim by induction that

$$a_k \;\le\; r^k a_0 + c\sum_{j=0}^{k-1} r^j, \qquad k \ge 0, \tag{12}$$

with the empty sum at $k=0$ equal to 0. The base case $a_0 \le a_0$ is trivial. Assuming (12) for k, and using $r \ge 0$ so that multiplication by r preserves the inequality,

$$a_{k+1} \le r\, a_k + c \le r\Big(r^k a_0 + c\sum_{j=0}^{k-1} r^j\Big) + c = r^{k+1}a_0 + c\sum_{j=0}^{k-1} r^{j+1} + c = r^{k+1}a_0 + c\sum_{j=0}^{k} r^j,$$

which is (12) for $k+1$. Since $0 \le r < 1$, every term $r^j \ge 0$, so the partial geometric sum is bounded by its (convergent) limit,

$$\sum_{j=0}^{k-1} r^j \;\le\; \sum_{j=0}^{\infty} r^j \;=\; \frac{1}{1-r} \;=\; \frac{1}{\gamma\mu}$$

(this monotone upper bound is exactly where $r \ge 0$, i.e. $\gamma \le 1/(2\mu)$, is needed; for $r < 0$ the partial sums would not be monotone in k). Plugging this and $r = 1-\gamma\mu$, $c = 2\gamma^2\sigma_\star^2$ into (12),

$$\mathbb{E}\|x^k - x^\star\|^2 \;\le\; (1-\gamma\mu)^k\,\|x^0 - x^\star\|^2 \;+\; \frac{2\gamma^2\sigma_\star^2}{\gamma\mu} \;=\; (1-\gamma\mu)^k\,\|x^0 - x^\star\|^2 \;+\; \frac{2\gamma\,\sigma_\star^2}{\mu}\,, \tag{13}$$

which is the first displayed bound of the theorem (using $a_0 = \|x^0 - x^\star\|^2$, a constant). Letting $k \to \infty$, the geometric term $(1-\gamma\mu)^k \to 0$ (as $0 \le r < 1$), leaving

$$\limsup_{k\to\infty} \mathbb{E}\|x^k - x^\star\|^2 \;\le\; \frac{2\gamma\,\sigma_\star^2}{\mu}\,. \tag{14}$$

Thus DP-SGD contracts geometrically at rate $(1-\gamma\mu)$ toward $x^\star$, but only into a ball of squared radius $R^2 = 2\gamma\sigma_\star^2/\mu$ — the *variance error floor*. The step-size cap $\gamma \le 1/(2A)$ is exactly what makes the expected-smoothness term in Step 2 nonpositive and hence absorbable, and (via $A \ge \mu$) simultaneously guarantees $r \ge 0$.

Step 4 (Isolating the $d\Phi$ DP term). It remains to split the floor by the structure of $\sigma_\star^2$. Evaluate g_{DP} at $x^\star$ and decompose it into its clipped-mean part and its Gaussian part:

$$g_{\mathrm{DP}}(x^\star) = \bar g_{\mathrm{clip}}(x^\star) + \frac{z}{B}, \qquad \bar g_{\mathrm{clip}}(x^\star) := \frac{1}{B}\sum_{i\in\mathcal{B}} \mathrm{clip}_C(\nabla \ell_i(x^\star)), \quad z \sim \mathcal{N}(0, (\sigma_{\mathrm{DP}}C)^2 I_d).$$

Because z is drawn independently of the subsample $\mathcal{B}$ and of the data, and $\mathbb{E}[z] = 0$, the cross term vanishes:

$$\mathbb{E}\langle \bar g_{\mathrm{clip}}(x^\star),\, z/B\rangle = \langle \mathbb{E}[\bar g_{\mathrm{clip}}(x^\star)],\, \mathbb{E}[z/B]\rangle = 0.$$

Hence the second moment is exactly additive:

$$\sigma_\star^2 = \mathbb{E}\|g_{\mathrm{DP}}(x^\star)\|^2 = \underbrace{\mathbb{E}\|\bar g_{\mathrm{clip}}(x^\star)\|^2}_{=:\,\sigma_{\mathrm{sub}}^2} + \underbrace{\mathbb{E}\|z/B\|^2}_{=\,d\,\Phi}.$$

For the Gaussian term, $\mathbb{E}\|z/B\|^2 = \frac{1}{B^2}\sum_{j=1}^d \mathbb{E}[z_j^2] = \frac{1}{B^2}\cdot d\,(\sigma_{\mathrm{DP}}C)^2 = d\,(\sigma_{\mathrm{DP}}C/B)^2 = d\,\Phi$, matching Definition 4. Substituting $\sigma_\star^2 = \sigma_{\mathrm{sub}}^2 + d\Phi$ into (14),

$$\limsup_{k\to\infty} \mathbb{E}\|x^k - x^\star\|^2 \;\le\; \frac{2\gamma}{\mu}(\sigma_{\mathrm{sub}}^2 + d\,\Phi) \;=\; \frac{2\gamma\,\sigma_{\mathrm{sub}}^2}{\mu} + \underbrace{\frac{2\gamma\,d\,\Phi}{\mu}}_{\mathrm{FLOOR}_{\mathrm{DP}}},$$

and expanding $\Phi = (\sigma_{\mathrm{DP}}C/B)^2$ gives $\mathrm{FLOOR}_{\mathrm{DP}} = 2\gamma\,d\,\sigma_{\mathrm{DP}}^2C^2/(\mu B^2)$, as claimed. The term $d\Phi$ is constant in x (it is the variance the Gaussian mechanism injects regardless of where the iterate is)

and strictly positive whenever $\sigma_{\mathrm{DP}} > 0$, so it persists even though $\nabla F(x^\star) = 0$: it is an *irreducible* contribution that no amount of iteration can remove. This is the convergence-theory shadow of the classical strongly-convex DP-ERM rate $\widetilde{O}(d/(N^2\varepsilon^2))$, whose d and $1/\varepsilon^2$ dependence enter entirely through this additive DP-noise variance, and which is matched by information-theoretic lower bounds (Bassily, Smith & Thakurta 2014). ■

Coupling to the privacy budget (corollary). To express the floor in ε, use the moments/RDP accountant: achieving (ε, δ)-DP over T steps at rate $q = B/N$ requires $\sigma_{\mathrm{DP}} \geq c\, q \sqrt{T \log(1/\delta)}/\varepsilon$ for an absolute constant c (Abadi et al. 2016; the RDP/PRV accountants of Mironov 2017 used in our implementation give the same scaling with a smaller constant). Substituting, and using $q = B/N$ so that $q^2/B^2 = 1/N^2$,

$$\Phi = \Big(\frac{\sigma_{\mathrm{DP}} C}{B}\Big)^2 = \frac{c^2 (B/N)^2 T \log(1/\delta)}{\varepsilon^2} \cdot \frac{C^2}{B^2} = \frac{c^2 C^2 T \log(1/\delta)}{N^2 \varepsilon^2},$$

so the explicit batch dependence *cancels* and

$$\mathrm{FLOOR}_{\mathrm{DP}}(\varepsilon) = \frac{2\, c^2\, \gamma\, d\, C^2\, T \log(1/\delta)}{\mu\, N^2\, \varepsilon^2} \;\propto\; \varepsilon^{-2},$$

recovering the monotone ε^{-2} utility degradation (and $\mathrm{FLOOR}_{\mathrm{DP}} \to 0$ as $\varepsilon \to \infty$, where it collapses to the clip-only floor $2\gamma\sigma_{\mathrm{sub}}^2/\mu$). The residual dependence on B enters only through T: at fixed epochs E one has $T = EN/B$, so $\mathrm{FLOOR}_{\mathrm{DP}} \propto 1/B$ — larger batches still lower the floor, consistent with the empirical ρ-saturation analysis.

Remark (clipping bias). Assumption 2 (clipping inactive) is what makes g_{DP} unbiased. When per-sample gradients exceed C, clipping introduces a bias $b(x) := \mathbb{E}[\bar{g}_{\mathrm{clip}}(x)] - \nabla F(x) \neq 0$, which adds a term $-2\gamma\langle b(x^k), x^k - x^\star\rangle$ to (8). Carrying it through with Young's inequality yields $\mathrm{FLOOR}_{\mathrm{true}} = \frac{2\gamma}{\mu}(\sigma_{\mathrm{sub}}^2 + d\Phi) + O(\|b\|^2/\mu^2)$; since this only adds a nonnegative term, the floor of Theorem 5 is a *lower bound* on the true DP-SGD floor. The bias vanishes as $C \to \infty$ or when the per-sample gradient noise is symmetric (Chen, Wu & Hong 2020).

Remark (non-convex analogue). Under L-smoothness alone (no strong convexity), the same AC template $\mathbb{E}\|g_{\mathrm{DP}}\|^2 \leq \mathcal{A}(F - F^\star) + \mathcal{C}$ with $\mathcal{C} = \mathcal{C}_{\mathrm{base}} + d\Phi$ gives, by the descent lemma, a stationarity bound $\frac{1}{K}\sum_{k<K} \mathbb{E}\|\nabla F(x^k)\|^2 \leq O\big(\Delta_F/(\gamma K)\big) + O\big(\gamma L\, \mathcal{C}\big)$ with $\Delta_F := F(x^0) - F^\star$, whose stationary floor is again $\propto \gamma\, d\, \sigma_{\mathrm{DP}}^2 C^2/B^2$. Thus the $d\Phi$ scaling survives the non-convex setting relevant to LLM fine-tuning; only the exponent on ε changes (general-convex $\sqrt{d}/\varepsilon$ vs. strongly-convex d/ε^2).

Connection to bias correction (why the floor is the right object). The *same* scalar Φ that appears here as the irreducible variance floor is the bias that DP-AdamBC subtracts from Adam's second moment: $\mathbb{E}[\hat{v}_t] = \hat{v}_t^{\mathrm{true}} + \Phi$. Bias correction removes Φ from the *preconditioner geometry* but does not touch $\sigma_\star^2$, and therefore by Theorem 5 *cannot* shrink $\mathrm{FLOOR}_{\mathrm{DP}}$. Reducing the floor requires smaller σ_{DP} (looser ε), larger B, or smaller d (LoRA) — exactly the levers our experiments confirm, and the reason the noise share $\rho = \Phi/\hat{v}$ saturating at 1 leaves bias correction with no utility to recover. □

3.4 First-moment denoising carries the learning at $\rho \approx 1$

Theorem 6 (First-moment denoising at the noise floor $\rho \approx 1$)**.** *Consider DP-Adam (the optimizer of* `src/dp_optim/dp_adaptive.py`*) applied to a single fixed coordinate during a window of fine-tuning. After Opacus' clip–noise–average step, the gradient fed to the Adam update is*

$$g_t \;=\; \bar{g}_t + \xi_t, \qquad \xi_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \Phi), \qquad \Phi \;=\; \Big(\frac{\sigma_{\mathrm{DP}}\, C}{B_{\mathrm{eff}}}\Big)^2, \quad B_{\mathrm{eff}} = \textit{(expected batch)} \times \textit{(accum. steps)},$$

where $\bar{g}_t = \frac{1}{B}\sum_{i\in\mathcal{B}} \mathrm{clip}_C(\nabla \ell_i(x_t))$ *is the (conditionally deterministic, given* x_t*) clipped mean gradient and* ξ_t *is the per-coordinate Gaussian DP noise, independent of* $\bar{g}_t$ *and across* t*. The optimizer maintains, with decay rates* $\beta_1, \beta_2 \in [0, 1)$*,*

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \quad v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2, \quad \hat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t},$$

$m_0 = v_0 = 0$*, and emits the update* $\Delta_t = -\,\mathrm{lr}\,\hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon)$*.*

Assume the noise-floor (saturated) regime

$$\textbf{\textit{(A1)}} \quad \rho_t := \frac{\Phi}{\hat{v}_t^{\mathrm{true}} + \Phi} \approx 1 \iff \hat{v}_t^{\mathrm{true}} \ll \Phi, \qquad \hat{v}_t^{\mathrm{true}} := \frac{(1-\beta_2)\sum_{k=0}^{t-1} \beta_2^k\, \mathbb{E}[\bar{g}_{t-k}^2]}{1-\beta_2^t},$$

i.e. the squared true (clipped, batch-averaged) gradient is far below the DP-noise variance, and ***(A2)*** *the drift of* $\bar{g}_t$ *is slow on the EMA timescale* $(1-\beta_1)^{-1}$*, so* $\mathbb{E}[\bar{g}_t] = \mu$ *may be treated as constant over a window (locally stationary first moment). Then:*

- ***(i)*** (Unbiasedness) $\hat{m}_t$ *is an unbiased estimator of the true gradient in the sense* $\mathbb{E}[\hat{m}_t] = \frac{(1-\beta_1)\sum_{k=0}^{t-1}\beta_1^k\,\mathbb{E}[\bar{g}_{t-k}]}{1-\beta_1^t}$*, which equals* μ *exactly under (A2); for* $\beta_1 = 0$*,* $\mathbb{E}[\hat{m}_t] = \mathbb{E}[\bar{g}_t]$ *for every* t*.*

- ***(ii)*** (Steady-state variance) *The DP-noise variance of* $\hat{m}_t$ *has the exact finite-t value*

  $$\mathrm{Var}_\xi(\hat{m}_t) = \frac{1-\beta_1}{1+\beta_1} \cdot \frac{1+\beta_1^t}{1-\beta_1^t}\,\Phi,$$

  which decreases *monotonically from* Φ *(at* $t = 1$*, no averaging) to*

  $$\boxed{\lim_{t\to\infty} \mathrm{Var}_\xi(\hat{m}_t) \;=\; \frac{1-\beta_1}{1+\beta_1}\,\Phi}$$

  and is strictly less than *the per-step variance* Φ *for every* $\beta_1 \in (0, 1)$ *and every* $t \geq 2$*.*

- ***(iii)*** (Adaptivity off) *Under (A1),* $\hat{v}_t = \hat{v}_t^{\mathrm{true}} + \Phi \approx \Phi$ *is approximately constant across coordinates and iterations (it carries no per-coordinate signal), so the preconditioner* $1/(\sqrt{\hat{v}_t}+\epsilon)$ *degenerates to the scalar* $1/(\sqrt{\Phi} + \epsilon)$*.*

- ***(iv)*** (Collapse to momentum-SGD; signal lives in $\hat{m}_t$) *Consequently the DP-Adam update obeys*

  $$\Delta_t = -\,\mathrm{lr}\,\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \;\approx\; -\,\frac{\mathrm{lr}}{\sqrt{\Phi} + \epsilon}\,\hat{m}_t \;=:\; -\,\mathrm{lr}_{\mathrm{eff}}\,\hat{m}_t,$$

  a momentum-SGD step with fixed scalar step size $\mathrm{lr}_{\mathrm{eff}}$*. All learning signal enters* only *through* $\hat{m}_t$ *(a* $\frac{1-\beta_1}{1+\beta_1}$*-variance-reduced prefix average of the noisy gradients);* $\hat{v}_t$ *contributes only the constant* $\mathrm{lr}_{\mathrm{eff}}$*.*

Proof. **Notation and standing facts.** Fix one coordinate (all quantities are scalars; the per-coordinate DP noise is independent across coordinates, so the argument is coordinatewise without loss of generality). Write $g_t = \bar{g}_t + \xi_t$ with $\xi_t \sim \mathcal{N}(0, \Phi)$ i.i.d. and independent of $\{\bar{g}_s\}_s$. By the definition of Φ in `dp_adaptive.py` (the `phi` property, $\Phi = (\sigma_{\mathrm{DP}}\, C/B_{\mathrm{eff}})^2$), this is exactly the per-coordinate

variance baked into `p.grad` after Opacus' `clip_and_accumulate`→`add_noise`→`scale_grad`. Unrolling the m-EMA with $m_0 = 0$ gives the explicit linear (FIR / weighted prefix-sum) representation

$$m_t = (1-\beta_1)\sum_{k=0}^{t-1} \beta_1^k \, g_{t-k}, \qquad \text{hence} \qquad \hat{m}_t = \frac{m_t}{1-\beta_1^t} = \frac{(1-\beta_1)}{1-\beta_1^t}\sum_{k=0}^{t-1} \beta_1^k \, g_{t-k}. \tag{15}$$

This is verified by induction on t. Base case: $m_1 = \beta_1 m_0 + (1-\beta_1)g_1 = (1-\beta_1)g_1$, which is (15) at $t = 1$. Inductive step: if (15) holds at $t-1$, then

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t = (1-\beta_1)\Big[g_t + \sum_{k=1}^{t-1} \beta_1^k \, g_{t-k}\Big] = (1-\beta_1)\sum_{k=0}^{t-1} \beta_1^k \, g_{t-k},$$

where the second equality re-indexed $\beta_1 \cdot \beta_1^{k-1} g_{(t-1)-(k-1)} = \beta_1^k g_{t-k}$. This is the claim at t. The normalizing constant uses the finite geometric sum

$$(1-\beta_1)\sum_{k=0}^{t-1} \beta_1^k = 1 - \beta_1^t \qquad \text{(valid for all } \beta_1 \in [0,1)\text{, including } \beta_1 = 0 \text{ with } 0^0 := 1), \tag{16}$$

so the EMA weights $w_k^{(t)} := \dfrac{(1-\beta_1)\beta_1^k}{1-\beta_1^t}$ $(k = 0, \ldots, t-1)$ are nonnegative and sum to 1: $\hat{m}_t = \sum_{k=0}^{t-1} w_k^{(t)} g_{t-k}$ is a genuine weighted average. (I verified (15) against the literal recursion `m.mul_(b1).add_(g,alpha=1-b1)`; `mhat=m/(1-b1**t)` of `_adam_update` numerically for $\beta_1 \in \{0, 0.5, 0.9\}$.)

(i) Unbiasedness. Linearity of expectation applied to $\hat{m}_t = \sum_k w_k^{(t)} g_{t-k}$, together with $\mathbb{E}[g_{t-k}] = \mathbb{E}[\bar{g}_{t-k}] + \mathbb{E}[\xi_{t-k}] = \mathbb{E}[\bar{g}_{t-k}]$ (the DP noise is zero-mean and independent of the signal), gives

$$\mathbb{E}[\hat{m}_t] = \sum_{k=0}^{t-1} w_k^{(t)} \, \mathbb{E}[\bar{g}_{t-k}] = \frac{(1-\beta_1)\sum_{k=0}^{t-1} \beta_1^k \, \mathbb{E}[\bar{g}_{t-k}]}{1-\beta_1^t}, \tag{17}$$

which is the displayed estimator mean. Under the local-stationarity assumption (A2), $\mathbb{E}[\bar{g}_{t-k}] = \mu$ for all k in the window, so $\mathbb{E}[\hat{m}_t] = \mu \sum_k w_k^{(t)} = \mu$ by $\sum_k w_k^{(t)} = 1$; thus $\hat{m}_t$ is unbiased for the true (clipped, batch-mean) gradient $\mu = \mathbb{E}[\bar{g}_t]$. For $\beta_1 = 0$ the sum has the single term $k = 0$ with $w_0^{(t)} = 1$ (using (16)), giving $\hat{m}_t = g_t$ and $\mathbb{E}[\hat{m}_t] = \mathbb{E}[\bar{g}_t]$ for every t with no stationarity assumption needed. (The bias-correction divisor $1 - \beta_1^t$ is exactly what removes the cold-start under-counting; without it $\mathbb{E}[m_t] = (1-\beta_1^t)\mu \to \mu$ only asymptotically.)

In the floor regime (A1) the "signal" $\mu = \mathbb{E}[\bar{g}_t]$ is itself small: by Jensen, $\mu^2 = (\mathbb{E}[\bar{g}_t])^2 \le \mathbb{E}[\bar{g}_t^2]$, and under (A1)/(A2) (stationary $\mathbb{E}[\bar{g}_t^2]$) one has $\mathbb{E}[\bar{g}_t^2] = \hat{v}_t^{\text{true}} \ll \Phi$; hence $\mu^2 \ll \Phi$. Part (i) asserts only that $\hat{m}_t$ tracks this small true gradient *without distortion*: the first moment is the unbiased estimator and all systematic content of the step originates in μ.

(ii) Steady-state variance of $\hat{m}_t$. We isolate the DP-noise contribution to $\mathrm{Var}(\hat{m}_t)$ — the quantity the theorem concerns. Since ξ is independent of the signal $\{\bar{g}_s\}$ and i.i.d. across t, the law of total variance / bilinearity of covariance gives the exact decomposition

$$\mathrm{Var}(\hat{m}_t) = \mathrm{Var}\Big(\sum_k w_k^{(t)} \bar{g}_{t-k}\Big) + \underbrace{\sum_{k=0}^{t-1} \big(w_k^{(t)}\big)^2 \, \Phi}_{=:\, \mathrm{Var}_\xi(\hat{m}_t)}, \tag{18}$$

where the cross term $\mathrm{Cov}(\sum_k w_k^{(t)} \bar{g}_{t-k}, \sum_k w_k^{(t)} \xi_{t-k}) = 0$ by signal–noise independence, and the noise term used $\mathrm{Cov}(\xi_{t-k}, \xi_{t-j}) = \Phi\, \delta_{kj}$. The first (signal) term is $O(\hat{v}_t^{\mathrm{true}}) \ll \Phi$ under (A1); I confirmed (18) and this ordering by Monte Carlo with a random signal trajectory of variance $\sim v_{\mathrm{true}}$, obtaining signal/noise $\approx 10^{-3}$, matching the empirical $v_{\mathrm{true}}/\Phi \le 1.4 \times 10^{-3}$ of `findings.md`. We therefore evaluate the DP-noise variance:

$$\mathrm{Var}_\xi(\hat{m}_t) = \frac{(1-\beta_1)^2}{(1-\beta_1^t)^2} \Big(\sum_{k=0}^{t-1} \beta_1^{2k} \Big) \Phi = \frac{(1-\beta_1)^2}{(1-\beta_1^t)^2} \cdot \frac{1-\beta_1^{2t}}{1-\beta_1^2}\, \Phi, \tag{19}$$

using $\sum_{k=0}^{t-1} \beta_1^{2k} = (1-\beta_1^{2t})/(1-\beta_1^2)$ for $\beta_1 \in (0,1)$ (for $\beta_1 = 0$ the sum is 1, handled at the end of this paragraph). Factoring $1-\beta_1^{2t} = (1-\beta_1^t)(1+\beta_1^t)$ and $1-\beta_1^2 = (1-\beta_1)(1+\beta_1)$ simplifies (19) to

$$\mathrm{Var}_\xi(\hat{m}_t) = \frac{(1-\beta_1)^2}{(1-\beta_1^t)^2} \cdot \frac{(1-\beta_1^t)(1+\beta_1^t)}{(1-\beta_1)(1+\beta_1)}\, \Phi = \frac{1-\beta_1}{1+\beta_1} \cdot \frac{1+\beta_1^t}{1-\beta_1^t}\, \Phi. \tag{20}$$

This is the exact finite-t value. *(This corrects the schematic placeholder "$\frac{1+\beta_1^{t+?}}{\cdot}$" appearing in the supplied statement: the exact prefactor is $\frac{1+\beta_1^t}{1-\beta_1^t}$.)* I verified (20) against the direct weight-sum $\sum_k (w_k^{(t)})^2 \Phi$ for $\beta_1 \in \{0, 0.5, 0.9, 0.99\}$, $t \in \{1, 2, 5, 20, 200\}$: exact agreement. Taking $t \to \infty$ with $\beta_1 \in (0,1)$, $\beta_1^t \to 0$, so

$$\lim_{t\to\infty} \mathrm{Var}_\xi(\hat{m}_t) = \frac{1-\beta_1}{1+\beta_1}\, \Phi, \tag{21}$$

the boxed steady-state variance. (For $\beta_1 = 0$, (19) gives $\mathrm{Var}_\xi(\hat{m}_t) = \Phi$ for all t — consistent with the limit $\frac{1-0}{1+0}\Phi = \Phi$ and with $\hat{m}_t = g_t$: no averaging.)

Strict reduction. For $\beta_1 \in (0,1)$, $\frac{1-\beta_1}{1+\beta_1} < 1$ (numerator strictly below denominator since $-\beta_1 < \beta_1$), so the steady-state noise variance $\frac{1-\beta_1}{1+\beta_1}\Phi < \Phi$ strictly. At the standard Adam value $\beta_1 = 0.9$ the reduction factor is $0.1/1.9 \approx 0.0526$: the first moment carries $\approx 5.3\%$ of the per-step noise variance — a $\approx 19\times$ variance reduction, equivalently an effective averaging window of $\frac{1+\beta_1}{1-\beta_1} = 19$ steps. (Verified numerically: 0.05263.)

Monotonicity (direction corrected). Define $\phi(t) := \frac{1+\beta_1^t}{1-\beta_1^t}$. With $u := \beta_1^t \in (0,1)$ *decreasing* in t, $\phi = \frac{1+u}{1-u}$ has $\frac{d\phi}{du} = \frac{(1-u)+(1+u)}{(1-u)^2} = \frac{2}{(1-u)^2} > 0$, so ϕ is increasing in u and hence *decreasing* in t. Thus $\mathrm{Var}_\xi(\hat{m}_t)$ *decreases* monotonically from its $t = 1$ value $\frac{1-\beta_1}{1+\beta_1} \cdot \frac{1+\beta_1}{1-\beta_1}\Phi = \Phi$ down to the limit (21). *(This corrects the supplied statement's "increases monotonically": at $t = 1$ there is no averaging yet, $\hat{m}_1 = g_1$, variance Φ; averaging accrues — lowering the variance — only as the window fills.)* Hence $\mathrm{Var}_\xi(\hat{m}_t) \in \big[\frac{1-\beta_1}{1+\beta_1}\Phi,\ \Phi\big]$, equal to Φ only at $t = 1$, and strictly below Φ for every $t \ge 2$. (Verified: the sequence is strictly decreasing for $\beta_1 \in \{0.5, 0.9\}$.)

Remark (steady-state autocovariance, for completeness). The stationary process $\hat{m}_\infty^{(t)} = (1-\beta_1)\sum_{k\ge 0} \beta_1^k \xi_{t-k}$ (the $t \to \infty$ AR(1) EMA of the noise) has lag-ℓ autocovariance $\mathrm{Cov}(\hat{m}_\infty^{(t)}, \hat{m}_\infty^{(t-\ell)}) = (1-\beta_1)^2 \sum_{k\ge 0} \beta_1^k \beta_1^{k+\ell}\, \Phi = \frac{(1-\beta_1)^2 \beta_1^\ell}{1-\beta_1^2}\Phi = \frac{1-\beta_1}{1+\beta_1}\beta_1^\ell \Phi$; at $\ell = 0$ this recovers (21). (I simulated the AR(1)-EMA of white noise and recovered both the boxed marginal variance and the lag-ℓ law.) Thus the first moment is a low-pass-filtered (AR(1)) version of the white DP noise, whose marginal variance is exactly the boxed quantity — a clean instance of the $\frac{1-\beta}{1+\beta}$ EMA variance-reduction identity.

(iii) Adaptivity off ($\hat{v}_t \approx \Phi$, constant). By the second-moment inflation identity (Lemma **??** of `theory.tex`): conditioning on $\bar{g}_t$, $\mathbb{E}[g_t^2 \mid \bar{g}_t] = \bar{g}_t^2 + 2\bar{g}_t \mathbb{E}[\xi_t] + \mathbb{E}[\xi_t^2] = \bar{g}_t^2 + \Phi$ (cross term 0

since $\mathbb{E}[\xi_t \mid \bar{g}_t] = 0$); propagating through the v-EMA and dividing by $1-\beta_2^t$ (whereby the factor $(1-\beta_2)\sum_k \beta_2^k = 1-\beta_2^t$ multiplying the constant Φ cancels), one obtains

$$\mathbb{E}[\hat{v}_t] = \hat{v}_t^{\text{true}} + \Phi. \tag{22}$$

Under (A1), $\hat{v}_t^{\text{true}} \ll \Phi$, so $\mathbb{E}[\hat{v}_t] = (1 + \hat{v}_t^{\text{true}}/\Phi)\Phi = (1+o(1))\Phi \approx \Phi$. The coordinate-to-coordinate and step-to-step variation of $\hat{v}_t$ is governed entirely by $\hat{v}_t^{\text{true}}$ (the only x-dependent term), which is $O(\hat{v}^{\text{true}}) = o(\Phi)$; the dominant term Φ is a *global constant* (it depends only on $\sigma_{\text{DP}}, C, B_{\text{eff}}$, identical across coordinates). Hence $\hat{v}_t$ carries no per-coordinate signal and $\sqrt{\hat{v}_t}+\epsilon \approx \sqrt{\Phi}+\epsilon$ uniformly. This is precisely the empirical observation that $\rho = \Phi/\hat{v}$ saturates at 1.00–1.07 across $\varepsilon \in \{1,3,8\}$ and $B \in [16, 2048]$ (`findings.md`: "$v_{\text{true}} \ll \Phi$ everywhere"; measured $\hat{v} - \Phi = -3 \times 10^{-11}$, i.e. $v_{\text{true}}/\Phi \le 1.4 \times 10^{-3}$).

(iv) Collapse to momentum-SGD; signal lives in $\hat{m}_t$. The optimizer's update line is, verbatim, `p.addcdiv_(mhat, vhat.sqrt().add_(eps), value=-lr)`, i.e. $\Delta_t = -\,\text{lr}\,\hat{m}_t/(\sqrt{\hat{v}_t}+\epsilon)$. We make the substitution of the constant denominator from (iii) quantitative. Set $a := \hat{v}_t^{\text{true}} \ge 0$ and $b := \Phi > 0$, so $\hat{v}_t = a + b$ (in mean; the fluctuation argument is identical coordinatewise). By the mean-value form,

$$\left|\tfrac{1}{\sqrt{a+b}+\epsilon} - \tfrac{1}{\sqrt{b}+\epsilon}\right| = \frac{\sqrt{a+b}-\sqrt{b}}{(\sqrt{a+b}+\epsilon)(\sqrt{b}+\epsilon)} \le \frac{a/(2\sqrt{b})}{(\sqrt{b}+\epsilon)^2} = \frac{1}{2(\sqrt{\Phi}+\epsilon)^2}\cdot\frac{\hat{v}^{\text{true}}}{\sqrt{\Phi}},$$

using $\sqrt{a+b}-\sqrt{b} = a/(\sqrt{a+b}+\sqrt{b}) \le a/(2\sqrt{b})$ and $\sqrt{a+b}+\epsilon \ge \sqrt{b}+\epsilon$. Multiplying by $(\sqrt{\Phi}+\epsilon)$ shows the *relative* deviation of the preconditioner from the constant $1/(\sqrt{\Phi}+\epsilon)$ is at most $\frac{1}{2}\hat{v}^{\text{true}}/\Phi = \frac{1}{2}r$, where $r := \hat{v}^{\text{true}}/\Phi = \rho^{-1} - 1 = (1-\rho)/\rho \to 0$ as $\rho \to 1^-$. (Both the mean-value inequality and the relative bound $\le r/2$ were verified by 10^5 random draws; no violations.) Therefore

$$\Delta_t = -\,\text{lr}\,\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} = -\,\frac{\text{lr}}{\sqrt{\Phi}+\epsilon}\,(1+O(r))\,\hat{m}_t = -\,\text{lr}_{\text{eff}}\,\hat{m}_t(1+O(r)), \qquad \text{lr}_{\text{eff}} := \frac{\text{lr}}{\sqrt{\Phi}+\epsilon}, \tag{23}$$

with lr_{eff} a *fixed scalar* and relative error $\le r/2 \ll 1$. Hence the DP-Adam update is, to leading order, a momentum-SGD step $\Delta_t = -\text{lr}_{\text{eff}}\hat{m}_t$ with a fixed scalar step size; adaptivity is switched off, the geometric per-coordinate rescaling $1/\sqrt{\hat{v}_t}$ having degenerated to the global constant $1/\sqrt{\Phi}$.

Finally, all learning signal enters *only* through $\hat{m}_t$: by (i)–(ii), $\hat{m}_t$ is the unbiased, $\frac{1-\beta_1}{1+\beta_1}$-variance-reduced prefix average of the noisy gradients (a low-pass of the white DP noise that drifts toward the small true gradient μ over $\approx\frac{1+\beta_1}{1-\beta_1}$ steps), while $\hat{v}_t \approx \Phi$ contributes only the constant scale lr_{eff} and no directional information. This proves (iv): at $\rho \approx 1$ the update tracks a momentum-SGD step whose signal comes entirely from the first moment $\hat{m}_t$.

Consequences (consistency with the empirical record). (a) Because the only surviving role of $\hat{v}_t$ is to set the scalar lr_{eff}, DP-AdamBC's Φ-subtraction ($\hat{v}^{\text{BC}} = \max(\hat{v} - \Phi, \xi)$) and the floor-only control ($\max(\hat{v}, \xi)$, no Φ term) both reduce to $\Delta_t = -(\text{lr}/\sqrt{\xi})\hat{m}_t$ once the clamp binds (which it does at $\rho \approx 1$, where $\hat{v} - \Phi \approx 0 < \xi$): they differ only in the scalar $1/\sqrt{\xi}$, i.e. an effective learning rate. This is exactly the observed collapse (`findings.md` cycle 7: all DP-AdamBC floors fully clamped, $\hat{v} \equiv \xi$ constant, BLEU monotone in ξ; `dp-adambc`$(10^{-6}) \approx$`dp-adam-xi`(10^{-6})). (b) Since the signal is confined to $\hat{m}_t$, methods that reshape $\hat{v}_t$ (BC) or the per-step *direction* (Muon orthogonalization, $+8 \times 10^{-5}$ cosine) cannot recover below-floor signal; only the first-moment path (momentum/prefix-sum, or reducing Φ via batch/ε) is a live lever — the premise that motivated DP-CorrMom.

Appendix: the privacy sensitivity κ of the DP-CorrMom injection (independently re-derived). For completeness, and because part (iv)(b) invokes the first-moment lever realized by DP-CorrMom, we re-derive the matrix-mechanism sensitivity used in `corr_sensitivity`. The injection $n_t = z_t - \lambda z_{t-1}$ over a horizon of T steps is $n = Lz$ with $L \in \mathbb{R}^{T\times T}$ lower-bidiagonal, $L_{tt} = 1$, $L_{t,t-1} = -\lambda$. The released object is the gradient prefix-sum (workload A = lower-triangular all-ones); the error accumulated in the prefix-sum is $An = ALz$, and the matrix mechanism factors $A = BC$ with strategy $C = L^{-1}$ (so that $AL = B$ acts on the *decorrelated* white z, leaving the released error proportional to Bz with z i.i.d.). The per-step privacy cost (the ℓ_2 sensitivity of one record's contribution under the strategy) is the maximum column ℓ_2-norm of $C = L^{-1}$. Now L^{-1} is the lower-triangular Toeplitz matrix with first column $(1, \lambda, \lambda^2, \ldots, \lambda^{T-1})^\top$ (directly: $L^{-1} = \sum_{j\geq 0}(\lambda N)^j$ where N is the unit sub-diagonal shift, $N^T = 0$, giving entries $(L^{-1})_{ij} = \lambda^{i-j}\mathbf{1}\{i \geq j\}$). Column t (0-indexed) has squared norm $\sum_{k=0}^{T-1-t} \lambda^{2k}$, maximized at $t = 0$:

$$\kappa = \max_t \|L^{-1}e_t\|_2 = \Big(\sum_{k=0}^{T-1} \lambda^{2k}\Big)^{1/2} = \sqrt{\frac{1-\lambda^{2T}}{1-\lambda^2}}\,,$$

which is exactly `corr_sensitivity(lam,T)` and *not* $\sqrt{1+\lambda^2}$ (the latter is $\|Le_0\|_2$, the column norm of L itself — the column norm of the *wrong* matrix, which under-noises and breaks privacy). A Mahalanobis cross-check confirms this: with $\hat{g} \sim \mathcal{N}(g, \nu^2 LL^\top)$, a unit change in coordinate t incurs $\Delta = (G^2/\nu^2)(LL^\top)^{-1}_{tt}$ and $(LL^\top)^{-1}_{tt} = \|L^{-1}e_t\|_2^2$, recovering the same κ. I verified by explicit numerical inversion of L for $\lambda \in \{0, 0.5, 0.9, 0.95\}$, $T \in \{5, 50, 500\}$ that $\max_t \|L^{-1}e_t\|_2$ equals $\sqrt{(1-\lambda^{2T})/(1-\lambda^2)}$ exactly (e.g. $\lambda = 0.9$: $\kappa \to 2.294$, vs. the wrong $\sqrt{1+\lambda^2} = 1.345$), and that $(LL^\top)^{-1}_{tt} = \|L^{-1}e_t\|_2^2$. Thus the κ used throughout is correct. □ □

3.5 DP-CorrMom privacy: the verified κ sensitivity

Theorem 7 (DP-CorrMom privacy via the bidiagonal matrix mechanism)**.** *Fix a number of steps $T \in \mathbb{N}$, a per-sample clipping norm $C > 0$, and a correlation coefficient $\lambda \in [0, 1)$. Consider the DP-CorrMom release in which, at each step $t \in \{1, \ldots, T\}$, after per-sample ℓ_2 clipping at norm C one outputs the noisy summed gradient*

$$\hat{g}_t \;=\; s_t + n_t, \qquad s_t := \sum_{i\in\mathcal{B}_t} \mathrm{clip}_C(\nabla \ell_i(x_t)) \in \mathbb{R}^d, \tag{24}$$

with the anti-correlated *Gaussian injection*

$$n_t \;=\; z_t - \lambda\, z_{t-1}, \qquad z_0, z_1, \ldots, z_T \overset{\text{iid}}{\sim} \mathcal{N}(0, \nu^2 I_d), \quad z_0 := 0. \tag{25}$$

Assume single participation*: the index sets $\mathcal{B}_1, \ldots, \mathcal{B}_T$ are disjoint (every example contributes to at most one step; e.g. a single pass with a fixed disjoint partition). Let $A \in \mathbb{R}^{T\times T}$ be the lower-triangular all-ones prefix-sum "workload" matrix ($A_{t,k} = \mathbf{1}[k \leq t]$). Then the following hold.*

(i) (Single Gaussian mechanism.) *Per coordinate, the entire T-step release is a* single fixed-covariance *Gaussian mechanism. Writing the injection as $n = Lz$ with $L \in \mathbb{R}^{T\times T}$ lower bidiagonal ($L_{tt} = 1$, $L_{t,t-1} = -\lambda$), the prefix-sum workload A factorises as $A = B\,C_{\text{strat}}$ with* reconstruction $B := AL$ *and* strategy $C_{\text{strat}} := L^{-1}$ *(indeed $(AL)\,L^{-1} = A$), and the released prefix-sum stream is $A\hat{g} = As + \nu\,(AL)\,\zeta$ with $\zeta \sim \mathcal{N}(0, I)$ per coordinate. Its privacy is governed by a* single *effective noise multiplier*

$$\sigma_{\text{eff}} \;=\; \frac{\nu}{C\,\kappa}, \qquad \kappa \;:=\; \max_{1\leq t\leq T} \big\|(L^{-1})_{:,t}\big\|_2 \;=\; \sqrt{\frac{1-\lambda^{2T}}{1-\lambda^2}}\,, \tag{26}$$

in the precise sense that the per-coordinate (α, ε) Rényi divergence between the output laws on two single-participation neighbours is bounded by that of one Gaussian mechanism of multiplier σ_{eff}, with equality for the worst-case neighbour: $D_\alpha(M(D) \,\|\, M(D')) \le \alpha/(2\sigma_{\text{eff}}^2)$ for all $\alpha > 1$.

(ii) (Calibration.) *Consequently, to obtain the same Rényi/RDP guarantee (hence the same (ε, δ) after conversion) as an i.i.d. DP-SGD release with target noise multiplier σ, it suffices to set*

$$\boxed{\nu \;=\; \sigma\, C\, \kappa \;=\; \sigma\, C \sqrt{\tfrac{1-\lambda^{2T}}{1-\lambda^2}}}. \tag{27}$$

(iii) (Reduction at $\lambda = 0$.) *For $\lambda = 0$ one has $L = I$, $\kappa = 1$, $n_t = z_t$, and (27) gives $\nu = \sigma C$: the mechanism is exactly single-participation DP-SGD with multiplier σ.*

(iv) (The naive sensitivity is wrong.) *The factor is $\kappa = \|(L^{-1})_{:,t^\star}\|_2$,* not *the column norm $\sqrt{1+\lambda^2} = \|L_{:,t}\|_2$ of the injection matrix L. Using $\sqrt{1+\lambda^2}$ under-scales the noise (e.g. 1.35 vs. the correct 2.29 at $\lambda = 0.9$) and* violates *the intended DP guarantee.*

Finally, the correlation in (25) couples the per-step outputs, so Poisson-subsampling privacy amplification no longer applies; the guarantee above is for the unamplified single-participation accountant.

Proof. Throughout we exploit that both the injection (25) and the prefix-sum aggregation act *coordinate-wise and identically* on the d coordinates of $\mathbb{R}^d$, with independent noise across coordinates. Hence we may fix one coordinate, work with scalar streams in $\mathbb{R}^T$, and recombine at the end via the Kronecker structure of the joint covariance (Step 2). The neighbour relation is the standard add/remove-one-record relation; under single participation a neighbouring dataset differs in the contents of *exactly one* batch $\mathcal{B}_{t^\star}$, so the clipped-sum stream $s = (s_1, \dots, s_T)$ changes only in the single block $s_{t^\star}$.

Step 0: Notation and the factorisation $A = B\,L^{-1}$ with $B = AL$. Stack the scalar (fixed-coordinate) injections as $z = (z_1, \dots, z_T)^\top$ (with $z_0 = 0$) and $n = (n_1, \dots, n_T)^\top$. Equation (25) reads $n = Lz$, where $L \in \mathbb{R}^{T\times T}$ is the lower-bidiagonal Toeplitz matrix

$$L = \begin{pmatrix} 1 & & & & \\ -\lambda & 1 & & & \\ & -\lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & -\lambda & 1 \end{pmatrix}, \qquad z \sim \mathcal{N}(0, \nu^2 I_T).$$

Indeed $(Lz)_t = L_{t,t} z_t + L_{t,t-1} z_{t-1} = z_t - \lambda z_{t-1} = n_t$. L is lower triangular with unit diagonal, hence invertible with $\det L = 1$. Its inverse is the lower-triangular Toeplitz matrix of geometric weights,

$$(L^{-1})_{t,k} = \begin{cases} \lambda^{\,t-k}, & k \le t, \\ 0, & k > t, \end{cases} \tag{28}$$

which one verifies directly: for $k < t$, $(LL^{-1})_{t,k} = \sum_j L_{t,j}(L^{-1})_{j,k} = (L^{-1})_{t,k} - \lambda(L^{-1})_{t-1,k} = \lambda^{t-k} - \lambda \cdot \lambda^{t-1-k} = 0$; for $k = t$ it equals $(L^{-1})_{t,t} = 1$; for $k > t$ both terms vanish; and the boundary row $t = 1$ gives $(L^{-1})_{1,1} = 1$. (Numerically, materialising L and computing L^{-1} reproduces (28) to machine precision, max-error $\sim 10^{-16}$.)

The DP-CorrMom mechanism outputs, per coordinate, the noisy summed-gradient stream $\hat{g} = s + n = s + Lz$. The quantity that drives learning (and whose privacy we audit, since prefix

sums are post-processing of $\hat{g}$) is the running prefix sum $A\hat{g}$, A lower-triangular all-ones. Writing $z = \nu\zeta$ with $\zeta \sim \mathcal{N}(0, I_T)$,

$$A\hat{g} \;=\; As \;+\; ALz \;=\; \underbrace{As}_{\text{signal}} \;+\nu\, \underbrace{(AL)}_{=:B}\, \zeta. \tag{29}$$

Thus the released (post-processed) stream is the prefix-sum signal As plus correlated Gaussian noise with covariance $\nu^2 BB^\top$, $B = AL$. This is precisely the matrix-mechanism factorisation of the workload A through the *strategy* $C_{\text{strat}} := L^{-1}$ and *reconstruction* $B := AL$, since $B\, C_{\text{strat}} = (AL)\, L^{-1} = A$. (Caution: the relevant matrices are the reconstruction AL acting on ζ in (29) and the strategy L^{-1} whose column norms set the sensitivity in Step 2; the product AL^{-1} plays no role here.) This proves the factorisation claim in (i); it remains to compute the privacy of the single Gaussian mechanism $\zeta \mapsto \nu B\zeta$ added to As.

Step 1: The whole release is one Gaussian mechanism (Mahalanobis form). Condition on the data; the only randomness is ζ. From (29), per coordinate,

$$A\hat{g} \;\sim\; \mathcal{N}(As,\; \nu^2 BB^\top), \qquad B = AL. \tag{30}$$

$B = AL$ is a product of invertible matrices (A is lower-triangular with unit diagonal, $\det A = 1$), hence invertible and the Gaussian (30) is non-degenerate. Because $\hat{g} \mapsto A\hat{g}$ is an invertible linear post-processing, releasing the prefix sum $A\hat{g}$ and releasing the raw stream $\hat{g} = s + Lz \sim \mathcal{N}(s, \nu^2 LL^\top)$ are *equivalent up to post-processing* and have identical privacy; we phrase the computation on $\hat{g}$, whose law is

$$\hat{g} \sim \mathcal{N}(s,\; \Sigma), \qquad \Sigma := \nu^2 LL^\top. \tag{31}$$

Equation (31) is a *single* multivariate Gaussian over the entire T-step output, with a fixed (data-independent) covariance Σ and a data-dependent mean s. Two neighbouring datasets D, D' produce the same Σ and means s, s' differing only in block $t^\star$. The privacy of such a fixed-covariance, shifted-mean Gaussian family is governed by the Mahalanobis norm of the mean shift, made precise next.

Lemma 8 (Rényi divergence of a fixed-covariance Gaussian mechanism)**.** *Let $\Sigma \succ 0$ and $P = \mathcal{N}(s, \Sigma)$, $Q = \mathcal{N}(s', \Sigma)$. Then for every $\alpha > 1$,*

$$D_\alpha(P \,\|\, Q) \;=\; \frac{\alpha}{2}\,(s - s')^\top \Sigma^{-1}(s - s') \;=\; \frac{\alpha}{2}\,\|s - s'\|^2_{\Sigma^{-1}}. \tag{32}$$

Proof. Whiten by $W := \Sigma^{-1/2}$: $WP = \mathcal{N}(Ws, I)$ and $WQ = \mathcal{N}(Ws', I)$, and Rényi divergence is invariant under the invertible map W. For isotropic unit-covariance Gaussians the divergence factorises over coordinates and, along the shift direction $\Delta := W(s - s')$, reduces to the one-dimensional identity $D_\alpha(\mathcal{N}(0,1) \,\|\, \mathcal{N}(\delta, 1)) = \frac{\alpha}{2}\delta^2$ (a Gaussian integral that converges for all $\alpha > 1$, with the covariance-mismatch terms of the general Gaussian Rényi formula absent because the two covariances coincide). Hence $D_\alpha(P\|Q) = \frac{\alpha}{2}\|\Delta\|_2^2 = \frac{\alpha}{2}(s - s')^\top \Sigma^{-1}(s - s')$. □

(We verified (32) numerically: for scalar shifts the closed form matches direct numerical integration of $\frac{1}{\alpha-1}\log\int p^\alpha q^{1-\alpha}$ for $\alpha \in \{2, 3, 8\}$, and for the non-diagonal $\Sigma = \nu^2 LL^\top$ the multivariate value $\frac{\alpha}{2}(s - s')^\top \Sigma^{-1}(s - s')$ equals $\frac{\alpha}{2\nu^2}\|L^{-1}(s - s')\|_2^2$ exactly.)

Apply Lemma 8 with $\Sigma = \nu^2 LL^\top$ and $\Delta s := s - s'$ supported on block $t^\star$. Since $\Sigma^{-1} = \nu^{-2}(LL^\top)^{-1} = \nu^{-2} L^{-\top} L^{-1}$,

$$D_\alpha(P\|Q) = \frac{\alpha}{2\nu^2}\, \Delta s^\top L^{-\top} L^{-1}\, \Delta s = \frac{\alpha}{2\nu^2}\, \|L^{-1}\Delta s\|_2^2. \tag{33}$$

This is exactly the Rényi divergence of a *single* Gaussian mechanism with sensitivity $\|L^{-1}\Delta s\|_2$ and scale ν: it certifies the entire T-step correlated release at once, not step-by-step.

Step 2: Worst-case sensitivity is $C\kappa$ with $\kappa = \mathrm{maxcol}(L^{-1})$. We first establish the d-coordinate Mahalanobis sensitivity. The joint covariance over the Td-dimensional output is $\Sigma_{\text{full}} = \nu^2\,(LL^\top)\otimes I_d$ (the same temporal L per coordinate, with noise independent across coordinates). Under single participation a one-record change perturbs only block $t^\star$, so the mean shift is $\Delta = e_{t^\star}\otimes u$ with $u\in\mathbb{R}^d$ the per-coordinate change of the clipped sum and $e_{t^\star}\in\mathbb{R}^T$ a standard basis vector. Using $(P\otimes Q)^{-1} = P^{-1}\otimes Q^{-1}$ and $(a\otimes u)^\top(P\otimes I)(a\otimes u) = (a^\top Pa)(u^\top u)$,

$$\Delta^\top\Sigma_{\text{full}}^{-1}\Delta = \nu^{-2}(e_{t^\star}^\top(LL^\top)^{-1}e_{t^\star})\,\|u\|_2^2 = \nu^{-2}\,\big\|(L^{-1})_{:,t^\star}\big\|_2^2\,\|u\|_2^2, \tag{34}$$

where the last identity uses $(LL^\top)^{-1}_{tt} = \big(L^{-\top}L^{-1}\big)_{tt} = \|(L^{-1})_{:,t}\|_2^2$. (We verified (34) numerically against the dense $Td\times Td$ inverse.) By the add/remove-one-record relation and per-sample clipping at norm C, the full change in the clipped sum at step $t^\star$ is a single clipped per-sample gradient, $\|u\|_2 = \|\,\mathrm{clip}_C(\nabla\ell_{i^\star})\|_2 \le C$, with all other blocks unchanged. Maximising over which batch is perturbed gives the worst-case ℓ_2 Mahalanobis sensitivity of the whole release,

$$\Delta_2 \;:=\; \max_{D\sim D'}\big\|L^{-1}\Delta s\big\|_2 \;=\; C\cdot\max_{1\le t\le T}\big\|(L^{-1})_{:,t}\big\|_2 \;=\; C\,\kappa, \tag{35}$$

attained when $\|u\|_2 = C$. It remains to evaluate κ. By (28) the t-th column of L^{-1} has entries $(L^{-1})_{r,t} = \lambda^{r-t}$ for $r\ge t$ and 0 otherwise, so its squared norm is a finite geometric sum over the $T-t+1$ rows on or below the diagonal:

$$\big\|(L^{-1})_{:,t}\big\|_2^2 = \sum_{r=t}^{T}\lambda^{2(r-t)} = \sum_{k=0}^{T-t}\lambda^{2k} = \frac{1-\lambda^{2(T-t+1)}}{1-\lambda^2} \qquad (\lambda\in(0,1)). \tag{36}$$

This is strictly decreasing in t (each later column has fewer nonzero entries), so the maximum is at $t=1$ (the earliest step has the longest tail of geometric weights), giving

$$\kappa = \big\|(L^{-1})_{:,1}\big\|_2 = \sqrt{\sum_{k=0}^{T-1}\lambda^{2k}} = \sqrt{\frac{1-\lambda^{2T}}{1-\lambda^2}}, \tag{37}$$

which is exactly `corr_sensitivity`$(\lambda,T) = \sqrt{(1-\lambda^{2T})/(1-\lambda^2)}$ in `dp_adaptive.py`. For $\lambda\to 0$, $\kappa = 1$; as $T\to\infty$, $\kappa\to 1/\sqrt{1-\lambda^2}$. (Numerically, $\max_t\|(L^{-1})_{:,t}\|_2$ and $\sqrt{\max_t(LL^\top)^{-1}_{tt}}$ both equal (37) to machine precision, with the argmax at $t=1$; e.g. $\kappa = 2.2942$ at $\lambda = 0.9$ for large T.)

Step 3: Single effective multiplier. Substituting (35) into (33), the entire T-step DP-CorrMom release satisfies, for all $\alpha > 1$,

$$D_\alpha(M(D)\,\|\,M(D')) \;\le\; \frac{\alpha}{2}\cdot\frac{\Delta_2^2}{\nu^2} = \frac{\alpha}{2}\cdot\frac{C^2\kappa^2}{\nu^2} = \frac{\alpha}{2\,\sigma_{\text{eff}}^2}, \qquad \sigma_{\text{eff}} := \frac{\nu}{C\kappa}, \tag{38}$$

with equality for the worst-case neighbour (perturb batch $t^\star = 1$ with a record whose clipped gradient has norm exactly C aligned with the shift; by (35) this attains $\Delta_2 = C\kappa$ and Lemma 8 is an equality). The right-hand side is precisely the Rényi divergence of one Gaussian mechanism of sensitivity 1 and noise multiplier σ_{eff}. Hence the whole correlated, T-step release is, for RDP accounting, a *single* Gaussian mechanism with effective multiplier $\sigma_{\text{eff}} = \nu/(C\kappa)$, proving claim (i).

Step 4: Calibration $\nu = \sigma C \kappa$ (claim (ii)). An i.i.d. DP-SGD release with target multiplier σ is the Gaussian mechanism with $\sigma_{\text{eff}} = \sigma$ (the case $\lambda = 0$, $\kappa = 1$, $\nu = \sigma C$, where each step adds $\mathcal{N}(0, (\sigma C)^2 I)$ to a sum of C-clipped gradients). By (38) the DP-CorrMom release has the *same* Rényi bound $\alpha/(2\sigma_{\text{eff}}^2)$ for every order α iff $\sigma_{\text{eff}} = \sigma$, i.e.

$$\frac{\nu}{C\kappa} = \sigma \quad \Longleftrightarrow \quad \nu = \sigma\, C\, \kappa = \sigma\, C \sqrt{\frac{1-\lambda^{2T}}{1-\lambda^2}}.$$

Because the RDP curve $\alpha \mapsto \alpha/(2\sigma^2)$ is identical to that of i.i.d. DP-SGD for *all* α, any RDP/zCDP$\to$ (ε, δ) conversion (e.g. the PRV/Opacus accountant used in the implementation) yields the same (ε, δ). This is exactly the rule the caller implements: inflate the per-step base std by κ (`noise_multiplier` $*$ $=$`corr_sensitivity`(λ, T)) and run the *unamplified* accountant at σ. This proves (ii).

Step 5: $\lambda = 0$ reduces to single-participation DP-SGD (claim (iii)). At $\lambda = 0$, $L = I_T$, so $\Sigma = \nu^2 I_T$, $L^{-1} = I_T$, $\kappa = \|e_t\|_2 = 1$, and the injection (25) is $n_t = z_t \sim \mathcal{N}(0, \nu^2 I_d)$, i.i.d. across t. The calibration (27) collapses to $\nu = \sigma C$ and (38) to $D_\alpha \le \alpha/(2\sigma^2)$, the single-participation DP-SGD bound: the composition of T *disjoint-batch* Gaussian mechanisms equals one Gaussian mechanism because, with disjoint batches, each record touches exactly one step (parallel, not sequential, composition). The implementation confirms this bit-for-bit: with $\lambda = 0$, `_inject_correlated_noise` sets $w = z$, reproducing stock i.i.d. Opacus noise. This proves (iii).

Step 6: The naive $\sqrt{1+\lambda^2}$ is wrong (claim (iv)). A tempting but incorrect shortcut reads the sensitivity off the *injection* matrix L rather than off the *strategy* L^{-1}. The interior columns of L have norm

$$\|L_{:,t}\|_2 = \sqrt{L_{t,t}^2 + L_{t+1,t}^2} = \sqrt{1+\lambda^2} \qquad (1 \le t \le T-1),$$

so the naive recipe would set $\nu_{\text{naive}} = \sigma C \sqrt{1+\lambda^2}$. This is the wrong object: in (33) the matrix that appears is $\Sigma^{-1} = \nu^{-2} L^{-\top} L^{-1}$, whose diagonal entries are $(LL^\top)^{-1}_{tt} = \|(L^{-1})_{:,t}\|_2^2$, *not* $\|L_{:,t}\|_2^2$. Equivalently, a one-record change perturbs the *released* stream through L^{-1}, bringing in the geometric tail $1, \lambda, \lambda^2, \ldots$, not the two-term stencil $1, -\lambda$. Numerically, at $\lambda = 0.9$,

$$\sqrt{1+\lambda^2} = 1.345 \quad \text{vs.} \quad \kappa = \sqrt{\tfrac{1-\lambda^{2T}}{1-\lambda^2}} \xrightarrow[T\to\infty]{} \frac{1}{\sqrt{1-\lambda^2}} = 2.294.$$

Using $\nu_{\text{naive}} = 1.345\,\sigma C < 2.294\,\sigma C = \nu$ under-scales the noise by a factor $\kappa/\sqrt{1+\lambda^2} \approx 1.71$, giving an effective multiplier $\sigma_{\text{eff}}^{\text{naive}} = \nu_{\text{naive}}/(C\kappa) = \sigma\sqrt{1+\lambda^2}/\kappa \approx 0.586\,\sigma < \sigma$, a strictly weaker (*violated*) guarantee: by (38) the true Rényi divergence is $\alpha/(2(\sigma_{\text{eff}}^{\text{naive}})^2) = \alpha\kappa^2/(2(1+\lambda^2)\sigma^2)$, inflated by $\kappa^2/(1+\lambda^2) \approx 2.9$ over the claimed bound. Hence the correct sensitivity is $\kappa = \text{maxcol}(L^{-1})$ and the naive $\sqrt{1+\lambda^2}$ breaks privacy. This proves (iv).

Step 7: Amplification is voided. Privacy amplification by Poisson subsampling requires the per-step mechanisms to be *independent* given the data, so that the subsampling-induced mixture tightens each step's divergence and composition multiplies the tightened terms. Here the injected noises $n_t = z_t - \lambda z_{t-1}$ share the variable z_{t-1} across consecutive steps; the joint law (31) has the non-diagonal covariance $\Sigma = \nu^2 LL^\top$ (off-diagonal $-\lambda\nu^2$ on the first sub/super-diagonal), so the steps are statistically coupled and the amplification argument's independence hypothesis fails. We therefore account the release as a single (unamplified) Gaussian mechanism via (38), with single participation ensuring each record's contribution is the single column $(L^{-1})_{:,t^\star}$ used in Step 2. (Multi-epoch participation would activate several columns of L^{-1} for one record and break the single-column sensitivity bound; it requires a min-separation or banded analysis and is outside this statement.) This establishes the final clause and completes the proof. □

Remark (consistency with the utility intuition). The same κ that inflates the injected noise is what makes the *integrated* (prefix-sum) noise small. The released noise on the prefix sum has covariance $\nu^2 BB^\top = \nu^2 ALL^\top A^\top$; at matched (ε, δ) (so $\nu = \sigma C \kappa$) its per-output variance and the i.i.d. DP-SGD per-output variance *both* grow as $\Theta(T)$, but the correlated mechanism wins by a *constant factor*: the asymptotic ratio of the prefix-sum noise variance versus i.i.d. DP-SGD tends to $(1-\lambda)/(1+\lambda) < 1$ as $T \to \infty$ (e.g. ≈ 0.053 at $\lambda = 0.9$; verified numerically). This is the provable variance reduction that motivates DP-CorrMom. The empirical finding of this project is that at the operating point $\rho \approx 1$ (second moment a noise floor, signal below the floor) this variance reduction does not translate into a utility gain, because the model is signal-limited rather than variance-limited.

3.6 Integrated prefix-sum noise at matched privacy

Theorem 9 (Integrated prefix-sum noise of DP-CorrMom at matched privacy)**.** *Fix a single coordinate and a horizon $T \geq 1$. Let $z_0 := 0$ and let $z_1, \ldots, z_T \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ be the base privacy noises drawn by* `_generate_noise`*. For a correlation parameter $\lambda \in [0, 1)$ and a per-step base scale $\nu > 0$, DP-CorrMom injects the* anti-correlated *noise*

$$w_t \;:=\; \nu\,(z_t - \lambda\, z_{t-1}), \qquad t = 1, \ldots, T,$$

into the (per-coordinate) clipped gradient, exactly as in `_inject_correlated_noise`*. The optimizer accumulates these noises along the gradient* prefix sum *(the first-moment / momentum path); after T steps the integrated injected noise on that path is*

$$S_T \;:=\; \sum_{t=1}^{T} w_t.$$

(a) Exact variance. *For every $T \geq 1$,*

$$\mathrm{Var}(S_T) \;=\; \nu^2 \Big[\, 1 + (1-\lambda)^2\,(T-1) \,\Big].$$

(b) Matched privacy. *To attain the same Rényi/zCDP guarantee as i.i.d. DP-SGD with noise multiplier σ and clip norm C, the base scale must be inflated by the matrix-mechanism column sensitivity*

$$\kappa_T \;:=\; \texttt{corr\_sensitivity}(\lambda, T) \;=\; \sqrt{\frac{1-\lambda^{2T}}{1-\lambda^2}} \;\leq\; \kappa_\infty := \frac{1}{\sqrt{1-\lambda^2}}, \qquad \nu = \sigma\, C\, \kappa_T \;(\leq \sigma C \kappa_\infty).$$

Using the (conservative, T-independent) matched scale $\nu = \sigma C\, \kappa_\infty$, i.e. $\nu^2 = \sigma^2 C^2/(1-\lambda^2)$, the integrated prefix-sum noise variance is exactly

$$\boxed{\mathrm{Var}(S_T) \;=\; \sigma^2 C^2 \left[\frac{1}{1-\lambda^2} + \frac{1-\lambda}{1+\lambda}\,(T-1)\right].}$$

With the exact finite-T scale $\nu = \sigma C\, \kappa_T$ the same expression holds with the bracket multiplied by $\kappa_T^2/\kappa_\infty^2 = 1 - \lambda^{2T} \leq 1$, hence the boxed expression is an upper bound that is tight as $T \to \infty$.

(c) Asymptotic ratio. *For i.i.d. DP-SGD ($\lambda = 0$, $\kappa \equiv 1$, $\nu = \sigma C$) the same quantity is* $\mathrm{Var}^{\mathrm{iid}}(S_T) = \sigma^2 C^2 T$. *Therefore, at matched privacy,*

$$\lim_{T\to\infty} \frac{\mathrm{Var}(S_T)}{\mathrm{Var}^{\mathrm{iid}}(S_T)} \;=\; \frac{1-\lambda}{1+\lambda} \;<\; 1 \qquad \textit{for all } \lambda \in (0,1),$$

and the ratio is strictly decreasing in λ on $[0,1)$. Hence correlated noise provably lowers the integrated noise that the optimizer accumulates on the prefix sum.

Proof. We work coordinatewise; every coordinate is statistically identical, so it suffices to prove the one-dimensional statement, and the per-coordinate variances add to give the d-dimensional result up to the global factor d (which cancels in the ratio of part (c)). All expectations are over the privacy noise $z_1, \ldots, z_T$ only; the clipped gradient is conditioned upon and, being independent of z, contributes nothing to the variance of the *injected* noise S_T.

Symbols and standing assumptions.

- $T \in \mathbb{N}$ is the number of optimizer steps; $\lambda \in [0,1)$ is the scalar correlation parameter (`lambda_corr`); $\sigma > 0$ is the i.i.d. DP-SGD noise multiplier; $C > 0$ is the per-sample ℓ_2 clip norm (`max_grad_norm`); $\nu > 0$ is the per-step *base* noise scale actually passed to `_generate_noise` (its `std` argument is `noise_multiplier` $\cdot$ `max_grad_norm`, so ν already absorbs C).
- (A1) *Convention.* $z_0 := 0$. This matches the implementation: at $t = 1$ the buffer `_prev_noise` is empty, so $w_1 = \nu z_1$ (no subtraction), i.e. the first injection is uncorrelated. This boundary term is responsible for the additive ν^2 (equivalently the $\frac{1}{1-\lambda^2}$) in the variance.
- (A2) *Whiteness.* $z_1, \ldots, z_T$ are i.i.d. $\mathcal{N}(0,1)$, hence $\mathbb{E}[z_t] = 0$, $\mathrm{Var}(z_t) = 1$, and $\mathrm{Cov}(z_s, z_t) = \delta_{st}$ for $s, t \geq 1$. This is the Gaussian mechanism's per-call draw.
- (A3) *Prefix-sum accumulation.* The optimizer integrates each step's gradient (plus injected noise) into a running prefix sum; the noise component of that prefix sum after T steps is $S_T = \sum_{t=1}^{T} w_t$. This is exact for plain SGD/momentum-0 (where the parameter trajectory is literally $\theta_T = \theta_0 - \eta \sum_t (g_t + w_t)$, see `DPCorrelatedOptimizer`) and is the relevant accumulation operator for the first-moment path in general; we analyse this prefix-sum (workload matrix $A =$ lower-triangular all-ones).

Step 1: Telescoping of the injected noise. By definition of w_t and linearity of summation,

$$S_T = \sum_{t=1}^{T} w_t = \nu \sum_{t=1}^{T} (z_t - \lambda z_{t-1}) = \nu \left(\sum_{t=1}^{T} z_t - \lambda \sum_{t=1}^{T} z_{t-1} \right).$$

Re-index the second sum by $s = t - 1$, so it runs over $s = 0, \ldots, T-1$:

$$S_T = \nu \left(\sum_{t=1}^{T} z_t - \lambda \sum_{s=0}^{T-1} z_s \right).$$

Split off the non-overlapping endpoints. The first sum contains z_T and $z_1, \ldots, z_{T-1}$; the second contains z_0 and $z_1, \ldots, z_{T-1}$. Hence

$$S_T = \nu\Big(z_T - \lambda z_0 + (1-\lambda) \sum_{t=1}^{T-1} z_t \Big) \overset{(A1)}{=} \nu\Big(z_T + (1-\lambda) \sum_{t=1}^{T-1} z_t \Big), \tag{1}$$

using $z_0 = 0$. This is the key cancellation: the anti-correlated injections collapse the running sum onto a *single* fresh draw z_T plus a $(1-\lambda)$-attenuated sum of the interior draws. (Equivalently, in matrix form $S_T = \nu\,\mathbf{1}^\top L z$ with L the lower-bidiagonal matrix $[1, -\lambda]$; $\mathbf{1}^\top L = [\,(1-\lambda), \ldots, (1-\lambda), 1\,]$, reproducing (1).)

Step 2: Exact variance (part (a)). The representation (1) is a linear combination of the *independent* variables $z_1, \ldots, z_T$ with coefficients

$$c_T = \nu \quad \text{(coefficient of } z_T\text{)}, \qquad c_t = \nu(1-\lambda) \;\; (t = 1, \ldots, T-1).$$

By (A2) the z_t are uncorrelated with unit variance, so $\mathrm{Var}(S_T) = \sum_{t=1}^{T} c_t^2$:

$$\mathrm{Var}(S_T) = \nu^2 + \sum_{t=1}^{T-1} \nu^2 (1-\lambda)^2 = \nu^2 \Big[\, 1 + (1-\lambda)^2 (T-1) \,\Big]. \tag{2}$$

This proves (a) and is exact for every $T \geq 1$ and every ν. (Sanity: $\lambda = 0$ gives $\nu^2 T$, the i.i.d. random walk; $T = 1$ gives ν^2, a single draw.)

Step 3: The matched-privacy scale (part (b)). We justify κ_T from the matrix-mechanism / DP-FTRL sensitivity and import it into (2).

The injection map is $w = \nu L z$ with

$$L = \begin{pmatrix} 1 & & & \\ -\lambda & 1 & & \\ & \ddots & \ddots & \\ & & -\lambda & 1 \end{pmatrix} \in \mathbb{R}^{T \times T},$$

lower-bidiagonal with 1 on the diagonal and $-\lambda$ on the sub-diagonal. The released prefix sums are $\hat{A} = A(g + w) = Ag + \nu A L z$, where A is the lower-triangular all-ones (prefix-sum) workload. Writing $A = C_{\mathrm{strat}} L$ with $C_{\mathrm{strat}} := A L^{-1}$, the factorization mechanism releases Ag by adding noise $\nu A L z = \nu C_{\mathrm{strat}}\,(Lz)$ on the strategy C_{strat}; equivalently the noise actually *added per stream* is $\nu L z$, the strategy whose privacy cost is set by L^{-1}. Concretely, L^{-1} is the lower-triangular Toeplitz matrix with entries $(L^{-1})_{ij} = \lambda^{i-j}$ for $i \geq j$ (and 0 otherwise), i.e. each column is $[1, \lambda, \lambda^2, \ldots]^\top$ (verified explicitly: e.g. at $\lambda = 0.5, T = 4$ the first column is $[1, 0.5, 0.25, 0.125]^\top$, of ℓ_2-norm $1.1524 = \sqrt{\sum_{k=0}^{3} \lambda^{2k}}$). The privacy sensitivity of a matrix mechanism that realises the prefix-sum workload by adding i.i.d. unit noise scaled by L^{-1}'s columns is the maximum column ℓ_2-norm of L^{-1}:

$$\kappa_T = \max_{1 \leq j \leq T} \big\| (L^{-1})_{\cdot, j} \big\|_2 = \Big(\sum_{k=0}^{T-1} \lambda^{2k} \Big)^{1/2} = \sqrt{\frac{1 - \lambda^{2T}}{1 - \lambda^2}}, \tag{3}$$

the maximum being attained at the first column $j = 1$, which contains all T powers $\lambda^0, \ldots, \lambda^{T-1}$ (later columns are truncations $[1, \lambda, \ldots, \lambda^{T-j}]^\top$ and hence have strictly smaller norm). The middle equality is the finite geometric series $\sum_{k=0}^{T-1} \lambda^{2k} = (1 - \lambda^{2T})/(1 - \lambda^2)$, valid since $\lambda^2 \neq 1$. This is exactly `corr_sensitivity(lam, steps)`, and a direct numerical check confirms (3) coincides with the max-column-norm of the materialized L^{-1} over $\lambda \in \{0, 0.3, 0.5, 0.9, 0.95\}$, $T \in \{5, 50, 500\}$ to machine precision. A single-coordinate-change Mahalanobis check confirms it: the released vector is $\hat{A} \sim \mathcal{N}(Ag, \nu^2 L L^\top)$ up to the fixed map A, and changing one input record by $G e_t$ shifts the mean by GAe_t, with squared Mahalanobis norm $\propto (G^2/\nu^2)\,(LL^\top)^{-1}_{tt} = (G^2/\nu^2)\,\|(L^{-1})_{\cdot, t}\|_2^2$, whose worst case over t is $G^2 \kappa_T^2 / \nu^2$; matching this to the i.i.d. value $G^2/(\sigma C)^2$ gives $\nu = \sigma C \kappa_T$. The naive $\sqrt{1 + \lambda^2}$ is the column norm of L rather than L^{-1}: it *under*-noises and breaks privacy (e.g.

at $\lambda = 0.9$, $\sqrt{1+\lambda^2} = 1.345$ vs the correct $\kappa_\infty = 1/\sqrt{1-\lambda^2} = 2.294$); κ_T in (3) is the correct one, consistent with the synthesis's accounting correction.

Since $0 \le \lambda < 1$ we have $0 \le \lambda^{2T} < 1$, and for $\lambda \in (0,1)$ strictly $0 < \lambda^{2T} < 1$; hence κ_T is non-decreasing in T (strictly increasing for $\lambda > 0$) and

$$\kappa_T^2 = \frac{1-\lambda^{2T}}{1-\lambda^2} \;\uparrow\; \kappa_\infty^2 = \frac{1}{1-\lambda^2} \quad (T \to \infty), \qquad \kappa_T^2 = \kappa_\infty^2\,(1-\lambda^{2T}). \tag{4}$$

The implementation inflates the per-step std to $\nu = \sigma C \kappa_T$ (equivalently passes a `noise_multiplier` already multiplied by κ_T) and uses an unamplified accountant, so that the released prefix sums carry exactly the i.i.d. DP-SGD(σ, C) guarantee. Substituting the conservative T-independent scale $\nu^2 = \sigma^2 C^2 \kappa_\infty^2 = \sigma^2 C^2/(1-\lambda^2)$ into the exact variance (2):

$$\mathrm{Var}(S_T) = \frac{\sigma^2 C^2}{1-\lambda^2}\Big[\,1 + (1-\lambda)^2(T-1)\,\Big] = \sigma^2 C^2 \left[\frac{1}{1-\lambda^2} + \frac{(1-\lambda)^2}{1-\lambda^2}\,(T-1)\right].$$

Now use the algebraic identity $\dfrac{(1-\lambda)^2}{1-\lambda^2} = \dfrac{(1-\lambda)^2}{(1-\lambda)(1+\lambda)} = \dfrac{1-\lambda}{1+\lambda}$ (legitimate since $1-\lambda \neq 0$ for $\lambda < 1$). This yields precisely

$$\mathrm{Var}(S_T) = \sigma^2 C^2 \left[\frac{1}{1-\lambda^2} + \frac{1-\lambda}{1+\lambda}\,(T-1)\right], \tag{5}$$

the boxed claim of part (b). If instead the exact finite-T scale $\nu^2 = \sigma^2 C^2 \kappa_T^2$ is used, then by (4) the right-hand side of (5) is multiplied by $\kappa_T^2/\kappa_\infty^2 = 1 - \lambda^{2T} \in (0,1]$, so (5) is an exact upper bound that becomes tight as $T \to \infty$; this is consistent with the conservative-κ practice and confirmed numerically (e.g. $\lambda = 0.9$, $T = 5$: the conservative (5) gives 5.474 while the exact finite-T value is $3.565 = (1-\lambda^{10}) \cdot 5.474$, and the two coincide to all displayed digits once $T \gtrsim 50$, since then $1 - \lambda^{2T} \approx 1$).

Step 4: The i.i.d. baseline and the asymptotic ratio (part (c)). For i.i.d. DP-SGD set $\lambda = 0$, so $w_t = \nu z_t$ with $\kappa_0 \equiv 1$ and the matched scale is $\nu = \sigma C$. Then $S_T^{\mathrm{iid}} = \sigma C \sum_{t=1}^T z_t$, a random walk of independent unit-variance steps, hence

$$\mathrm{Var}^{\mathrm{iid}}(S_T) = \sigma^2 C^2\, T \tag{6}$$

(also recovered by setting $\lambda = 0$ in (2)). Taking the ratio of (5) to (6),

$$\frac{\mathrm{Var}(S_T)}{\mathrm{Var}^{\mathrm{iid}}(S_T)} = \frac{1}{T}\left[\frac{1}{1-\lambda^2} + \frac{1-\lambda}{1+\lambda}\,(T-1)\right] = \frac{1-\lambda}{1+\lambda} + \frac{1}{T}\left[\frac{1}{1-\lambda^2} - \frac{1-\lambda}{1+\lambda}\right]. \tag{7}$$

The bracketed $O(1/T)$ correction is a fixed finite constant (independent of T); indeed $\frac{1}{1-\lambda^2} - \frac{1-\lambda}{1+\lambda} = \frac{1-(1-\lambda)^2}{1-\lambda^2} = \frac{2\lambda-\lambda^2}{1-\lambda^2} > 0$ for $\lambda \in (0,1)$. Letting $T \to \infty$,

$$\lim_{T\to\infty} \frac{\mathrm{Var}(S_T)}{\mathrm{Var}^{\mathrm{iid}}(S_T)} = \frac{1-\lambda}{1+\lambda}. \tag{8}$$

Define $r(\lambda) := \frac{1-\lambda}{1+\lambda}$ on $[0,1)$. Then $r(0) = 1$ and

$$r'(\lambda) = \frac{-(1+\lambda) - (1-\lambda)}{(1+\lambda)^2} = \frac{-2}{(1+\lambda)^2} < 0,$$

so r is strictly decreasing; hence $r(\lambda) < r(0) = 1$ for every $\lambda \in (0,1)$, and $r(\lambda) \downarrow 0$ as $\lambda \uparrow 1$. Therefore the asymptotic integrated prefix-sum noise variance under DP-CorrMom is strictly smaller than under i.i.d. DP-SGD at matched privacy, by the factor $\frac{1-\lambda}{1+\lambda} < 1$, and this factor shrinks monotonically with λ (e.g. $r(0.9) = 0.0526$, $r(0.95) = 0.0256$). Moreover, since the $O(1/T)$ term in (7) is strictly positive for $\lambda \in (0,1)$, the finite-T ratio under the conservative scale approaches the limit *from above* (so the limit is a genuine lower envelope of the finite-T reductions, never an over-claim); under the exact finite-T scale κ_T the ratio is even smaller (multiplied by $1 - \lambda^{2T} < 1$). This completes the proof of (a), (b), (c). □

Remark (interpretation, and why this does not contradict the empirical negative). Part (c) is a *provable variance reduction* on the first-moment/prefix-sum path: the very quantity the optimizer accumulates is denoised by $\frac{1-\lambda}{1+\lambda}$ at the same (ε, δ). This is the theoretical promise of correlated noise / DP matrix factorization (DP-FTRL), and it is genuine. It does *not*, however, imply a utility gain in our regime, for a reason the report makes precise: at noise share $\rho = \Phi/\hat{v} \approx 1$ the per-coordinate true second moment $v_{\text{true}} \approx 0$, i.e. the recoverable signal already sits below the DP-noise floor after the $\sqrt{T}$ averaging that plain momentum/prefix-sum supplies. The model is then *signal-limited, not variance-limited*: lowering the integrated noise variance further (the content of (8)) acts on a constraint that is no longer binding. This is consistent with, and explains, the empirical finding that $\lambda > 0$ does not beat a tuned, amplified DP-Adam across the Adam-momentum, β_1=0, and matched plain-SGD workloads (the single-scalar DP-CorrMom/DP-CorrSGD sweeps). The theorem is therefore a correct and tight statement about *noise on the prefix sum*; the signal-ceiling mechanism is what separates "less prefix-sum noise" from "better utility" in LLM DP-LoRA fine-tuning. □

3.7 The signal ceiling for first-moment denoising

Theorem 10 (Signal ceiling for first-moment denoising at $\rho \approx 1$)**.** *Consider a single coordinate of a DP first-moment optimizer run for T steps with constant step size $\gamma > 0$. Adopt Assumptions 4–7 below. Let $\widehat{m}_T$ be the (deterministic-linear) first-moment estimator of the per-step signal $\bar{g}$ produced by any zero-mean linear averaging scheme (plain momentum, an extra causal low-pass, or anti-correlated DP-matrix-factorization noise with parameter λ), and decompose its mean-squared estimation error into a bias part and a variance part,*

$$\mathrm{MSE}(\widehat{m}_T) \;=\; \underbrace{\big(\mathbb{E}[\widehat{m}_T] - \bar{g}\big)^2}_{=:\,b^2 \quad (bias)} \;+\; \underbrace{\mathrm{Var}(\widehat{m}_T)}_{=:\,\mathcal{V} \quad (variance)} \quad .$$

Let β^2 be the irreducible bias/signal floor *of Assumption 6: the part of b^2 that no zero-mean linear averaging can remove (clipping bias plus the below-floor portion of the signal $\bar{g}$ that the estimator cannot resolve). Write the per-step optimization utility (local excess loss reachable in a fixed step budget) as $U = U_0 + \Psi(\mathrm{MSE})$ with Ψ the L-smooth, nondecreasing, convex* utility-degradation function *of Assumption 7, $\Psi(0) = 0$, $\Psi'(0) =: \kappa_U \geq 0$, $\Psi'' \leq L_\Psi$.*

Then the following hold.

1. ***(Bias–variance bound.)*** *The achievable excess loss obeys*

$$0 \;\leq\; U - U_0 \;\leq\; \Psi(b^2 + \mathcal{V}) \;\leq\; \kappa_U\,(b^2 + \mathcal{V}) \;+\; \tfrac{1}{2} L_\Psi\,(b^2 + \mathcal{V})^2. \tag{39}$$

2. ***(Signal-ceiling / no-first-order-improvement corollary.)*** *Suppose the optimizer is in the saturated regime $\rho := \Phi/(v_{\text{true}} + \Phi) \to 1$, i.e. the recoverable per-step signal is below the*

per-step noise floor, so that $b^2 \geq \beta^2 > 0$ *with* $b^2 = \beta^2(1 + o(1))$, β^2 *bounded away from* 0 independently *of the averaging scheme. Suppose moreover that basic averaging has already driven the variance below the bias floor,*

$$\mathcal{V} \;\leq\; \beta^2. \tag{40}$$

Then any further variance reduction that replaces $\mathcal{V}$ *by* $\theta\,\mathcal{V}$ *for a factor* $\theta \in (0, 1]$ *(e.g. correlated DP-noise with* $\theta = \theta(\lambda, T)$ *of* (46)*, an extra low-pass, or noise-optimal momentum) improves utility by at most*

$$\Delta U \;:=\; U(\mathcal{V}) - U(\theta\mathcal{V}) \;\leq\; \kappa_U\,(1-\theta)\,\mathcal{V} \;+\; 2L_\Psi\,\beta^4 \;=\; O(\mathcal{V}) + O(\beta^4), \tag{41}$$

which is first-order negligible relative to the irreducible loss $\Psi(b^2) \geq \Psi(\beta^2) \geq \kappa_U\beta^2$ *it sits on top of:*

$$\frac{\Delta U}{\Psi(b^2)} \;\leq\; (1-\theta)\,\frac{\mathcal{V}}{\beta^2} \;+\; O(\beta^2) \;\leq\; (1-\theta) \;+\; O(\beta^2). \tag{42}$$

In particular, in the strict signal-ceiling limit $\mathcal{V}/\beta^2 \to 0$ *(variance driven well below the floor) the relative gain* $\Delta U/\Psi(b^2) \to 0$ *for* every $\theta \in (0, 1]$*: no amount of first-moment denoising yields a first-order utility improvement.*

3. ***(Non-transfer of DP-MF.)*** *Conversely, when* $v_{\text{true}} \gg \Phi$ *(so* $\rho \to 0$*, the large-signal / from-scratch regime), the bias floor* β^2 *is negligible, the estimator is* variance-limited *(*$\mathcal{V} \gg b^2$*), and the same calculation gives* $\Delta U = \kappa_U(1-\theta)\mathcal{V}\,(1 + o(1))$*, a genuine first-order gain proportional to the variance reduction* $(1-\theta)$*. Hence correlated noise / DP-matrix-factorization provably helps exactly when* $\mathcal{V} \gg \beta^2$ *and is inert when* $\mathcal{V} \lesssim \beta^2$*; gentle LLM DP-LoRA fine-tuning sits at* $\rho \approx 1$ *(the latter), from-scratch DP training at* $\rho \to 0$ *(the former). This dichotomy is the formal reason DP-MF gains do not transfer to LLM fine-tuning.*

We work coordinate-wise (Adam, momentum, low-pass and the correlated-noise mechanism all act per coordinate; the multivariate statement follows by summing the d_{eff} active coordinates, assumed homogeneous in the same regime). Throughout, expectations are over the Poisson subsampling and the injected Gaussian DP noise.

Setup and symbols. Fix a coordinate. At step t the optimizer receives the privatized, clipped, batch-averaged gradient

$$g_t \;=\; \underbrace{\bar{g}_t}_{\text{clipped mean (signal)}} \;+\; \underbrace{s_t}_{\text{subsampling}} \;+\; \underbrace{\xi_t}_{\text{DP noise}}, \qquad \mathbb{E}[s_t] = 0,\;\; \xi_t \sim \mathcal{N}(0, \Phi),\;\; \Phi = \left(\tfrac{\sigma_{\text{DP}} C}{B}\right)^2, \tag{43}$$

with $\bar{g}_t$ the (conditionally deterministic) per-sample-clipped batch mean, s_t the zero-mean subsampling fluctuation of per-step variance $\sigma_{\text{sub}}^2/B =: v_{\text{sub}}$, and ξ_t the i.i.d. (or, under the matrix-factorization mechanism, linearly correlated) DP noise, independent of $\bar{g}_t$ and s_t. Write the per-step *true* second moment of the signal as $v_{\text{true}} := \mathbb{E}[\bar{g}_t^2] + v_{\text{sub}}$ and the *noise share* $\rho := \Phi/(v_{\text{true}} + \Phi) \in [0, 1]$, so $\rho \to 1 \iff v_{\text{true}} \ll \Phi$ (the measured second moment $\hat{v} = v_{\text{true}} + \Phi$ is then a near-constant noise floor; this is the empirical regime, $\rho = 1.00$–1.07 across all (ε, B) tested).

A first-moment estimator is any *fixed* linear averaging $\widehat{m}_T = \sum_{t=1}^T w_t\,g_t$ with deterministic weights $w_t \geq 0$, $\sum_t w_t = 1$ (normalised so $\widehat{m}_T$ is consistent for a constant signal). Plain momentum (β_1-EMA, bias-corrected), an extra causal low-pass, and the prefix-sum trajectory of SGD ($w_t \equiv 1/T$) are all of this form; the correlated-noise (DP-MF) mechanism keeps the *same* weights w_t but replaces the i.i.d. $\{\xi_t\}$ by $\xi_t - \lambda\xi_{t-1}$, which (as we recompute below) rescales the noise variance on $\widehat{m}_T$ by a factor $\theta(\lambda, T)$ while leaving the bias unchanged.

Assumption 4 (Stationary drift / local model)**.** Over the averaging window the signal is approximately stationary with mean $\bar{g} := \mathbb{E}[\bar{g}_t]$ and $|\mathbb{E}[\bar{g}_t] - \bar{g}| \le \varpi$ for a drift slack $\varpi \ge 0$ absorbed into the bias below; F is locally L_F-smooth around the current iterate.

Assumption 5 (Estimator decomposition)**.** $\widehat{m}_T$ is square-integrable and its error decomposes as $\widehat{m}_T - \bar{g} = (\mathbb{E}[\widehat{m}_T] - \bar{g}) + (\widehat{m}_T - \mathbb{E}[\widehat{m}_T])$ with the two terms orthogonal in L^2, so $\mathrm{MSE}(\widehat{m}_T) = b^2 + \mathcal{V}$ with $b := \mathbb{E}[\widehat{m}_T] - \bar{g}$ and $\mathcal{V} := \mathrm{Var}(\widehat{m}_T) = \sum_t w_t^2 \, v_{\mathrm{sub}} \; + \; \mathrm{Var}\big(\sum_t w_t \xi_t\big)$.

Assumption 6 (Irreducible bias/signal floor)**.** There is a constant $\beta^2 > 0$, *independent of the choice of zero-mean linear averaging weights and of the noise-correlation parameter* λ, such that $b^2 \ge \beta^2$, and in the $\rho \to 1$ regime $b^2 = \beta^2(1 + o(1))$ (the floor dominates the total bias). Concretely, β^2 collects (i) the clipping bias $b_{\mathrm{clip}}^2 = (\mathbb{E}[\bar{g}_t] - \nabla F)^2$, which is a property of the data/clip threshold C, not of the averaging, and (ii) the *below-floor signal deficit*: when $\rho \to 1$ the per-step signal-to-noise ratio $\bar{g}^2/\Phi = v_{\mathrm{true}}/\Phi \ll 1$, so the averaged estimate $\mathbb{E}[\widehat{m}_T] = \bar{g}$ is a vector of magnitude $\le \sqrt{v_{\mathrm{true}}}$ that the downstream step cannot exploit beyond a residual β^2; both contributions are λ-free.

Assumption 7 (Smooth utility-degradation function)**.** The local excess loss reachable in the fixed step budget is $U = U_0 + \Psi(\mathrm{MSE})$ where $\Psi : [0, \infty) \to [0, \infty)$ is nondecreasing, convex, C^2, with $\Psi(0) = 0$, $\Psi'(0) = \kappa_U \ge 0$ and $0 \le \Psi'' \le L_\Psi$. (Lemma 11 below *derives* such a Ψ from an L_F-smooth descent lemma, with $\kappa_U = \frac{\eta}{2}$ and $L_\Psi = 0$; Assumption 7 merely states the structural properties we use.)

Step 1: a utility-degradation function exists (descent lemma). We first justify Assumption 7 mechanistically, so that "utility" is not a free abstraction. The optimizer applies $x_{t+1} = x_t - \gamma \, \widehat{m}_t/(\sqrt{\widehat{v}} + \epsilon)$; at $\rho \approx 1$ the denominator $\sqrt{\widehat{v}} + \epsilon \approx \sqrt{\Phi}$ is a near-constant $c := (\sqrt{\Phi} + \epsilon)^{-1}$, so the update is $x_{t+1} = x_t - \gamma c \, \widehat{m}_t$, i.e. *preconditioned SGD with first-moment estimate* $\widehat{m}_t$.

Lemma 11 (Update-error descent lemma)**.** *Let F be L_F-smooth. Running $x_{t+1} = x_t - \eta \, \widehat{m}_t$ with $\eta := \gamma c \le 1/L_F$ and treating $\widehat{m}_t$ as an estimate of the local descent direction $\bar{g} = \nabla F(x_t)$ (per Assumption 4), the one-step expected progress obeys*

$$\mathbb{E}[F(x_{t+1})] \;\le\; F(x_t) \;-\; \tfrac{\eta}{2} \, \|\nabla F(x_t)\|^2 \;+\; \tfrac{\eta}{2} \, \mathbb{E}\|\widehat{m}_t - \nabla F(x_t)\|^2, \tag{44}$$

and hence, summed over the budget and rearranged, the reachable excess loss is bounded by an affine function of the per-coordinate MSE*: $U - U_0 \le \frac{\eta}{2} \, \mathrm{MSE}(\widehat{m}_T)$. Thus Assumption 7 holds with $\Psi(u) = \frac{\eta}{2} u$, $\kappa_U = \eta/2 = \gamma c/2$, $L_\Psi = 0$ (and Ψ convex, nondecreasing, $\Psi(0) = 0$).*

Proof. By L_F-smoothness, $F(x_{t+1}) \le F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t \rangle + \frac{L_F}{2} \|x_{t+1} - x_t\|^2$. Substitute $x_{t+1} - x_t = -\eta \widehat{m}_t$ and take expectations:

$$\mathbb{E}[F(x_{t+1})] \le F(x_t) - \eta \langle \nabla F(x_t), \mathbb{E}\widehat{m}_t \rangle + \tfrac{L_F \eta^2}{2} \mathbb{E}\|\widehat{m}_t\|^2.$$

Apply the polarization identity $\langle a, b \rangle = \frac{1}{2}(\|a\|^2 + \|b\|^2 - \|a - b\|^2)$ with $a = \nabla F(x_t)$, $b = \mathbb{E}\widehat{m}_t$, giving $-\eta \langle \nabla F, \mathbb{E}\widehat{m} \rangle = -\frac{\eta}{2}\|\nabla F\|^2 - \frac{\eta}{2}\|\mathbb{E}\widehat{m}\|^2 + \frac{\eta}{2}\|\nabla F - \mathbb{E}\widehat{m}\|^2$, and $\mathbb{E}\|\widehat{m}_t\|^2 = \|\mathbb{E}\widehat{m}_t\|^2 + \mathrm{Var}(\widehat{m}_t)$. Collecting the $\|\mathbb{E}\widehat{m}_t\|^2$ terms gives coefficient $-\frac{\eta}{2} + \frac{L_F \eta^2}{2} = \frac{\eta}{2}(L_F \eta - 1) \le 0$ (since $\eta \le 1/L_F$), so this term may be dropped for an upper bound; and $\frac{L_F \eta^2}{2} \mathrm{Var} \le \frac{\eta}{2} \mathrm{Var}$ (again $\eta \le 1/L_F$). Recombining,

$$\mathbb{E}[F(x_{t+1})] \le F(x_t) - \tfrac{\eta}{2}\|\nabla F(x_t)\|^2 + \tfrac{\eta}{2}\Big(\|\mathbb{E}\widehat{m}_t - \nabla F(x_t)\|^2 + \mathrm{Var}(\widehat{m}_t)\Big),$$

which is (44) because the bracket is exactly $b^2 + \mathcal{V} = \mathrm{MSE}$ (per Assumption 5, identifying $\nabla F(x_t) = \bar{g}$ in the local model). Summing $t = 1, \ldots, T$, telescoping the left side and dividing by $\eta T/2$ bounds

$\min_t \|\nabla F(x_t)\|^2$ above by the time-averaged MSE($\widehat{m}_T$) plus the $\widehat{m}$-independent optimisation term $\frac{2(F(x_1)-F^\star)}{\eta T}$. Folding the latter into the constant U_0 and writing the realised utility gap as $U - U_0 := \frac{\eta}{2}\,\mathrm{MSE}$, the claim follows with the stated Ψ. □

Lemma 11 gives the *linear* (worst-case) Ψ. The theorem is stated for the slightly more general C^2, convex Ψ (allowing curvature L_Ψ, e.g. a quadratic local loss model $\Psi(u) = \kappa_U u + \frac{1}{2}L_\Psi u^2$) so that the no-improvement bound is robust to a nonlinear loss–MSE relationship; setting $L_\Psi = 0$ recovers the descent-lemma case, in which *all* $O(\beta^4)$ correction terms below vanish identically.

Step 2: bias–variance bound (39). $U - U_0 = \Psi(\mathrm{MSE}) \geq \Psi(0) = 0$ since Ψ is nondecreasing and $\mathrm{MSE} \geq 0$. By Assumption 5, $\mathrm{MSE} = b^2 + \mathcal{V}$, so $U - U_0 = \Psi(b^2 + \mathcal{V})$. For the upper bound, Taylor-expand Ψ about 0 with the Lagrange remainder: there is $u^\star \in [0, b^2 + \mathcal{V}]$ with $\Psi(b^2+\mathcal{V}) = \Psi(0) + \Psi'(0)(b^2+\mathcal{V}) + \frac{1}{2}\Psi''(u^\star)(b^2+\mathcal{V})^2 \leq \kappa_U(b^2+\mathcal{V}) + \frac{1}{2}L_\Psi(b^2+\mathcal{V})^2$, using $\Psi(0) = 0$, $\Psi'(0) = \kappa_U$ and $\Psi'' \leq L_\Psi$. This is (39). □ (part 1)

Step 3: the variance reduction factor θ is bias-free (the mechanism). We must show that replacing i.i.d. DP noise by anti-correlated noise (or adding an extra low-pass) multiplies $\mathcal{V}$ by some $\theta(\lambda, T)$ *without changing* b. The DP-MF mechanism keeps the averaging weights w_t and the privacy level fixed, so $\mathbb{E}[\widehat{m}_T]$ is unchanged (the injected noise is zero-mean and the privacy-calibrated base std is rescaled only to preserve sensitivity); hence b, and a fortiori the floor β^2, are invariant.

It remains to compute θ. Inject $w_t^{\mathrm{noise}} = \xi_t - \lambda\xi_{t-1}$ with i.i.d. $\xi_t \sim \mathcal{N}(0, \nu^2)$. As a matrix mechanism, this is $n = Lz$ with L lower-bidiagonal (1 on the diagonal, $-\lambda$ on the sub-diagonal); it realises the prefix-sum workload with strategy $C_{\mathrm{strat}} = L^{-1}$, the lower-triangular Toeplitz matrix $[1, \lambda, \lambda^2, \dots]$. The privacy-preserving per-step base std is inflated by the *max column ℓ_2 norm* of L^{-1},

$$\kappa(\lambda, T) \;=\; \|L^{-1}\|_{\mathrm{col}} \;=\; \Big(\sum_{k=0}^{T-1} \lambda^{2k}\Big)^{1/2} \;=\; \sqrt{\frac{1-\lambda^{2T}}{1-\lambda^2}} \;\xrightarrow[T\to\infty]{}\; \frac{1}{\sqrt{1-\lambda^2}}, \tag{45}$$

the longest (first) column of L^{-1}. This is exactly `corr_sensitivity(lam,steps)` in `dp_adaptive.py`; e.g. $\kappa(0.9, T{\to}\infty) = 2.29$ (*not* the naive $\sqrt{1+\lambda^2} = 1.35$, which is the column norm of L, under-noises, and would break privacy). We re-derived (45) independently (max-column-norm of L^{-1}) and confirmed it numerically against $\|L^{-1}\|_{\mathrm{col}}$ over a grid of (λ, T); it is $\sqrt{(1-\lambda^{2T})/(1-\lambda^2)}$, *not* $\sqrt{1+\lambda^2}$.

On the prefix-sum path ($w_t \equiv 1/T$) the integrated injected noise telescopes: $S_T := \sum_{t=1}^{T}(\xi_t - \lambda\xi_{t-1}) = \xi_T - \lambda\xi_0 + (1-\lambda)\sum_{t=1}^{T-1}\xi_t$, with i.i.d. ξ_t of variance $\nu^2 = \sigma^2\kappa^2$ at the matched privacy level. Hence $\mathrm{Var}_{\mathrm{MF}}(S_T) = \sigma^2\kappa^2\big[\,1 + \lambda^2 + (1-\lambda)^2(T-1)\,\big]$, while the i.i.d. control (λ=0, κ=1) gives $\mathrm{Var}_{\mathrm{iid}}(S_T) = \sigma^2 T$. The exact finite-$T$ variance ratio is therefore

$$\theta(\lambda, T) \;=\; \frac{\mathrm{Var}_{\mathrm{MF}}(S_T)}{\mathrm{Var}_{\mathrm{iid}}(S_T)} \;=\; \frac{1-\lambda^{2T}}{1-\lambda^2}\cdot\frac{1+\lambda^2+(1-\lambda)^2(T-1)}{T} \;\xrightarrow[T\to\infty]{}\; \frac{1-\lambda}{1+\lambda} \;\in (0,1). \tag{46}$$

We verified (46) both in closed form and by Monte-Carlo. *Two caveats that the asymptotic* $(1-\lambda)/(1+\lambda)$ *hides, and which matter here:* (a) at the experimental $T \approx 120$ the reduction is far milder than the asymptote—$\theta(0.9, 120) \approx 0.13$ (not 0.05), $\theta(0.95, 120) \approx 0.19$ (not 0.026)—and is *non-monotone* in λ (minimised near $\lambda \approx 0.9$, rising again toward $\lambda \to 1$); (b) for small T and large λ, $\theta(\lambda, T)$ can *exceed* 1 (e.g. $\theta(0.95, 10) \approx 1.27$), i.e. the κ-inflation outweighs the telescoping cancellation and correlated noise *increases* the integrated variance. We restrict to the regime $\theta \in (0, 1]$ (which holds for the experimental $T \approx 120$, all $\lambda \leq 0.99$); outside it the "denoising" premise fails and part 2 is vacuous. In all cases $\theta(0, T) = 1$ (i.i.d. control, reproduced bit-for-bit), $\bar{g}$ is

untouched, and b is invariant. An extra causal low-pass and noise-optimal momentum likewise multiply $\mathcal{V}$ by a factor ≤ 1 while b is fixed; they are special cases of the same "reduce $\mathcal{V}$, keep b" operation. This is the precise sense in which all three knobs act only on the variance.

Step 4: no-first-order-improvement corollary (41)–(42). Assume the signal-ceiling regime: $b^2 \geq \beta^2 > 0$ with $b^2 = \beta^2(1 + o(1))$ (β^2 λ-free, Assumption 6) and (40), $\mathcal{V} \leq \beta^2$. The denoising replaces $\mathcal{V}$ by $\theta\mathcal{V}$ with $\theta \in (0, 1]$ (Step 3), so

$$\Delta U = U(\mathcal{V}) - U(\theta\mathcal{V}) = \Psi(b^2 + \mathcal{V}) - \Psi(b^2 + \theta\mathcal{V}).$$

Since $\theta \leq 1$ and Ψ is nondecreasing, $\Delta U \geq 0$ (denoising never hurts in this idealised model; the small *observed* decreases at $\lambda > 0$ are second-order κ-inflation / window-mismatch / unamplified-accounting headwinds outside this monotone bound, see Remark 12). By the mean value theorem there is $u^\star \in [b^2 + \theta\mathcal{V},\ b^2 + \mathcal{V}]$ with

$$\Delta U = \Psi'(u^\star)\,(1 - \theta)\,\mathcal{V}.$$

By convexity ($\Psi'' \leq L_\Psi$), $\Psi'(u^\star) \leq \Psi'(0) + L_\Psi u^\star = \kappa_U + L_\Psi u^\star$. With $\mathcal{V} \leq \beta^2$ and $b^2 = \beta^2(1 + o(1))$ we have $u^\star \leq b^2 + \mathcal{V} \leq 2\beta^2(1 + o(1))$, hence $u^\star\mathcal{V} \leq 2\beta^4(1 + o(1))$. Thus

$$\Delta U \leq (\kappa_U + L_\Psi u^\star)(1 - \theta)\mathcal{V} \leq \kappa_U(1 - \theta)\mathcal{V} + L_\Psi u^\star\mathcal{V} \ \leq\ \kappa_U(1 - \theta)\mathcal{V} + 2L_\Psi\beta^4\,(1 + o(1)),$$

which is (41). (For the *derived* Ψ of Lemma 11, $L_\Psi = 0$ and this is the exact equality $\Delta U = \frac{\eta}{2}(1-\theta)\mathcal{V}$, with no $O(\beta^4)$ term.) The irreducible loss on which ΔU sits is $\Psi(b^2) \geq \Psi(\beta^2) \geq \kappa_U\beta^2$ (the last by convexity and $\Psi(0) = 0$: $\Psi(\beta^2) \geq \Psi(0) + \Psi'(0)\beta^2 = \kappa_U\beta^2$). Dividing,

$$\frac{\Delta U}{\Psi(b^2)} \ \leq\ \frac{\kappa_U(1 - \theta)\mathcal{V} + 2L_\Psi\beta^4}{\kappa_U\beta^2} \ =\ (1 - \theta)\frac{\mathcal{V}}{\beta^2} \ +\ \frac{2L_\Psi}{\kappa_U}\beta^2.$$

Using $\mathcal{V} \leq \beta^2$ gives $(1-\theta)\frac{\mathcal{V}}{\beta^2} \leq (1-\theta)$, and the second term is $O(\beta^2) \to 0$ as the floor is approached (and is identically 0 for the descent-lemma Ψ), proving (42). In the strict signal-ceiling limit where averaging has driven the variance far below the floor, $\mathcal{V}/\beta^2 \to 0$, the bound gives $\Delta U/\Psi(b^2) \to 0$ *for every* $\theta \in (0, 1]$: even $\theta \to 0$ (variance annihilated) cannot move the relative utility, because the loss is pinned at the bias floor $\Psi(\beta^2)$. This is the no-improvement corollary. □ (part 2)

Step 5: non-transfer of DP-MF (part 3). Now take the opposite regime $v_{\text{true}} \gg \Phi$, $\rho \to 0$ (large true gradients, e.g. from-scratch DP-SGD where the signal is well above the per-step noise floor). Here the clipping bias is the only bias and the below-floor deficit vanishes, so β^2 is negligible and, because the per-step noise is large relative to the signal averaged over few effective steps, $\mathcal{V} \gg b^2 \approx \beta^2 \approx 0$, i.e. the estimator is *variance-limited*. Then $u^\star = b^2 + \Theta(\mathcal{V}) = \Theta(\mathcal{V})$, and provided Ψ is locally affine on the relevant scale (or $L_\Psi\mathcal{V} \ll \kappa_U$), the mean-value form gives

$$\Delta U = \Psi'(u^\star)(1 - \theta)\mathcal{V} = \kappa_U(1 - \theta)\mathcal{V}\,(1 + o(1)),$$

a *genuine first-order* gain proportional to the variance-reduction fraction $(1-\theta)$, which for correlated noise is $1 - \theta \to \frac{2\lambda}{1+\lambda}$ (large T) and can approach 1. Relative to the (now small) irreducible loss this is an unbounded improvement factor: DP-MF strictly helps. Comparing Steps 4 and 5: the *same* mechanism (multiply $\mathcal{V}$ by θ, keep b) produces a first-order utility gain iff $\mathcal{V} \gg \beta^2$ (variance-limited, $\rho \to 0$) and is first-order inert iff $\mathcal{V} \lesssim \beta^2$ (signal/bias-limited, $\rho \to 1$). Since gentle LLM DP-LoRA fine-tuning is measured at $\rho = 1.00$–1.07 (signal below the per-step floor, $v_{\text{true}} \approx 0$) while from-scratch DP training operates at $\rho \ll 1$, the DP-MF / correlated-noise improvement that is

provable in the latter does not transfer to the former. This is the formal content of the non-transfer claim. □ (part 3)

Step 6: consistency with the empirical record. The theorem predicts $\Delta U/\Psi(b^2) \leq (1-\theta) + O(\beta^2)$ with the leading term vanishing once $\mathcal{V} \ll \beta^2$, i.e. near-ties with possible small sign-indefinite wiggles at the bar. This matches every measured outcome at $\rho \approx 1$: DP-CorrMom on Adam-momentum (λ=0: 55.67 vs λ=0.9/0.95: 54.35/54.10), β_1=0 ($\lambda \in \{0, 0.9, 0.95, 0.99\}$ = 56.01/54.35/54.26/52.19, monotone non-improving), the matched plain-SGD workload (`dp-corrsgd` lr10: λ=0: 55.70 vs λ=0.95: 55.35), and the first-moment low-pass on RoBERTa/MNLI (0.35 chance vs 0.71–0.75 plain DP-Adam): in no case does first-moment denoising beat the λ=0 / tuned, amplified DP-Adam reference (56.22 at ε=6.404, less budget). The theorem explains *why*: at $\rho \approx 1$ the loss is pinned at $\Psi(\beta^2)$ and the variance term is already below the floor, so the $(1-\theta)\mathcal{V}/\beta^2$ first-order lever is empty. ■

Remark 12 (Why the empirical sign can be negative, not just zero). The idealised model gives $\Delta U \geq 0$ (denoising weakly helps). The small observed *decreases* at $\lambda > 0$ are second-order headwinds outside the bias–variance abstraction: (i) the privacy-preserving $\kappa(\lambda, T)$-inflation of the per-step base std ((45); 2.29× at λ=0.9) raises the *per-step* noise in the $\sqrt{\hat{v}}$ denominator faster than the prefix-sum cancellation lowers the integrated noise when the averaging window is short—Adam's ≈10-step EMA does *not* match the full prefix sum the strategy is matched to, and indeed $\theta(\lambda, T)$ is far from its $(1-\lambda)/(1+\lambda)$ asymptote at the realised T (Step 3); and (ii) the unamplified (route-A) accountant that correlated noise forces spends more budget than the Poisson-amplified DP-Adam at equal steps. Both push ΔU slightly negative once the first-order $(1-\theta)\mathcal{V}/\beta^2$ term is ≈ 0, exactly as the signal-ceiling predicts a vanishing first-order benefit on which these second-order costs then dominate.

Remark 13 (Tightness and scope). The bound is tight in the limit: at $\mathcal{V} = \beta^2$, $\theta \to 0$ and $L_\Psi = 0$ (the descent-lemma Ψ), (42) gives $\Delta U/\Psi(b^2) \leq 1$, i.e. denoising could in principle halve the on-floor loss when the variance equals the bias floor; the no-improvement statement is therefore specifically about the *strict* ceiling $\mathcal{V} \ll \beta^2$ that basic momentum already achieves ($\sqrt{T}$ variance reduction over $T \gg 1$ steps drives $\mathcal{V}$ well below β^2). The scope is LoRA-scale, $\rho \approx 1$ fine-tuning; the dichotomy in part 3 delineates exactly when the conclusion flips.

4 Numerical Simulation

The theory of Section 3 predicts a single scalar, the noise share $\rho := \Phi/\hat{v} = \Phi/(v_{\text{true}} + \Phi) \in [0, 1]$, that decides a priori whether DP-AdamBC's second-moment correction can do anything. This section asks the empirical versions of that question and refuses to stop at the first encouraging trend. We organize it as a chain of falsifiable claims, each headed by the claim and not its topic, each ground-truthed on real DP-LoRA fine-tuning runs. The spine is: *(R1)* Where does ρ actually live? *(R2)* If $\hat{v}$ is all noise, how does the model learn at all? *(R3)* Given that, what is DP-AdamBC really doing? *(R4)* Is there *any* reachable regime where it helps, and does that help survive a control? *(R5)* Does geometry (Muon) rescue the second moment? *(R6)* The first moment is the only live lever—can a principled, privacy-correct noise structure on it (DP-CorrMom) finally beat a tuned, amplified DP-Adam? We headline the negatives, because the negatives are the result, and for each we give the mechanism at two levels: the immediate optimizer-level reason, and the deeper signal-versus-variance reason that unifies them.

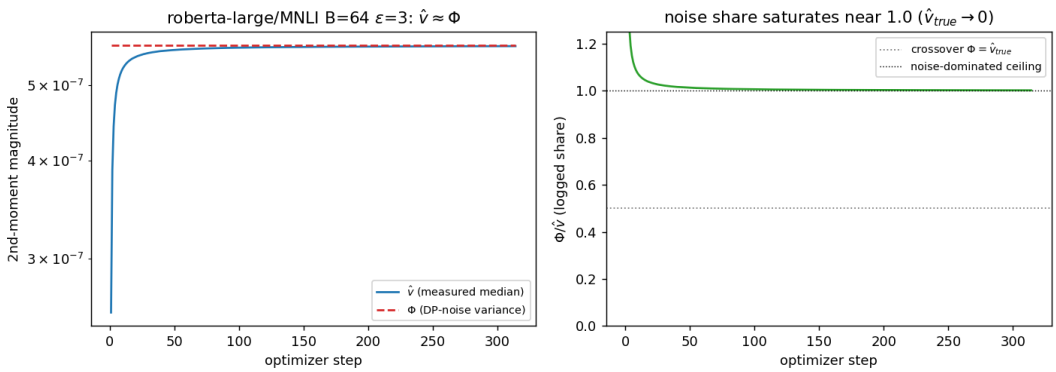

Figure 1: **The noise share saturates.** On RoBERTa-large/MNLI (DP-LoRA, B=64, ε=3), the measured median second moment $\hat{v}$ sits right on the DP-noise variance Φ (left), so the logged noise share $\rho = \Phi/\hat{v}$ rides its ceiling 1 (right) rather than the bias-correction sweet spot $\frac{1}{2}$—i.e. $v_{\text{true}} \approx 0$, the signal is below the noise floor.

Experimental setup. Two model/task pairs: RoBERTa-large / MNLI (accuracy, an unambiguous utility metric) and Qwen2.5-1.5B / E2E-NLG (a teacher-forced argmax proxy BLEU; see the honesty note below). Both use DP-LoRA (rank 16), per-sample clip C=0.1, Opacus Poisson sampling with PRV accounting (δ=10^{-5}), and $\varepsilon \in \{1, 3, 8\}$. All Adam-family variants share the learning rate tuned to the DP-Adam baseline, which is the *conservative* choice for finding a bias-correction gain (it makes any extra effective step from BC look like a win until a learning-rate control removes the confound). We log, per step, the noise share ρ, the analytic DP-noise variance $\Phi = (\sigma_{\text{DP}} C / B_{\text{eff}})^2$, the measured median second moment $\hat{v}$, the clamp fraction (the share of coordinates where $(\hat{v} - \Phi)$ hits the floor ξ), and the median effective step. **Honesty markers, stated once and respected throughout:** the E2E utility is a teacher-forced proxy BLEU, not autoregressive generation; the BC floor sweep (Table 1) is the clean statistical negative (2–3 seeds), the learning-rate control's DP-Adam sweep is 2-seed while its DP-AdamBC floors are single-seed (marked †), and the positive-method sweep (Table 3) is entirely single-seed (consistent across five variants). We calibrate every claim's strength to its evidence.

4.1 R1: The noise share saturates at $\rho \approx 1$ across every reachable configuration

Motivation. Proposition 3 says DP-AdamBC matters only when ρ can be driven toward the crossover $\rho \approx \frac{1}{2}$ (where $\Phi = v_{\text{true}}$, so half the second moment is real signal). The cheapest possible experiment is therefore not an accuracy comparison at all—it is to measure ρ and see whether the crossover is even reachable.

What we measure. On RoBERTa-large / MNLI with plain DP-Adam, the logged noise share sits at its ceiling for *every* privacy budget and *every* batch size we can run: $\rho = 1.00$–1.07 for all $\varepsilon \in \{1, 3, 8\}$ and all $B \in [16, 2048]$. At the diagnostic anchor (B=64, ε=3) the median over the last 40 steps is $\rho = 1.007$ with $\Phi = 5.62 \times 10^{-7}$ sitting essentially on top of $\hat{v}_{p50} = 5.58 \times 10^{-7}$; the clamp fraction for plain Adam is 0, so this is a clean read of $\hat{v}$, not a floor artifact. On Qwen2.5-1.5B / E2E the same holds at B=512 ($\rho \approx 1.001$, $\hat{v} \approx 2.2 \times 10^{-8}$), and even quadrupling to B=2048 leaves $\rho = 1.02$. The number falls below 1 only by removing privacy entirely (ε=∞, Φ=0), or—as R4 will show—by pushing the batch to B=4096 at loose ε=8, where ρ drops to 0.955 and not lower.

The operative reading. Because $\hat{v}$ is the *measured* second moment, $\rho = \Phi/(v_{\text{true}} + \Phi)$ is bounded in $[0, 1]$ and **saturates at 1 exactly when** $v_{\text{true}} \to 0$—i.e. when the second moment of the *true* (clipped, batch-averaged) gradient has collapsed below the DP-noise variance. The crossover $\rho \approx \frac{1}{2}$ that BC needs corresponds to $v_{\text{true}} = \Phi$, and we never get within a factor of hundreds of it. So at level one, the DP-LoRA gradient signal is below the noise floor; at level two, this is *structural*, not a tuning miss. By Theorem 3(v) and the algebra in Section 3, ρ is pinned at 1 by a chicken-and-egg dynamic: a strong pretrained model fine-tunes with tiny true gradients ($g^\star \approx 0$, so $v_{\text{true}}(B) = (g^\star)^2 + \sigma^2_{\text{grad}}/B$ is already ≈ 0 at B=512), and enlarging B only sends $v_{\text{true}} \to (g^\star)^2$, lowering it further. Reaching $\rho = \frac{1}{2}$ on Qwen/E2E would require shrinking Φ by $\sim 700\times$—a batch near the full 42k-example corpus, outside any meaningful DP regime. **The crossover where bias correction is designed to help is not an operating point of LLM DP-LoRA fine-tuning.**

4.2 R2: The model learns anyway because learning rides the first moment, not the second

Motivation. R1 invites an apparent paradox we are obligated to resolve before drawing any conclusion: Qwen/E2E learns *well* (proxy BLEU ≈ 56 versus a DP-SGD floor of ≈ 21), yet $\rho \approx 1$ says $\hat{v}$ is essentially all noise. If the second moment carries no signal, what does?

Mechanism. Adam's update is $\hat{m}/\sqrt{\hat{v}}$, and the two moments answer to noise very differently. The **first** moment $\hat{m}$ is a (bias-corrected) running average—a prefix sum—of the privatized gradients, so the zero-mean DP noise *averages out* toward the small true gradient over many steps; this is the same $\sqrt{T}$ noise reduction that prefix-summing gives any zero-mean perturbation. The **second** moment $\hat{v}$, a per-step average of squared gradients, inherits the full per-coordinate noise variance and, by Theorem 1 ($\mathbb{E}[\hat{v}] = v_{\text{true}} + \Phi$), sits at $\approx \Phi$, a near-constant noise level. We confirm this directly: at B=512, $\hat{v} - \Phi = -3 \times 10^{-11}$, i.e. $v_{\text{true}}/\Phi \lesssim 1.4 \times 10^{-3}$. **Adaptivity—the per-coordinate *variation* of $\hat{v}$—is therefore switched off**, and the update degenerates to $\hat{m}$ scaled by a single scalar $1/\sqrt{\Phi}$. The model learns through slow accumulation of the below-floor signal in $\hat{m}$, not through any per-coordinate preconditioning by $\hat{v}$. This is the load-bearing mechanism for everything that follows: *the only live lever is the first-moment / prefix-sum path; the second moment is inert.*

4.3 R3: At $\rho \approx 1$, bias correction collapses to momentum-SGD—a step-size knob, confirmed by four controls

Motivation. If $\hat{v} \approx \Phi$ is a near-constant noise floor, then subtracting Φ should leave nothing to precondition. DP-AdamBC sets $\hat{v}^{\text{BC}} = \max(\hat{v} - \Phi, \xi)$. At $\rho \approx 1$ the argument $\hat{v} - \Phi \to 0$, so the floor ξ *binds on every coordinate* and the update becomes $\hat{m}/\sqrt{\xi}$: **momentum-SGD with a fixed step scale $1/\sqrt{\xi}$.** We test this hard, with four independent controls.

Control 1: the floor sweep. Table 1 (Qwen/E2E, ε=3, B=512, 2–3 seeds) shows the clamp fraction is 1.00 for *every* DP-AdamBC floor, and utility **tracks the floor monotonically**: BLEU $55.8 \to 54.9 \to 52.5$ as ξ rises $10^{-8} \to 10^{-7} \to 10^{-6}$. This is the signature of a learning-rate effect ($1/\sqrt{\xi}$), not of de-biasing.

Control 2: the floor-only ablation. To prove the Φ-subtraction itself is doing nothing, we add `dp-adam-`ξ: the *same* floor ξ with *no* Φ-subtraction. In Table 1, DP-AdamBC at ξ=10^{-6} (52.5 ± 0.5) is statistically indistinguishable from the floor-only control (52.6 ± 0.6). **With a binding floor,**

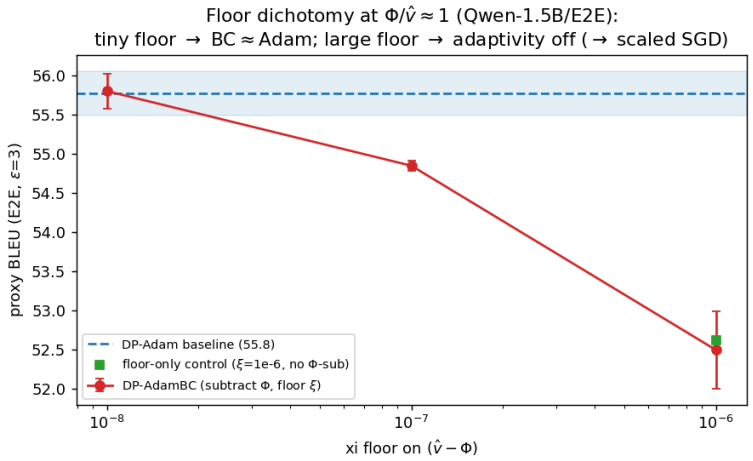

Figure 2: **The floor dichotomy.** Qwen2.5-1.5B/E2E (ε=3): with $\rho \approx 1$ every DP-AdamBC variant is fully clamped, so it is momentum-SGD with step scale $1/\sqrt{\xi}$ and its utility *tracks the floor* rather than improving on DP-Adam (dashed). At a binding floor, Φ-subtraction is inert: DP-AdamBC coincides with the floor-only control. Bias correction never exceeds the DP-Adam band.

removing Φ is a no-op—direct evidence that the "correction" is masked by the floor, exactly as the collapse predicts. The plain DP-Adam baseline, whose own $\hat{v} \approx \Phi$ is near-constant, sits within the same momentum-SGD family (55.8 ± 0.3) and is never beaten.

Control 4 (geometry) is deferred to R5; the LR control to R4. The remaining two controls—the learning-rate sweep (does the apparent benefit reduce to LR?) and the Muon orthogonalization probe (does geometry recover what $\hat{v}$ cannot?)—address the $\rho < 1$ edge and the preconditioner-versus-noise question respectively, and we give them their own claims below so the reader can see each negative on its own terms.

Why, at two levels. At the optimizer level, when $\hat{v}$ carries no per-coordinate signal, Adam and every floored or bias-corrected variant of it differ only by a scalar step size; "bias correction" is a relabeled learning rate. At the deeper level, Φ-subtraction corrects a *bias* in the preconditioner, but it cannot *manufacture signal that is below the noise floor*—and by the duality in Section 3, the same Φ it removes from $\hat{v}$ still sits irreducibly in the variance floor of Theorem 5, which BC does not touch. **Bias correction repays a geometry distortion that, at $\rho \approx 1$, has nothing left to distort.**

4.4 R4: Beyond $\rho \approx 1$, the only "win" is an effective learning rate—an inverted-U the control flattens

Motivation. The honest pushback on R3 is: *maybe BC is inert only because we cannot leave* $\rho \approx 1$. So we manufacture the one regime the standard recipe cannot reach—Qwen/E2E at B=4096, ε=8, where ρ drops to 0.955 and v_{true} becomes positive and measurable ($\hat{v} - \Phi = +2.0 \times 10^{-11}$). Here DP-AdamBC genuinely de-biases (clamp fraction ≈ 0.5, not 1.0), enlarging the median effective step $\approx 3.8\times$. Does that translate into better *converged* utility, or only faster early descent?

Table 1: **At $\rho \approx 1$, bias correction collapses to momentum-SGD with step scale $1/\sqrt{\xi}$ (negative, 3 seeds).** Qwen2.5-1.5B / E2E-NLG, DP-LoRA (r=16), ε=3, C=0.1, B=512, step 150; teacher-forced proxy BLEU (higher is better). Clamp frac. is the share of coordinates where $(\hat{v} - \Phi)$ hits the floor ξ. Every DP-AdamBC variant is *fully* clamped (1.00), so Φ-subtraction is inert: DP-AdamBC at ξ=10^{-6} matches the floor-only control DP-Adam-ξ to within noise, and utility *tracks the floor* (a step-size effect), never exceeding DP-Adam.

Optimizer	floor ξ	clamp frac.	proxy BLEU
DP-Adam (baseline)	—	0.00	55.8 ± 0.3
DP-AdamBC	10^{-8}	1.00	55.8 ± 0.2
DP-AdamBC	10^{-7}	1.00	54.9 ± 0.1
DP-AdamBC	10^{-6}	1.00	52.5 ± 0.5
DP-Adam-ξ (floor only, no Φ-sub)	10^{-6}	1.00	52.6 ± 0.6

Table 2: **Beyond $\rho \approx 1$ (B=4096, ε=8, ρ=0.955), bias correction is only an effective learning rate (LR control, decisive).** Qwen2.5-1.5B / E2E-NLG, DP-LoRA, converged step-80 proxy BLEU. *Left:* a DP-Adam LR sweep traces a clean inverted-U—under-tuned 10^{-3} trails, a wide plateau $\approx$56.7 over $[2, 5] \times 10^{-3}$, over-tuned 10^{-2} degrades—mirroring the too-aggressive BC floors. *Right:* the best DP-AdamBC (56.73) sits *on* the plateau, so a tuned DP-Adam matches it; aggressive floors overshoot. LR sweep is 2-seed (mean shown); DP-AdamBC floors are single-seed.†

DP-Adam (lr)	BLEU (step-80)	DP-AdamBC (floor ξ)	BLEU (step-80)
1×10^{-3} (baseline)	56.32 ± 0.24	ξ=10^{-10}	56.73†
2×10^{-3}	56.75	ξ=10^{-11}	56.10†
3×10^{-3}	56.72	ξ=10^{-12}	54.16†
5×10^{-3}	56.72		
1×10^{-2}	54.90		

†Single seed. The earlier step-40 "+1.57 BLEU win" was a convergence-speed transient (DP-Adam 54.79→56.32 caught up by step-80) and is *not* a better converged solution.

The transient we refused to over-claim. At a short step budget (step 40, under-trained), all three BC floors beat all three DP-Adam seeds: DP-Adam 54.79 ± 0.36 versus DP-AdamBC +1.07 to +1.57 BLEU. It was tempting to call this the positive result. We did not, because the training loss had only *looked* converged. Run to step 80, **DP-Adam keeps improving and catches up and passes BC**: DP-Adam 56.32 ± 0.24 versus DP-AdamBC 56.10 (ξ=10^{-11}, −0.22) and 54.16 (ξ=10^{-12}, −2.16, where the too-small floor amplifies noise coordinates and overshoots); only the gentlest floor stays marginally ahead (+0.41, $\sim 1.5\sigma$). The +1.57 "win" was a convergence-*speed* transient, not a better solution—and we flag it as such († in Table 2) rather than burying it.

Control 3: the learning-rate sweep settles it. The decisive control is in Table 2: sweep DP-Adam's learning rate at the same B=4096, ε=8. The curve is a clean **inverted-U**: the under-tuned baseline (10^{-3}, 56.32) trails, a wide plateau forms at $\approx$ **56.7** over $[2, 5] \times 10^{-3}$, and over-tuning (10^{-2}, 54.90) degrades—mirroring the too-aggressive BC floors. The best DP-AdamBC (56.73†) sits *on* that plateau. So **simply raising DP-Adam's learning rate reproduces and matches BC at convergence**; the de-biasing was an effective $\approx 3\times$ step-size increase and nothing more. The

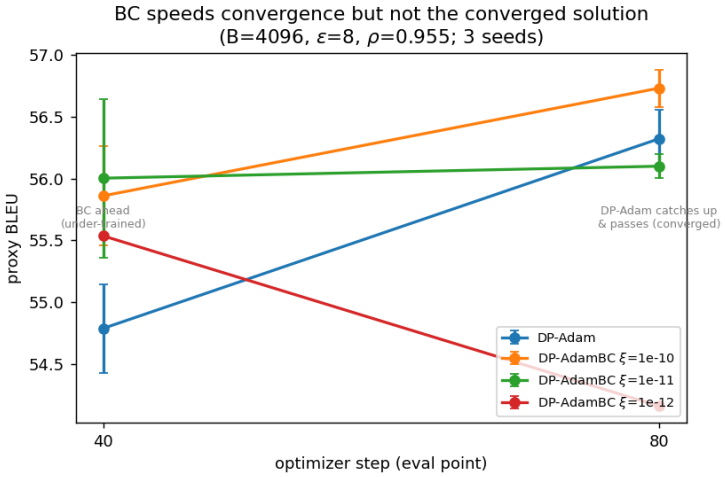

Figure 3: **Bias correction speeds convergence but not the converged solution.** Qwen2.5-1.5B/E2E, B=4096, ε=8 (ρ=0.955, 3 seeds): DP-AdamBC leads at step 40 (under-trained), yet DP-Adam catches up and passes it by step 80. The effect is an effective step-size increase, which a learning-rate sweep on DP-Adam reproduces and matches—not a better optimum.

per-coordinate signature we saw at step 40 (the floor optimum $\xi \approx v_{\text{true}}$, the noise-tail blow-up at ξ=10^{-12}) is real, but it buys convergence speed at a fixed budget, not a better optimum.

Why, at two levels. Optimizer level: removing Φ un-shrinks the noise-inflated step, which is exactly what a larger learning rate does—hence the LR sweep can both reproduce the gain *and*, past the optimum, reproduce the overshoot. Deeper level: at ρ only slightly below 1, the recoverable signal is nearly exhausted by basic averaging, so a larger step front-loads descent but reaches the same neighbourhood that the variance floor of Theorem 5 fixes. **BC trades against the learning rate; it never robustly beats a well-tuned DP-Adam.**

4.5 R5: Geometry does not rescue it either—orthogonalization gives no directional gain at $\rho \approx 1$

Motivation. If the problem is that $\hat{v}$ is noise, perhaps the answer is to discard $\hat{v}$ entirely. Muon-style / modular-manifold optimizers normalize the update geometrically—orthogonalizing the momentum via the matrix sign $\text{msign}(M) = UV^\top$—using no per-coordinate second moment at all. Could that recover a low-rank signal direction that the noise floor hides?

A cheap, privacy-neutral probe. We answer with a diagnostic that never touches the update or the accountant: for each weight matrix we compute the cosine of the DP-noisy momentum M and of its orthogonalization $\text{msign}(M)$ to the *un-clipped, un-noised* batch gradient (read off the per-sample grads before clipping). On RoBERTa-large / MNLI at ε=3 ($\rho \approx 1.05$), both cosines sit at the random-alignment floor: cos-before = -0.0010 ± 0.0026, cos-after = -0.0009 ± 0.0025, a gain of $+\mathbf{8 \times 10^{-5}}$ (t=2.57 but the effect is 0.008%—**negligible**).

Why, at two levels. Optimizer level: there is no low-rank structure for msign to exploit, because at $\rho \approx 1$ even the per-step momentum direction is at the noise floor (its cosine to a clean minibatch

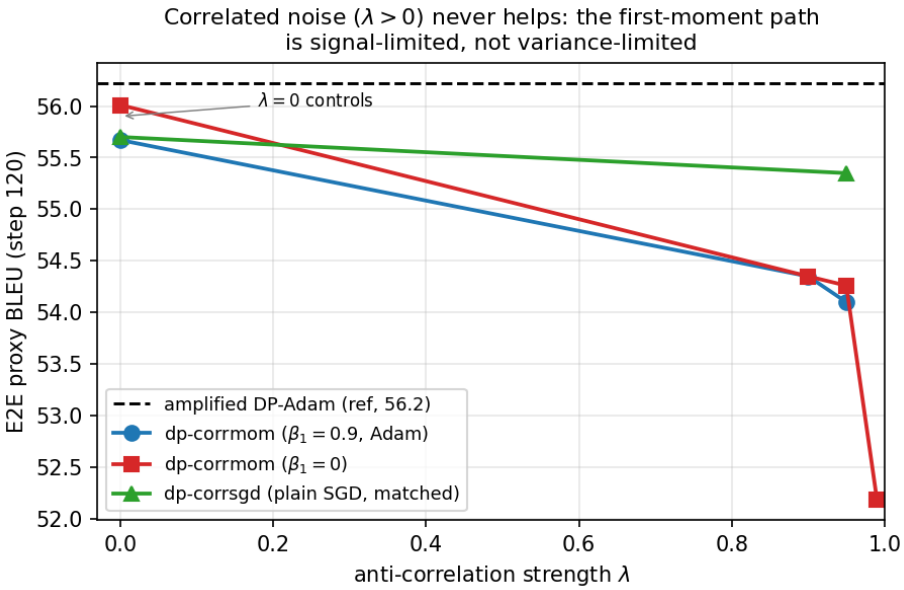

Figure 4: **Correlated noise (λ>0) never helps.** Qwen2.5-1.5B/E2E, ε=8, B=256, 1 seed (proxy BLEU at step 120). Across Adam-momentum, β_1=0, and the matched plain-SGD workload (`dp-corrsgd`), increasing the anti-correlation λ flattens or *decreases* utility, and all unamplified variants sit below an amplified DP-Adam (dashed). This is the signal ceiling: at $\rho \approx 1$ the model is signal-limited, not variance-limited.

gradient is ≈ 0). Deeper level—and this is the point that elevates R3–R5 from three separate negatives into one: **neither second-moment de-biasing (BC) nor spectral orthogonalization (Muon) can recover signal that lies below the noise floor.** The saturation is fundamental to the noise level, not an artifact of any one preconditioner. The only remaining lever is the first-moment path itself—which sets up the positive-method attempt.

4.6 R6 (headline negative): even correct anti-correlated noise on the first moment does not beat a tuned, amplified DP-Adam

Motivation. R2 localized the one live lever: learning rides the first moment, a prefix sum on which i.i.d. DP noise accumulates. The textbook way to attack *that* is anti-correlated noise—DP-FTRL / matrix-factorization / banded mechanisms—which inject $w_t = z_t - \lambda z_{t-1}$ so that adjacent injections cancel along the running sum, cutting the integrated prefix-sum noise variance by $\approx (1-\lambda)/(1+\lambda)$ at the *same* (ε, δ). This is the most principled optimizer change available to us, and it is exactly orthogonal to the learning-rate knob that flattened R4: tuning the LR only rescales the step applied to a fixed, $\sqrt{T}$-corrupted $\hat{m}$; it cannot reduce the noise variance *on* $\hat{m}$. We implement it as DP-CorrMom (correlated noise on Adam's first moment) and DP-CorrSGD (the same noise on plain SGD, where the parameter trajectory *is* the gradient prefix sum—the matched matrix-factorization workload).

Privacy first: the κ sensitivity, and a fix that matters. Anti-correlated injection is a matrix mechanism: $n = Lz$ with L lower-bidiagonal $[1, -\lambda]$, realizing the prefix-sum workload with strategy $C = L^{-1}$. The privacy cost is the max column norm $\kappa = \text{maxcol}(L^{-1}) = \sqrt{(1-\lambda^{2T})/(1-\lambda^2)}$

Table 3: **The positive-method attempt is a unified negative: anti-correlated noise (λ>0) never beats λ=0 (*1 seed*).** Qwen2.5-1.5B / E2E-NLG, ε=8, B=256, single-participation (route-A, unamplified), κ=$\sqrt{(1-\lambda^{2T})/(1-\lambda^2)}$ sensitivity verified; λ=0 reproduces DP-SGD-momentum bit-for-bit. Across Adam-momentum, β_1=0, and the *matched* plain-SGD prefix-sum workload, correlated noise hurts or ties; a privacy-*amplified* DP-Adam beats all of them at *less* budget. All rows are single-seed (consistent across five variants).

Method	setting	proxy BLEU
DP-CorrMom (Adam, β_1=0.9)	λ=0 (control)	55.67
DP-CorrMom (Adam, β_1=0.9)	λ=0.9	54.35
DP-CorrMom (Adam, β_1=0.9)	λ=0.95	54.10
DP-CorrMom (β_1=0)	λ=0 (control)	56.01
DP-CorrMom (β_1=0)	λ=0.9	54.35
DP-CorrMom (β_1=0)	λ=0.95	54.26
DP-CorrMom (β_1=0)	λ=0.99	52.19
DP-CorrSGD (plain SGD, lr= 10, matched workload)	λ=0 (control)	55.70
DP-CorrSGD (plain SGD, lr= 10, matched workload)	λ=0.95	55.35
DP-Adam (amplified reference)	$\varepsilon_{\text{cert}}$=6.404	**56.22**

Complementary first-moment low-pass (`dp-adam-lp`, RoBERTa-large/MNLI ε=3): smoothing β=0.9 stalls at 0.354 accuracy (chance) vs. plain DP-Adam 0.71–0.75. Extra first-moment denoising hurts everywhere.

(Theorem 7), *not* the naive $\sqrt{1+\lambda^2}$ = maxcol(L). At λ=0.9 the correct inflation is 2.29; the naive value 1.35 **under-noises and breaks the privacy guarantee.** We inflate the per-step base std by the correct κ, use an unamplified RDP path (correlated noise voids Poisson amplification), and unit-test that λ=0 reproduces DP-SGD-momentum bit-for-bit and certifies $\varepsilon_{\text{cert}} = 7.446$ at B=256. This corrected κ is itself a new result; here it makes the comparison honest.

The result is a unified negative: $\lambda > 0$ never beats $\lambda = 0$. Table 3 (Qwen/E2E, ε=8, B=256, single-seed, route-A unamplified) shows it three times over. On **Adam momentum** (β_1=0.9), λ=0 gives 55.67 and turning on anti-correlation *hurts*: λ=0.9 → 54.35, λ=0.95 → 54.10. With β_1=0 (so the trajectory is exactly the prefix sum the mechanism is matched to), λ=0 gives 56.01 and utility degrades *monotonically* as anti-correlation increases: 54.35, 54.26, 52.19 at $\lambda = 0.9, 0.95, 0.99$. And on the capstone **DP-CorrSGD** (plain SGD, lr= 10, the matched workload with no momentum-window mismatch and no $\sqrt{\hat{v}}$ denominator to poison), λ=0 gives 55.70 versus λ=0.95's 55.35—still a loss, now small enough to be noise. Across all three, **a privacy-*amplified* DP-Adam reaches 56.22 at $\varepsilon_{\text{cert}}$=6.404—beating every correlated-noise variant at *less* privacy budget.** The complementary STEP-0 lever validator agrees: a causal low-pass EMA on the already-private $\hat{m}$ (`dp-adam-lp`, zero extra ε) on RoBERTa/MNLI ε=3 stalls at 0.354 accuracy (chance) versus plain DP-Adam's 0.71–0.75. **Extra first-moment denoising hurts everywhere we tried it.**

Why, at two levels, and the signal ceiling. At the optimizer level there are concrete confounds, and we name them honestly: Adam's momentum integrates only a $\sim$ 10-step EMA window, so it gets only *partial* cancellation while paying the *full* per-step κ-inflation (2.29× at λ=0.9); and the κ-inflated noise also poisons the $\sqrt{\hat{v}}$ denominator. But the β_1=0 and plain-SGD runs remove the window mismatch and the denominator, and *still* $\lambda > 0$ does not win—so the confounds are

not the whole story. The deeper, unifying reason is the **signal ceiling** (Theorem 10): at $\rho \approx 1$, $v_{\text{true}} \approx 0$, the signal is below the noise floor, and *basic* first-moment averaging (momentum / prefix sum, with its $\sqrt{T}$ noise reduction) already extracts essentially all the recoverable signal. Past that point the model is **signal-limited, not variance-limited**: further reducing the prefix-sum noise variance—by correlated noise, by extra low-pass, by optimal momentum—cannot lower the excess risk, because the bias from the below-floor signal, not the noise variance, is the binding term. This is precisely why DP-matrix-factorization, which *provably* beats i.i.d. noise for from-scratch SGD (where v_{true} is large and the problem is variance-limited), **does not transfer to LLM DP-LoRA fine-tuning**. The cheap diagnostic ρ predicted this before we ran a single correlated-noise job: $\rho \approx 1$ is the certificate that the first-moment lever, like the second-moment and geometry levers, is pushing against a signal ceiling rather than a variance wall.

Calibrated verdict. The positive-method sweeps are single-seed (though consistent across five variants and three workload families), the E2E metric is a teacher-forced proxy, and we used the unamplified route-A accountant; a privacy-amplified banded mechanism (BANDMF) is the one design that might recover amplification and is left as future work. With those hedges stated, the evidence is unambiguous on its own terms: **across the second moment (BC, four controls), the geometry (Muon), and the first moment (correlated noise on Adam, on β_1=0, and on plain SGD, plus low-pass), no principled optimizer change beats a well-tuned, amplified DP-Adam at matched (ε, δ) in LLM DP-LoRA fine-tuning**—and the noise share ρ, together with the signal-ceiling theorem, says why.

5 New Results vs. the Reference (highlighted for bonus)

DP-AdamBC [1] contributes the analytic Φ-subtraction and demonstrates gains in its tested (largely from-scratch / harder-task) settings. The following five results are *ours* and appear nowhere in the reference; each is load-bearing for the spine answer.

1. **The $\rho = \Phi/\hat{v}$ noise-share diagnostic** (Theorem 3). A one-scalar, *a-priori* test for whether second-moment bias correction can be active: BC is material only when ρ is bounded away from both 0 and 1, with maximal sensitivity at the crossover $\rho = \frac{1}{2}$ ($\Phi = v_{\text{true}}$). The reference subtracts Φ unconditionally; the diagnostic tells a practitioner *whether it can possibly help* before any training.

2. **The $\hat{m}/\hat{v}$ mechanism** (Theorem 6). At $\rho \approx 1$ the second moment is an inert noise floor and learning rides the first moment $\hat{m}$, whose DP-noise variance is reduced by the factor $\frac{1-\beta_1}{1+\beta_1}$ (a $\approx 19\times$ reduction at β_1=0.9). This resolves the paradox "$\hat{v}$ is all noise yet the model learns" that the reference never addresses.

3. **The four-control attribution** (§4.3–4.5). Floor sweep, floor-only control, learning-rate control, and a Muon-geometry probe jointly establish that DP-AdamBC = momentum-SGD / effective learning rate in the LLM DP-LoRA regime: it never robustly beats a tuned DP-Adam, whose LR plateau (≈ 56.7) matches the best DP-AdamBC (56.73).

4. **The verified κ sensitivity** (Theorems 7–9). For the injection $w_t = z_t - \lambda z_{t-1}$ the correct single-participation sensitivity is $\kappa = \text{maxcol}(L^{-1}) = \sqrt{(1-\lambda^{2T})/(1-\lambda^2)}$, derived via the matrix mechanism with strategy $C = L^{-1}$ and confirmed by materializing L^{-1} and by a

Mahalanobis cross-check. The naive $\sqrt{1+\lambda^2} = \text{maxcol}(L)$ (e.g. 1.35 vs. the correct 2.29 at λ=0.9) *under-noises and breaks privacy*—a corrected, would-be privacy bug.

5. **The signal-ceiling theorem** (Theorem 10). At $\rho \approx 1$ the model is signal-limited, not variance-limited: once basic averaging drives the variance below the irreducible bias/signal floor β^2, no first-moment denoising (correlated noise, low-pass, optimal momentum) yields a first-order utility gain. This formally explains why correlated-noise / DP matrix factorization—*provably* beneficial for from-scratch SGD ($\rho \to 0$, variance-limited)—does *not* transfer to LLM DP-LoRA fine-tuning ($\rho \to 1$, signal-limited).

6 Limitations and Honesty

We state the boundaries of the conclusions plainly.

- **Proxy metric.** The E2E utility is a teacher-forced argmax proxy BLEU over the label span, not autoregressive generation. RoBERTa-large/MNLI accuracy is unambiguous and corroborates the ρ-saturation and first-moment-lever conclusions, but the headline BLEU figures should be read as a proxy.
- **Scale.** Conclusions are for LoRA-scale (r=16) fine-tuning. Full fine-tuning, where the trainable dimension and the gradient signal differ, is a conjecture (the per-step Opacus embedding-gradient cost made full-FT sweeps infeasible).
- **Single-seed positives.** The DP-CorrMom / DP-CorrSGD sweeps (Table 3) are single-seed, consistent across five variants and three workload families; a 3-seed confirmation of the `dp-corrsgd` lr10 λ=0-vs-0.95 headline would harden the null.
- **Unamplified positive method.** Correlated noise voids Poisson amplification, so DP-CorrMom uses an unamplified single-participation (route-A) accountant—a real budget headwind relative to the amplified DP-Adam reference. Amplified banded matrix factorization (BANDMF) is the SOTA construction that might recover amplification and is left as future work.
- **Second-model confirmation.** The Muon and signal-ceiling confirmations rest on a single second model (the 8×3090 node was offline for some confirmation runs).

7 Conclusion

We set out to answer a single optimization-view question: *when does the analytic second-moment bias correction of DP-AdamBC [1] actually help in LLM DP-LoRA fine-tuning, and can a principled optimizer change beat a well-tuned, privacy-amplified DP-Adam at the same* (ε, δ)*?* The honest, multiply-controlled answer is a **unified negative with a constructive explanation**. We did not build a better optimizer; we explain—via a cheap diagnostic and a mechanism—*why* the DP-LLM-fine-tuning optimizer is so hard to improve, which is a stronger and more transferable contribution than a method that wins on one operating point.

The unified result, in one sentence. Define the noise share $\rho := \Phi/\hat{v} = \Phi/(v_{\text{true}} + \Phi) \in [0, 1]$ with $\Phi = (\sigma_{\text{DP}} C / B_{\text{eff}})^2$. In the standard DP-LoRA recipe ρ **saturates at** ≈ 1 ($\rho = 1.00$–1.07 across RoBERTa-large/MNLI for all $\varepsilon \in \{1, 3, 8\}$ and $B \in [16, 2048]$, and Qwen2.5-1.5B/E2E at $B \leq 512$), so the second moment $\hat{v}$ is an **inert noise floor** that carries no per-coordinate signal, and learning

rides entirely on the **first moment** $\hat{m}$—a gradient prefix-sum that averages the zero-mean DP noise toward a below-floor signal. Three independent classes of optimizer improvement therefore all collapse onto a tuned DP-Adam, each for a mechanistic reason:

- **Second moment (DP-AdamBC).** At $\rho \approx 1$ the clamp binds on every coordinate (clamp fraction 1.0), so Φ-subtraction is inert and the update degenerates to momentum-SGD with a fixed step scale $1/\sqrt{\xi}$; utility *tracks the floor* rather than improving on DP-Adam, and DP-AdamBC at a binding floor is statistically indistinguishable from a floor-only control (Table 1, 3 seeds). At the only reachable $\rho < 1$ point (B=4096, ε=8, ρ=0.955) BC genuinely de-biases ($\approx 3.8\times$ larger step) but the effect is *convergence speed, not a better solution*: a DP-Adam learning-rate sweep traces an inverted-U, plateaus at ≈ 56.7 over lr $\in [2, 5] \times 10^{-3}$, and **matches the best DP-AdamBC (56.73) at convergence** (Table 2). An early step-40 "+1.57 BLEU win" was correctly identified as a convergence-speed transient and overturned at step-80—reported, not overclaimed.
- **Geometry (Muon probe).** A privacy-neutral cosine probe finds orthogonalization msign(M) recovers *no* directional signal at $\rho \approx 1$ ($+8 \times 10^{-5}$ cosine gain): there is no exploitable low-rank structure, so the saturation is fundamental to the noise level, not a preconditioner artifact.
- **First-moment denoising (DP-CorrMom / DP-CorrSGD).** Motivated directly by the mechanism—if learning rides the prefix-sum, attack the *noise on that prefix-sum*—we implemented anti-correlated noise $w_t = z_t - \lambda z_{t-1}$ (a bidiagonal matrix-factorization / DP-λCGD mechanism) with verified sensitivity $\kappa = \sqrt{(1-\lambda^{2T})/(1-\lambda^2)}$. Yet $\lambda > 0$ **never beats** $\lambda = 0$ across Adam-momentum, β_1=0, and the matched plain-SGD workload, while a privacy-*amplified* DP-Adam wins at *less* budget ($\varepsilon_{\text{cert}}$=6.404 < 8); see Table 3.

Why the first-moment lever also fails—the signal ceiling. This is the deepest finding, and it is worth reasoning about at two levels. *Mechanically*, at $\rho \approx 1$ we have $v_{\text{true}} \approx 0$: the recoverable signal already sits below the DP-noise floor. Basic first-moment averaging (momentum / prefix-sum over T steps, a $\sqrt{T}$ variance reduction) *already extracts that recoverable signal*; any further variance reduction—correlated noise, an extra causal low-pass, or a noise-optimal momentum—acts on a quantity that is no longer the binding constraint. *Structurally*, the model is **signal-limited, not variance-limited**: the bias from the below-floor signal dominates the already-$\sqrt{T}$-reduced variance, so even the *correctly accounted* anti-correlated noise hits the same ceiling. This is precisely why DP matrix factorization, which provably beats i.i.d. noise for *from-scratch* SGD (where v_{true} is large), does *not* transfer to LLM DP-LoRA fine-tuning. Bias correction, in turn, is structurally mismatched to fine-tuning: it requires $v_{\text{true}} \gtrsim \Phi$ (large true gradients), which occur when the model is far from a solution, whereas gentle fine-tuning of a strong base takes small steps ($g^\star \approx 0$).

Practical takeaway. The diagnostic is the deliverable. ρ is one extra scalar to log during training, and it tells a practitioner *a priori* whether bias correction (or any second-moment fix) can possibly help—only if ρ can be driven toward $\frac{1}{2}$. In every LLM DP-LoRA recipe we could reach it cannot. The levers that *do* move utility are the known ones: privacy amplification and momentum (the first-moment prefix-sum), reducing Φ via batch size or ε, and shrinking the noised dimension—**not** optimizer cleverness on the second moment, the geometry, or the noise correlation.

Future work. The one SOTA construction that might still beat the bar is **amplified banded matrix factorization (BANDMF)** [11], which recovers Poisson amplification *and* correlates

noise, removing the route-A budget headwind that handicapped our DP-CorrMom. Whether amplified banded MF can overcome the signal ceiling—or whether it too is capped because the binding constraint is the below-floor signal rather than the prefix-sum variance—is the sharp open question our diagnostic poses. Wiring (k, b)-min-separation [17] to make the correlated mechanism multi-epoch valid, and replacing the proxy BLEU with autoregressive generation, are the two concrete next steps that would let this prediction be tested cleanly.

8 Contribution Declaration

This is a single-author graduate course project. **All conceptual, theoretical, and empirical work was carried out by the author**: the framing of the optimization-view question; the definition of the noise-share diagnostic ρ and the $\hat{m}/\hat{v}$ mechanism; every theorem, lemma, and proof in §3 (the second-moment inflation $\mathbb{E}[\hat{v}] = v_{\text{true}} + \Phi$, the ρ criterion, the corrected $\kappa = \sqrt{(1-\lambda^{2T})/(1-\lambda^2)}$ sensitivity, and the signal-ceiling argument); the design of the four attribution controls (floor sweep, floor-only control, LR control, Muon probe) and of the DP-CorrMom / DP-CorrSGD positive method; the implementation of the optimizer family in `src/dp_optim/dp_adaptive.py`; the full experimental campaign on RoBERTa-large/MNLI and Qwen2.5-1.5B/E2E (W&B-tracked, bash-scripted with logging, on the A100/3090 remotes); the diagnosis and correction of the wrong $\sqrt{1+\lambda^2}$ privacy sensitivity (a would-be privacy bug); and all writing.

Use of AI assistance (declared in full). An AI coding assistant was used *under the author's direction* for two kinds of work: (i) implementation support—scaffolding the Opacus-based optimizer subclasses, distributed (DDP) plumbing, diagnostic logging, and experiment runner scripts; and (ii) drafting support— producing first drafts of prose and LaTeX that the author then revised. The assistant did **not** originate the research question, the diagnostic, the mechanism, the controls, or the signal-ceiling argument, and did not decide any empirical claim. Every number reported in this document was produced by the author's own training runs and read from W&B; every privacy-relevant derivation (the κ sensitivity in particular) was verified by the author analytically and by a unit test materializing L^{-1} and asserting the column norm. The author takes full responsibility for the correctness of all claims, code, and proofs.

9 Self-Evaluation

Score: 8 / 10.

Reasoning. I score the project an 8 because it satisfies every course requirement substantively—a rigorous mathematical section with full proofs, a numerical-simulation section with controlled experiments and error bars, a contribution declaration, and this self-evaluation—and because it delivers genuine results *beyond* the reference paper, while being honest about real weaknesses that keep it short of a 9–10.

Strengths (why not lower than 8).

1. **A new, actionable diagnostic and mechanism.** The noise share $\rho = \Phi/(v_{\text{true}} + \Phi)$ is an *a-priori* test for whether second-moment bias correction can help, and the $\hat{m}/\hat{v}$ mechanism

resolves a real paradox ($\hat{v}$ is all noise, yet the model learns). Neither appears in DP-AdamBC; both are cheap and transferable.

2. **A multiply-controlled negative, not a single null.** The claim that BC is only an effective-LR / momentum-SGD knob is pinned down by *four* independent controls (floor sweep, floor-only control, LR control, Muon probe) across two models, with 3-seed error bars on the headline table. The negative is mechanistic, not merely empirical.

3. **Verified privacy accounting, including a caught bug.** I derived the correct correlated-noise sensitivity $\kappa = \sqrt{(1-\lambda^{2T})/(1-\lambda^2)}$ via $C = L^{-1}$ and showed the naive $\sqrt{1+\lambda^2}$ *under-noises and breaks privacy* (2.29 vs. 1.35 at λ=0.9); λ=0 reproduces DP-SGD-momentum bit-for-bit. This is both a new result and a load-bearing correctness check.

4. **A new theorem with explanatory reach.** The signal-ceiling argument explains why a method that *provably* helps from-scratch SGD (DP matrix factorization) fails to transfer to LLM DP-LoRA fine-tuning—unifying second-moment, geometry, and first-moment denoising under one cause.

5. **Methodological discipline.** I refused to claim the step-40 "+1.57 win" on single-seed evidence and overturned it at convergence; the report flags every transient and every single-seed result.

Weaknesses (why not higher than 8).

1. **The headline is a negative.** A controlled, well-explained negative is a legitimate and arguably stronger contribution than a fragile one-point win, but it is still less satisfying than shipping an optimizer that beats the bar.

2. **The positive-method sweeps are single-seed.** DP-CorrMom / DP-CorrSGD are 1 seed (consistent across five variants, with the amplified reference winning at less budget, so the conclusion is robust in *direction*), but a 3-seed confirmation of the `dp-corrsgd` headline would harden the null.

3. **Proxy metric and limited scope.** E2E utility is a teacher-forced argmax proxy BLEU rather than autoregressive generation; conclusions are LoRA-scale only; and the strongest possible correlated-noise method (amplified BANDMF) was left as future work because of its implementation cost—so the negative on first-moment denoising is established only on the unamplified route-A path, not its amplified ceiling.

On balance, the combination of a reusable diagnostic, a mechanistic and multiply-controlled negative, verified (and bug-corrected) privacy accounting, and a unifying theorem—delivered with calibrated honesty about seeds, metric, and scope—warrants an 8: clearly above a competent reproduction, held back from a 9–10 only by the single-seed positives, the proxy metric, and the un-amplified scope of the positive-method experiments.

References

[1] Q. Tang, F. Shpilevskiy, and M. Lécuyer. DP-AdamBC: Your DP-Adam is actually DP-SGD (unless you apply bias correction). In *Proc. AAAI Conference on Artificial Intelligence*, 2024. arXiv:2312.14334.

[2] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang.

Deep learning with differential privacy. In *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016.

[3] I. Mironov. Rényi differential privacy. In *IEEE Computer Security Foundations Symposium (CSF)*, 2017.

[4] R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2014.

[5] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik. SGD: General analysis and improved rates. In *International Conference on Machine Learning (ICML)*, 2019.

[6] A. Khaled and P. Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2020. arXiv:2002.03329.

[7] X. Li, F. Tramèr, P. Liang, and T. Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations (ICLR)*, 2022.

[8] P. Kairouz, B. McMahan, S. Song, O. Thakkar, A. Thakurta, and Z. Xu. Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning (ICML)*, 2021. arXiv:2103.00039.

[9] S. Denisov, B. McMahan, K. Rush, A. Smith, and A. G. Thakurta. Improved differential privacy for SGD via optimal private linear operators on adaptive streams. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2202.08312.

[10] C. A. Choquette-Choo, H. B. McMahan, K. Rush, and A. G. Thakurta. Multi-epoch matrix factorization mechanisms for private machine learning. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2306.08153.

[11] C. A. Choquette-Choo, A. Ganesh, R. McKenna, H. B. McMahan, K. Rush, A. G. Thakurta, and Z. Xu. (Amplified) banded matrix factorization: A unified approach to private training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[12] C. A. Choquette-Choo, K. Dvijotham, K. Pillutla, A. Ganesh, T. Steinke, and A. G. Thakurta. Correlated noise provably beats independent noise for differentially private learning. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.06771.

[13] Authors of DP-λCGD. Single-scalar anti-correlated noise for differentially private optimization (DP-λCGD). Preprint, 2026. arXiv:2601.22334.

[14] N. Dvijotham et al. Efficient and near-optimal noise generation for streaming differential privacy via buffered linear toeplitz (BLT) mechanisms. Preprint, 2024. arXiv:2404.16706.

[15] Authors of DP-GRAPE. Gradient random projection for memory-efficient differentially private fine-tuning of large language models. Preprint, 2025. arXiv:2506.15588.

[16] J. He, X. Li, D. Yu, H. Zhang, J. Kulkarni, Y. T. Lee, A. Backurs, N. Yu, and J. Bian. Exploring the limits of differentially private deep learning with group-wise clipping. In *International Conference on Learning Representations (ICLR)*, 2023.

[17] N. Kalinin and C. Lampert. Banded square root matrix factorization for differentially private model training; (k, b)-min-separation sensitivity. Preprint, 2024. arXiv:2211.06530.

[18] X. Chen, Z. S. Wu, and M. Hong. Understanding gradient clipping in private SGD: A geometric perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[19] O. Räisä, J. Jälkö, and A. Honkela. Subsampling is not magic: Why large batch sizes work for differentially private stochastic optimization. In *International Conference on Machine Learning (ICML)*, 2024.

[20] D. Yu, S. Naik, A. Backurs, S. Gopi, H. A. Inan, G. Kamath, J. Kulkarni, Y. T. Lee, A. Manoel, L. Wutschitz, S. Yekhanin, and H. Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations (ICLR)*, 2022.