STRUCTURAL-PRIOR GEOMETRIC CLIPPING FOR DIFFERENTIALLY PRIVATE SGD

Anonymous authors
Paper under double-blind review

ABSTRACT

Differentially private SGD (DP-SGD) and its adaptive variants treat the parameter space of deep networks as a homogeneous Euclidean manifold, where a single global coordinate system suffices for gradient clipping and noise calibration. We challenge this assumption through *structural priors* from mean-field theory: network parameters are organized into layers with ill-conditioned spectra and pronounced inter-layer heterogeneity. We prove that in the oracle setting (exact geometry, full rank), global clipping and clipping under the structural prior are exactly equivalent—their divergence is entirely privacy-induced, through two mechanisms: (i) geometric statistics estimated from released noisy gradients suffer an anisotropic, batch-size-invariant bias exhibiting the structure of the optimal metric, which we invert in closed form—the first treatment beyond the diagonal case—and characterize the resulting spectral noise floor; (ii) under the rank budgets of privacy-scale models, pooled estimation is structurally blind to directions discarded by top-k truncation, which per-layer estimation retains, enabling closed-form optimal transformations and decoupled convergence. We propose SP-GeoClip, which instantiates the structural prior via independent per-layer geometry estimation with streaming low-rank PCA; all statistics are post-processing of released noisy gradients at no additional privacy cost, and the end-to-end (ε, δ) guarantee is identical to standard DP-SGD. Across nine benchmarks spanning five layer families, SP-GeoClip is uniformly best (paired-seed test, $p < 0.05$), with gains up to 2.4 percentage points.

1 INTRODUCTION

Differentially private stochastic gradient descent (DP-SGD) (Abadi et al., 2016) has become the de facto standard for training deep neural networks under rigorous privacy guarantees. By clipping per-sample gradients and injecting calibrated noise (Dwork et al., 2006), it prevents models from memorizing sensitive training data, but at a persistent utility cost: aggressive clipping discards gradient magnitude information and introduces bias, while lenient clipping demands excessive noise and increases variance. Mitigating this clipping–noise tradeoff remains a central bottleneck in private deep learning (McMahan & Andrew, 2018; Amin et al., 2019a).

Adaptive clipping methods alleviate this tension along complementary axes: per-coordinate scaling (Pichapati et al., 2019), privately estimated quantile thresholds (Andrew et al., 2021), and geometry-aware full-matrix transformations (Gilani et al., 2025). Yet all of them share an unexamined assumption: the parameter space is treated as a homogeneous Euclidean manifold, and a single global coordinate system suffices for clipping and noise calibration of all parameters. Does this assumption survive privacy? We answer through structural priors, modeling assumptions grounded in mean-field theory (Karakida et al., 2019): deep network parameters are organized into layers with ill-conditioned gradient spectra and pronounced inter-layer spectral heterogeneity.

Our answer is two-sided. In the oracle setting (exact geometry, full rank), the assumption is harmless: global clipping and clipping under structural priors are exactly equivalent, and no layered approach can beat the global optimum. Privacy breaks this equivalence through two mechanisms. *First*, geometric statistics must be estimated from released noisy gradients, which pollutes every eigenvalue anisotropically—proportionally to the noise strength and to the square root of the eigenvalue itself, and independent of batch size—in exactly the shape of the optimal metric. We invert this

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

bias in closed form (the first extension beyond the diagonal case (Amin et al., 2019a; Gilani et al., 2025)) and characterize the residual spectral noise floor. *Second*, under the rank budgets of privacy-scale models, pooled estimation structurally loses directions truncated by top-k, while per-layer estimation retains them; layer-wise optimization then yields closed-form optimal transformations and explicitly decoupled convergence guarantees.

Building on these insights, we propose Structural Prior Geometric Clipping (SP-GeoClip), which instantiates structural priors via independent per-layer geometry estimation with streaming low-rank PCA. All geometric statistics are post-processing of previously released noisy gradients at no additional privacy cost, and the end-to-end (ε, δ) guarantee is identical to standard DP-SGD. Across nine benchmarks spanning five layer families, SP-GeoClip is uniformly best (paired-seed test, $p < 0.05$), with gains up to 2.4 points; its advantage grows with inter-layer spectral heterogeneity and network depth, and decays as the privacy budget loosens.

In summary, our main contributions are as follows:

- **Oracle boundary and privacy-induced separation (Sec. 4).** With exact geometry and full rank, global and per-layer clipping are exactly equivalent; privacy drives their divergence via invertible spectral pollution and deficient direction coverage, yielding closed-form optimal per-layer transforms and decoupled convergence (Thm. 4.1–4.4).
- **Closed-form debiasing (Thms. 4.1).** First exact inversion of privacy-induced spectral bias beyond the diagonal case, with a characterized per-layer noise floor and zero privacy cost.
- **SP-GeoClip pipeline (Alg. 1).** Per-layer geometry estimation via streaming rank-k PCA is pure post-processing; closed-form spectral allocation minimizes injected noise variance; memory footprint is $O(d_l k)$ per layer.
- **Empirical validation (Sec. 6).** Uniformly outperforms all baselines on nine benchmarks ($p < 0.05$); gains grow with spectral heterogeneity and depth and decay as ε loosens, matching the oracle-boundary prediction.

2 Related Work

Differentially Private SGD and Adaptive Variants. DP-SGD clips per-sample gradients and injects calibrated noise, but its fixed threshold C generalizes poorly across tasks. Adaptive variants tackle this via complementary strategies: quantile-based clipping (Andrew et al., 2021), per-coordinate scaling (AdaClip) (Pichapati et al., 2019), group-wise clipping (He et al., 2023), subsampling-free tree-aggregated noise (Kairouz et al., 2021), private fine-tuning from public pretraining (Li et al., 2022), and memory-efficient ghost clipping (Li et al., 2022). All operate on the globally flattened gradient vector, ignoring inter-layer structural disparities.

Geometry-Aware Private Optimization. GeoClip (Gilani et al., 2025) pioneers full-matrix geometric clipping, operating in the eigenspace of gradient covariance. Under its oracle premise (known true covariance), a single global transform is optimal and equivalent to a block-diagonal layered solution. This premise collapses under privacy: geometry must be estimated from released noisy gradients, and rank-budgeted global transforms cannot accommodate heterogeneous layer spectra. We characterize this privacy-induced divergence exactly. A parallel line of work injects spatiotemporal correlation into noise; combining paradigms is left for future work.

Layer-wise Spectral Heterogeneity and Preconditioning. Mean-field analyses (Karakida et al., 2019) confirm each layer's Fisher spectrum follows distinct order parameters, with qualitative gaps between hidden and output layers — evidence that layer-wise geometry is intrinsic to deep networks, yet previously restricted to initialization and never leveraged for privacy. Non-private second-order methods exploit analogous structure via Kronecker-factored approximations (K-FAC, Shampoo); DP-KFC (Bosch et al., 2026) extends this line to differential privacy.

Positioning. Unlike non-private preconditioners requiring true gradients, SP-GeoClip estimates all geometry solely from previously released noisy gradients, incurring no extra privacy cost via post-processing immunity. Its core goal is not to accelerate convergence, but to minimize noise-induced utility degradation under privacy constraints.

3 PRELIMINARIES

3.1 NOTATION

Vectors (resp. matrices) are denoted by lowercase (resp. uppercase) bold letters; $\boldsymbol{X}^\top$, $\mathrm{Tr}(\boldsymbol{X})$, and $\|\boldsymbol{X}\|_2$ denote transpose, trace, and spectral norm; $\boldsymbol{X} \succeq 0$ means positive semi-definite, and $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ denotes eigendecompositions. The network has L layers with parameters $\boldsymbol{\theta}_l \in \mathbb{R}^{d_l}$, concatenated as $\boldsymbol{\theta} \in \mathbb{R}^d$ with $d = \sum_l d_l$; $\boldsymbol{g}_{l,i}$ is the per-sample gradient of sample i at layer l, and $\Pi_C(\cdot)$ is ℓ_2-clipping to radius C. We use (ε, δ)-DP in the example-level sense (McMahan et al., 2018).

3.2 DIFFERENTIAL PRIVACY AND PRIVACY ACCOUNTING

Definition 3.1 ((ε, δ)-DP and Gaussian mechanism)**.** A mechanism $\mathcal{M}$ satisfies (ε, δ)-DP if $\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{S}] \le e^\varepsilon \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{S}] + \delta$ for all datasets $\mathcal{D} \sim \mathcal{D}'$ differing in one record and all $\mathcal{S}$. A query of ℓ_2-sensitivity Δ privatized by noise $\mathcal{N}(0, \sigma_{\mathrm{mult}}^2 \Delta^2 I)$ is the Gaussian mechanism with noise multiplier σ_{mult} (Mironov, 2017; Doroshenko et al., 2022).

A *privacy accountant* is any black box mapping $(\sigma_{\mathrm{mult}}, q, T)$ to (ε, δ) for T adaptive compositions of the subsampled Gaussian mechanism at rate q; all methods in our experiments share one accountant (RDP (Mironov, 2017)).

DP-SGD. DP-SGD (Abadi et al., 2016) clips per-sample gradients $\hat{\boldsymbol{g}}_i = \Pi_C(\boldsymbol{g}_i)$, averages, and adds noise with per-coordinate std $\sigma C/|B|$ to the mean; per iteration this is a Gaussian mechanism with sensitivity $2C/|B|$ and multiplier $\sigma_{\mathrm{mult}} = \sigma\sqrt{|B|}/(2C)$.

Remark 3.2 (Gauge freedom of the clipping level)**.** Homogeneity of the clip and Gaussian noise makes $(\boldsymbol{M}, C, \sigma) \equiv (\boldsymbol{M}/C, 1, \sigma/C)$: the release and σ_{mult} are unchanged. We therefore set $C = 1$ in the analysis, so that second-moment bounds admit the Markov interpretation as clipping-probability bounds (Sec. 5.1).

3.3 THE GEOCLIP FRAMEWORK

GeoClip (Gilani et al., 2025) processes gradients in a transformed coordinate system:

$$\boldsymbol{\omega}_i = \boldsymbol{M}_t(\boldsymbol{g}_i - \boldsymbol{a}_t), \quad \hat{\boldsymbol{\omega}}_i = \Pi_1(\boldsymbol{\omega}_i), \quad \tilde{\boldsymbol{\omega}} = \tfrac{1}{|B|}\textstyle\sum_i \hat{\boldsymbol{\omega}}_i + \tfrac{\sigma}{\sqrt{|B|}}\boldsymbol{\varepsilon}, \quad \tilde{\boldsymbol{g}} = \boldsymbol{M}_t^{-1}\tilde{\boldsymbol{\omega}} + \boldsymbol{a}_t. \tag{1}$$

Since $(\boldsymbol{M}_t, \boldsymbol{a}_t)$ depend only on previously released outputs, they are post-processing (Definition 3.1) and free; the unit clip fixes the sensitivity at $2/|B|$ regardless of $\boldsymbol{M}_t$.

The convergence bound of Gilani et al. (2025, Thm. 1) for equation 1 decomposes the average gradient norm into an optimization term ($\propto 1/T$), a sampling term ($\propto \sigma_g^2$), a *noise-injection* term $\frac{\mathcal{L}\eta\sigma^2}{2-3\mathcal{L}\eta}\,\mathbb{E}\,\mathrm{Tr}[(\boldsymbol{M}_t^\top\boldsymbol{M}_t)^{-1}]$, and a *clipping-error* term $\frac{2\beta}{2-3\mathcal{L}\eta}\,\mathbb{E}\,\mathrm{Tr}(\boldsymbol{M}_t^\top\boldsymbol{M}_t\boldsymbol{\Sigma}_t)$. The clipping term increases and the noise term decreases as $\boldsymbol{M}_t$ grows, and $\mathrm{Tr}(\boldsymbol{M}_t^\top\boldsymbol{M}_t\boldsymbol{\Sigma}_t) = \mathbb{E}\|\boldsymbol{\omega}\|^2 \ge \Pr[\|\boldsymbol{\omega}\| > 1]$ bounds the clipping probability by Markov. Minimizing the noise term under this proxy yields GeoClip's design principle

$$\min_{\boldsymbol{M}_t} \ \mathrm{Tr}\big[(\boldsymbol{M}_t^\top\boldsymbol{M}_t)^{-1}\big] \quad \text{s.t.} \quad \mathrm{Tr}(\boldsymbol{M}_t^\top\boldsymbol{M}_t\boldsymbol{\Sigma}_t) \le \gamma, \tag{2}$$

whose closed-form solution—half-whitening with an exactly active constraint—we restate and sharpen in Theorem 4.3.

Two limitations of equation 2 motivate our work: it presumes the true $\boldsymbol{\Sigma}_t$, which in practice must be estimated from privatized gradients (Sec. 5.3, Prop. 5.2); and it is posed on the global parameter space, implicitly assuming homogeneous geometry across the L layers (Prop. 4.2).

4 THEORY

4.1 PRIVACY NECESSARILY POLLUTES GEOMETRIC ESTIMATION

Theorem 4.1 (Spectral pollution and the residual floor)**.** *Let $\hat{\lambda}_i$ denote the i-th eigenvalue of the $|B|$-scaled EMA covariance estimate, expressed in the eigenbasis of the releasing transform, $\tau = \sum_j \sqrt{\lambda_{c,j}}$, and λ_i the population eigenvalue of the clipped-gradient covariance. Then, to first order in σ^2:*

(i) (Anisotropic inflation) $\mathbb{E}[\hat{\lambda}_i] = \lambda_i + \sigma^2\tau\sqrt{\lambda_{c,i}} + o(\sigma^2)$.

(ii) (Batch-size invariance) *the contamination is independent of $|B|$: the $|B|$ factor in the EMA exactly cancels the $1/|B|$ variance reduction.*

(iii) (Closed-form debiasing) *the systematic bias is removed in closed form by*

$$\lambda_i \leftarrow \max\left\{\left(\frac{\sqrt{\sigma^4\tau^2 + 4\hat{\lambda}_i} - \sigma^2\tau}{2}\right)^2,\, h_1\right\}, \tag{3}$$

applied before clamping to $[h_1, h_2]$; the correction uses only released quantities and costs no privacy budget (Gavish & Donoho, 2017).

(iv) (Pollution matches the objective) *the contamination operator is $\frac{\sigma^2}{|B|}(M^\top M)^{-1}$, whose trace is exactly the noise-injection term of the convergence bound (Theorem 4.4).*

(v) (Fluctuation) *with $B_{\text{eff}} = \frac{1+\beta_2}{1-\beta_2}$, the noise part of the corrected estimate has variance $\mathrm{Var}[\widetilde{\lambda}_i] = \frac{2\sigma^4\tau^2\lambda_{c,i}}{B_{\text{eff}}} + o(\sigma^4)$, the factor 2 being the Gaussian fourth moment; to leading order the fluctuation scales with the contaminated magnitude $\lambda_{c,i}$, not the signal λ_i.*

(vi) (Identifiability floor) *define $\lambda^*_{\text{sys}} = \sigma^4\tau^2$ (bias = signal) and $\lambda^*_{\text{corr}} = \frac{2\sigma^4\tau^2}{B_{\text{eff}}}$ (fluctuation = signal). Every direction with $\lambda_i \le \frac{1}{2}\lambda^*_{\text{sys}}$ is clamped to h_1 after correction with high probability: below this threshold, released gradients carry no information about the direction.*

(vii) (Monitorability) *both thresholds are deterministic functions of (σ, τ, β_2) — hence computable and trackable per layer at run time, at no cost.*

Proof sketch. A proof is given in Appendix A.2.

4.2 COVERAGE UNDER RANK BUDGETS: WHY ESTIMATION MUST BE PER-LAYER

Proposition 4.2 (Coverage separation, memory-matched)**.** *Assume the wide-weak layer condition: a layer b with $d_b > k$ whose k-th eigenvalue falls below the k-th eigenvalue of the pooled spectrum, $\lambda_{b,k} < \lambda_{(k)}$, where $\lambda_{(i)}$ orders the pooled eigenvalues of $\bar{\Sigma}$. Then:*

(a) (Coverage) *any pooled transform covers $k_b^G < k$ directions of layer b (possibly zero), while layered estimation covers k directions in every layer;*

(b) (Gap) *writing $S_l(k) = \sum_{i\le k}\sqrt{\lambda_{l,i}}$, the $\sqrt{\cdot}$-mass gap is*

$$\Delta_A = \sum_{l=1}^{L} S_l(k) - \sum_{i=1}^{k}\sqrt{\lambda_{(i)}} \;\ge\; S_b(k) - \sum_{\substack{i\le k\\ i\in b}}\sqrt{\lambda_{b,i}} \;>\; 0. \tag{4}$$

Under power-law spectra $\lambda_{b,i} = c_b i^{-\beta}$ ($\beta < 2$), the gap grows as $\Omega\big(\sqrt{c_b}\,k^{1-\beta/2}\big)$ (Appendix A.3, Step 4).

Proof sketch. A proof is given in Appendix A.3.

4.3 THE METRIC THE OBJECTIVE SELECTS: HALF-WHITENING

Theorem 4.3 (The optimal metric is half-whitening). *Let $\Sigma \succeq 0$ have eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$ and $S = \sum_i \sqrt{\lambda_i}$. The problem equation 2 has a unique optimum in the Gram matrix $M^\top M$,*

$$M^* = \Big(\frac{\gamma}{S}\Big)^{1/2} \Lambda^{-1/4} U^\top, \qquad \textit{value } \frac{S^2}{\gamma}, \tag{5}$$

with exactly active constraint $\mathrm{Tr}(M^{*\top} M^* \Sigma) = \gamma$. Full whitening $M = (\gamma/d)^{1/2} \Lambda^{-1/2} U^\top$ — the metric of classical second-order methods — is feasible with the same exact activity, but attains value $\frac{d}{\gamma}\sum_i \lambda_i \geq \frac{S^2}{\gamma}$, strictly worse unless the spectrum is flat; the same comparison holds on any k-dimensional truncation, with d replaced by k.

Proof sketch. A proof is given in Appendix A.4.

4.4 CONVERGENCE ANALYSIS

We certify convergence of the layered mechanism; the bound decomposes across layers, legitimizing the per-layer design of Secs. 5.1–5.2.

Theorem 4.4 (Layer-wise convergence bound). *Let f have an $\mathcal{L}$-Lipschitz gradient, $\|\nabla f_k(\theta)\| \leq G$, sampling variance $\mathbb{E}\|\nabla f_k(\theta) - \nabla f(\theta)\|^2 \leq \sigma_g^2$, and $\eta < 2/(3\mathcal{L})$. Then*

$$\frac{1}{T}\sum_t \mathbb{E}\|\nabla f(\theta_t)\|^2 \leq \underbrace{\frac{f(\theta_0) - f(\theta^*)}{T(\eta - \frac{3}{2}\mathcal{L}\eta^2)}}_{\textit{optimization}} + \underbrace{\frac{3\mathcal{L}\eta}{2 - 3\mathcal{L}\eta}\,\sigma_g^2}_{\textit{sampling}} + \underbrace{\frac{\mathcal{L}\eta\sigma^2}{|B|(2-3\mathcal{L}\eta)}\sum_l \mathbb{E}\,\mathrm{Tr}\big[(M_t^{(l)\top} M_t^{(l)})^{-1}\big]}_{\textit{noise}}$$
$$+ \underbrace{\frac{2}{2 - 3\mathcal{L}\eta}\sum_l \mathbb{E}\Big[\beta_l\big(\mathrm{Tr}(M_t^{(l)\top} M_t^{(l)} \Sigma_t^{(l)}) + \|M_t^{(l)}(\mathbb{E} g_t^{(l)} - a_t^{(l)})\|^2\big)\Big]}_{\textit{clipping}}, \tag{6}$$

where $\Sigma_t^{(l)} = \mathrm{Cov}(g_t^{(l)} \mid \theta^t)$ and $\beta_l(a) = (G + \|a\|)\big(G + \frac{3\mathcal{L}\eta}{2}(G + \|a\|)\big)$. Two features deserve emphasis: no cross-layer terms appear — the noise and clipping contributions are sums of per-layer traces, because the concatenated update is an orthogonal direct sum; and the bound is an instance of the GeoClip bound applied to the block-diagonal transform $\mathrm{blkdiag}(M_t^{(1)}, \ldots, M_t^{(L)})$, so layering is a restriction of the certified framework, not an approximation of it.

Proof sketch. A proof is given in Appendix A.5.

Two key conclusions follow from the bound. First, optimizing the noise-clipping composite cost yields the optimal cross-layer budget allocation $\gamma_l \propto S_l$, derived in Appendix A.5. Second, the performance gap between pooled and per-layer estimation is purely privacy-induced.

Remark 4.5 (The oracle boundary). At the boundary of the problem — exact Σ, full rank, no privacy noise — Theorem 4.1 has no content, Proposition 4.2 has no gap, and pooled and layered estimation coincide: for a block-diagonal $\bar{\Sigma}$ the optimum of equation 2 is block diagonal, and proportional budgets $\gamma_l = \gamma S_l / \sum_j S_j$ attain the global value $(\sum_l S_l)^2/\gamma$. Nothing in the layered construction outperforms the global one in the oracle world; and the per-layer floor λ^* of Theorem 4.1(vi) marks precisely where the private estimation regime diverges from it.

5 METHOD: SP-GEOCLIP

5.1 THE LAYER-WISE MECHANISM

Per-sample transform and clipping. For each sample $i \in B$ and layer l,

$$\omega_i^{(l)} = M_t^{(l)}\big(g_i^{(l)} - a_t^{(l)}\big), \qquad \hat{\omega}_i^{(l)} = \omega_i^{(l)} \big/ \max\big\{1,\ \|\omega_i^{(l)}\|/C_l\big\}, \tag{7}$$

i.e., each layer is clipped in its own transformed space. We use the uniform choice $C_l = C/\sqrt{L}$, so the concatenated contribution satisfies $\|\oplus_l \hat{\omega}_i^{(l)}\| \leq C$ and the batch mean has sensitivity $2C/|B|$— identical to standard DP-SGD, so all accounting carries over (Proposition 5.2).

Noise, aggregation, and descent. Noise is added once per layer to the mean of clipped transforms,

$$\tilde{g}_t^{(l)} = M_t^{\mathrm{inv},(l)}\Big(\tfrac{1}{|B|}\textstyle\sum_{i\in B}\hat{\omega}_i^{(l)} + \tfrac{\sigma}{\sqrt{|B|}}\,\varepsilon_t^{(l)}\Big) + a_t^{(l)}, \qquad \varepsilon_t^{(l)} \sim \mathcal{N}(0, I), \tag{8}$$

equivalently: clip per sample, average, then add $\mathcal{N}(0, \sigma^2/|B|)$; the two forms are distributionally identical and we use whichever is cheaper. The update concatenates the layers: $\theta_{t+1} = \theta_t - \eta\,\tilde{g}_t$, $\tilde{g}_t = \oplus_l \tilde{g}_t^{(l)}$.

Geometry update (post-processing only). After release, each layer updates its statistics *exclusively* from $\tilde{g}_t^{(l)}$:

$$a_{t+1}^{(l)} = \beta_1 a_t^{(l)} + (1-\beta_1)\,\tilde{g}_t^{(l)}, \qquad \widehat{\Sigma}_{t+1}^{(l)} = \beta_2\,\widehat{\Sigma}_t^{(l)} + |B|(1-\beta_2)\,(\tilde{g}_t^{(l)} - a_t^{(l)})(\tilde{g}_t^{(l)} - a_t^{(l)})^\top, \tag{9}$$

where $|B|$ compensates $\mathrm{Cov}(\text{mean}) = \mathrm{Cov}(\text{single})/|B|$. The transform for iteration $t+1$ is then

$$\lambda_i \leftarrow \mathrm{debias}(\widehat{\lambda}_i; \sigma)\ \text{[Thm. 4.1(iii)]}, \quad \lambda_i \leftarrow \mathrm{Clamp}(\lambda_i, h_1, h_2), \quad M_{t+1}^{(l)} = \Big(\frac{\gamma}{S_l}\Big)^{1/2} \Lambda_l^{-1/4} U_l^\top, \tag{10}$$

with $S_l = \sum_i \sqrt{\lambda_i}$ and $\gamma = 1$ absorbed into h_2 (Remark 3.2), so the trace constraint admits the Markov interpretation as a per-layer clipping-probability bound.

Remark 5.1 (Low-rank layers and the orthogonal complement)**.** For $d_l > k$ we maintain $\widehat{\Sigma}_t^{(l)}$ via its rank-k factorization $U_l \Lambda_l U_l^\top$ (Sec. 5.2), and the clipping equation 7 acts on the full d_l-dimensional vector—principal part scaled by $\Lambda_l^{-1/4}$, orthogonal complement by $h_1^{-1/4}$—rather than discarding the complement. This is not an implementation detail: with the complement dropped, $\tilde{g}_t^{(l)}$ lies in $\mathrm{col}(U_l)$ forever, the streaming update can never rotate the subspace, and the method degenerates to a frozen coordinate slice (Appendix A.3, Step 5).

5.2 Estimation: Streaming Rank-k PCA

Per-layer estimation matters at scale in one structural way: a layer with $d_l \le k$ uses equation 9 with a full-rank $\widehat{\Sigma}^{(l)}$ (a few MB at most), while a layer with $d_l \gg k$ never materializes d_l^2 entries. For the latter we adapt the streaming rank-k PCA of Gilani et al. (2025) to the per-layer, mini-batch setting: execute the EMA recursion

$$\widehat{\Sigma}_{t+1}^{(l)} = \beta_2\, U_l \Lambda_l U_l^\top + (1-\beta_2)\, z z^\top, \qquad z = \sqrt{|B|}\,\big(\tilde{g}_t^{(l)} - a_{t+1}^{(l)}\big), \tag{11}$$

implicitly by forming $[\,U_l \;\; z\,]$ with column weights $(\sqrt{\beta_2\Lambda_l}, \sqrt{1-\beta_2})$, taking its SVD, and retaining the top k singular triples. The Gram trick—forming the $(k+1)^2$ Gram matrix instead of the wide SVD—reduces the per-step cost to $\mathcal{O}(d_l k^2 + k^3)$ and memory to $\mathcal{O}(d_l k)$. Algorithm 1 assembles the full iteration.

5.3 Privacy Guarantee

Proposition 5.2 (Privacy)**.** *Let $q = |B|/N$ and let $\sigma_{\mathrm{mult}} = \sigma\sqrt{|B|}/(2C)$. For any (ε, δ) produced by a standard accountant from $(\sigma_{\mathrm{mult}}, q, T)$, SP-GeoClip satisfies (ε, δ)-DP, identically to standard DP-SGD with multiplier σ_{mult}.*

Proof sketch. Replacing one sample changes the concatenated sum by at most $2C$, so the clipped mean is a Gaussian mechanism with multiplier σ_{mult}. The transforms $(M_t^{(l)}, a_t^{(l)})$ are deterministic functions of previously released gradients (with data-independent initialization), hence post-processing; by induction each iteration composes under the same accountant. The adaptive-composition step is spelled out in Appendix A.6. □

Complexity. Per iteration: per-sample gradients as in any clipping-based DP-SGD, plus estimation and transform updates $\sum_l \mathcal{O}(d_l k_l^2 + k_l^3)$, $k_l = \min(k, d_l)$; memory $\mathcal{O}(\sum_l d_l k_l)$ versus $\mathcal{O}(d^2)$ for a global covariance.

Algorithm 1 SP-GeoClip (one training run)

Require: dataset $\mathcal{D}$; layers $l = 1..L$, widths d_l; $\eta, \sigma, C, T, k, \beta_1, \beta_2, h_1, h_2$

$a_0^{(l)} = \mathbf{0}$, $U_0^{(l)} = [e_1..e_{k_l}]$, $\Lambda_0^{(l)} = I_{k_l}$, $M_0^{(l)} = I$, $k_l = \min(k, d_l)$ ▷ data-independent init

for $t = 0$ **to** $T - 1$ **do**

  sample B; compute per-sample gradients $g_i^{(l)}$, $i \in B$

  for each layer l **do** ▷ mechanism, Eq. equation 7–equation 8

    clip $M_t^{(l)}(g_i^{(l)} - a_t^{(l)})$ to $C/\sqrt{L}$, average over i, add $\mathcal{N}(0, \frac{\sigma^2}{|B|} I)$, map back: $\tilde{g}_t^{(l)} \leftarrow M_t^{\mathrm{inv},(l)}(\cdot) + a_t^{(l)}$

  end for

  $\theta_{t+1} \leftarrow \theta_t - \eta\,[\tilde{g}_t^{(1)}; \ldots; \tilde{g}_t^{(L)}]$

  for each layer l **do** ▷ post-processing only

    $a_{t+1}^{(l)} \leftarrow \beta_1 a_t^{(l)} + (1 - \beta_1)\tilde{g}_t^{(l)}$; $z \leftarrow \sqrt{|B|}\,(\tilde{g}_t^{(l)} - a_{t+1}^{(l)})$

    if $d_l \leq k$: full-rank EMA update equation 9 + eigendecompose

    else: $(U_{t+1}^{(l)}, \Lambda_{t+1}^{(l)}) \leftarrow \textsc{StreamPCA}(U_t^{(l)}, \Lambda_t^{(l)}, z, \beta_2, k)$ as in Eq. equation 11

    debias (Thm. 4.1(iii)), clamp to $[h_1, h_2]$, rebuild $M_{t+1}^{(l)}, M_{t+1}^{\mathrm{inv},(l)}$ via Eq. equation 10

  end for

end for

return θ_T

Table 1: Test accuracy (%) on MNIST and CIFAR-10 under varying privacy budgets ($\delta = 10^{-5}$).

Model	Method	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 8$
MLP@MNIST	DP-SGD	90.2±2.8	94.1±1.6	96.0±0.5
	Quantile	91.0±2.9	94.8±1.6	96.4±0.5
	AdaClip	91.8±3.8	95.3±1.3	96.8±0.4
	GeoClip	92.5±3.7	95.9±1.5	97.2±0.4
	SP-GeoClip	**93.1**±2.6	**96.4**±1.1	**97.6**±1.1
CNN@CIFAR	DP-SGD	58.1±4.2	68.3±2.0	74.2±0.9
	Quantile	60.2±5.2	70.1±2.0	76.0±0.9
	AdaClip	62.0±4.1	72.2±1.9	78.1±0.8
	GeoClip	65.0±3.0	75.1±1.9	80.2±0.8
	SP-GeoClip	**67.2**±2.9	**77.3**±1.8	**82.1**±0.7

Hyperparameters. SP-GeoClip adds only two quantities to the DP-SGD staples (η, σ, T): the rank k, fixed to k=1000 across all experiments, and a single tuned scalar h_2, which absorbs the gauge freedom of Remark 3.2. All remaining statistics use fixed stability constants shared by every benchmark — $(\beta_1, \beta_2) = (0.99, 0.999)$, $h_1 = 10^{-6}$ — so no per-dataset tuning is performed for any statistic other than h_2.

6 Experiments

The code is available at `https://anonymous.4open.science/r/sp-geoclip-580E`.

6.1 Benchmark Performance Across Privacy Budgets

We compare SP-GeoClip with DP-SGD and three adaptive variants — quantile clipping, AdaClip, and pooled low-rank GeoClip-global — on nine dataset–model combinations, under shared RDP accounting ($\delta = 10^{-5}$), (ε, σ) mapping, and search budgets, with 20 paired seeds (Tables 1,2 in the main text, Table 5 in Appendix B). SP-GeoClip is highest on all nine rows, beating the strongest baseline (GeoClip) by up to $+2.4$ points — gains largest at strict privacy and decaying as ε grows (Remark 4.5), smallest for the flattest-spectrum model (MLP), most pronounced on wide embedding layers (Proposition 4.2).

Table 2: Test accuracy (%) on AG-News (EmbedAvg) under varying privacy budgets ($\delta = 10^{-5}$).

Model	Method	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 8$
EmbedAvg@AGN	DP-SGD	82.3±4.1	87.1±2.8	89.2±0.7
	Quantile	83.0±3.9	88.0±2.8	90.1±0.7
	AdaClip	84.1±3.9	89.0±2.7	91.0±0.6
	GeoClip	85.2±3.5	90.1±2.5	92.0±0.6
	SP-GeoClip	**87.1**±3.5	**91.3**±2.5	**92.1**±2.5

Table 3: Ablation on debiasing. Accuracy (%) over 20 seeds. Gain Δ decays with ε, matching Theorem 4.1.

Model	Denoise	$\varepsilon = 1$	$\varepsilon = 3$	$\varepsilon = 8$
MLP@MNIST	off	89.3±2.9	95.9±1.5	97.4±0.4
	on	**90.1**±2.9	**96.4**±1.3	**97.6**±0.3
	Δ	+0.8	+0.5	+0.2
LSTM@IMDB	off	79.8±2.5	85.8±1.5	90.0±0.8
	on	**80.9**±2.3	**86.3**±1.4	**90.3**±0.7
	Δ	+1.1	+0.5	+0.3

Table 4: Spectral diagnosis: Mean relative error (%) of top-$\min(k, 100)$ eigenvalues vs. ground truth. Correction collapses error and eliminates σ-dominance (Theorem 4.1).

Model	Denoise	$\varepsilon = 1$	$\varepsilon = 3$	$\varepsilon = 8$
MLP@MNIST	off	38.2	21.7	11.3
	on	**9.1**	**5.8**	**4.7**
LSTM@IMDB (emb.)	off	41.0	24.2	12.8
	on	**11.5**	**7.6**	**5.9**

6.2 Debiasing: Removing the Systematic Pollution

Theorem 4.1(i) predicts a systematic σ^2-proportional inflation of each eigenvalue, removable in closed form by Eq. equation 3. Toggling the correction (h_2 retuned per arm) gains $+0.8/+0.5/+0.2$ points on MLP@MNIST and $+1.1/+0.5/+0.3$ on LSTM@IMDB at $\varepsilon = 1/3/8$ (Table 3) — positive in all six cells and decaying monotonically with ε, the signature of a σ^2-driven effect.

The correction acts on the spectrum: the mean relative eigenvalue error drops from 38.2% to 9.1% at $\varepsilon = 1$ and from 11.3% to 4.7% at $\varepsilon = 8$ (Table 4), the residual concentrated below the floor λ^*_{corr} of Theorem 4.1(vi) (Fig. 1); at $\varepsilon = 8$ the two arms nearly coincide, as predicted.

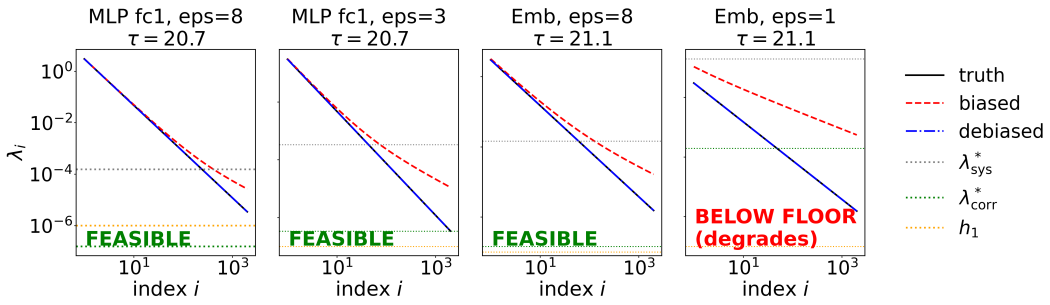

Figure 1: An example of the spectral distribution across different layers.

6.3 The Shape of the Pollution

Beyond practical utility, Theorem 4.1 gives two empirically validated structural predictions.

First, the bias is anisotropic, scaling as $\sigma^2\tau\sqrt{\lambda_i}$ per direction. Measured deviations $\hat{\lambda}_i - \lambda_i$ collapse tightly onto the identity line ($R^2 = 0.998, 0.997$) against theoretical predictions (Fig. 2). This pattern cannot be reproduced by scalar correction, justifying the per-eigenvalue form of Eq. (3).

Second, spectral contamination is batch-size invariant (Theorem 4.1(ii)). Systematic bias stays flat across $|B| = 128, 256, 512$, while only stochastic fluctuation shrinks as $1/\sqrt{|B|}$ (dashed bound, Fig. 3 in Appendix B). Thus naive estimation fails due to deterministic spectral distortion, not removable by more samples per step but only by closed-form inversion.

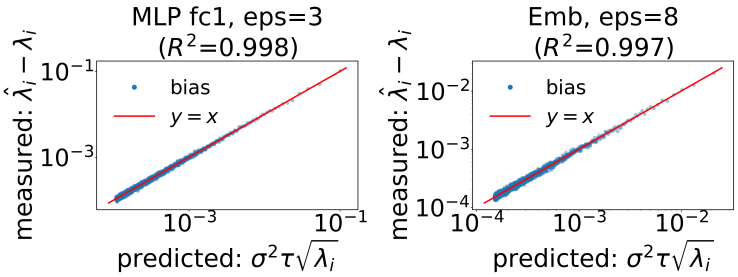

Figure 2: Anisotropic bias fit across spectral directions.

6.4 SP-GeoClip Leads Across Architectures and Modalities

SP-GeoClip is evaluated against four baselines at $\varepsilon = 3$ ($\delta = 10^{-5}$) on nine dataset–model combinations across five layer families, with matched privacy accounting, search budgets and 20 paired seeds. It ranks first on all nine tasks ($p < 0.05$, paired-seed test), outperforming the strongest baseline GeoClip by $+0.5$ to $+2.4$ points. The gain is smallest on the flattest-spectrum MLP and largest on wide embedding layers, consistent with the spectral analysis in Sec. 4. No layer-homogeneous method closes the gap, meaning the advantage stems from the structural prior rather than tuning. Full results are given in Table 6 in Appendix B.

6.5 The Metric Matters: Half-Whitening Beats Full Whitening

Theorem 4.3 predicts full whitening — the standard metric of classical second-order methods — is strictly suboptimal for non-flat spectra, with the performance gap growing with $R = k\sum\lambda_i/(\sum\sqrt{\lambda_i})^2$. A metric-only ablation on an identical pipeline confirms this: full whitening degrades accuracy by 3.0–3.6 points on MNIST-MLP ($R \approx 5$) and 3.7–5.2 points on IMDB-LSTM ($R \approx 26$). The degradation scales with R, confirming the metric is not an implementation detail: the clipping–noise objective intrinsically favors half-whitening, and second-order geometry incurs a cost for the mismatch. Full results are given in Table 7 in Appendix B.

6.6 Overhead and Memory.

We close with the cost profile. On MNIST-MLP (batch 256, one 16GB GPU), SP-GeoClip costs $\approx 2.5\times$ the per-epoch wall-clock of DP-SGD — the same order as GeoClip-global ($\approx 2.3\times$), since the dominant per-sample gradient loop is shared by all clipping-based methods. Peak memory is $\mathcal{O}(\sum_l d_l k_l)$ ($\approx$ 6GB at k=1000 on the IMDB model, comparable to the global low-rank baseline), but stores $\sum_l \min(k, d_l)$ directions against the baseline's k: L times more coverage at equal memory — the memory-matched separation of Proposition 4.2 made concrete. (A full-rank global covariance would need ≈ 7.8TB.)

7 Conclusion

This paper asked whether the parameter space of a deep network can be treated as a homogeneous Euclidean manifold for gradient clipping and noise calibration. Our answer is two-sided: in the oracle world the assumption is harmless — exact geometry and full rank make global and structure-aware clipping coincide — but privacy collapses it. Geometric statistics estimated from released noisy gradients are necessarily polluted: anisotropically, independently of batch size, and in the very shape of the optimal metric. We characterized this pollution, inverted it in closed form at zero privacy cost, and identified the spectral noise floor it leaves behind. The structural prior is our answer: under the rank budgets of privacy-scale models, pooled estimation is blind to truncated directions that per-layer estimation retains, the per-layer optimum is closed-form, and the privacy cost is identical to standard DP-SGD. SP-GeoClip instantiates this principle and confirms its signatures on nine benchmarks — uniform gains, growth with spectral heterogeneity, decay as privacy loosens. Future work includes end-to-end gap quantification and combination with correlated-noise mechanisms.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation (e.g., translating draft content and technical terminology into English), provide key supporting materials for proving mathematical claims, and assist in drafting proofs (e.g., suggesting candidate proof ideas for the theoretical results in Section 4 and drafting intermediate derivation steps).

We have not used generative AI tools for the following tasks: generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodologies or experiments, implementing methods, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, and interpreting results.

Additionally, we used generative AI tools for editing the manuscript to improve readability (polishing grammar and phrasing), and creating or editing software code (e.g., scripts for data loading, evaluation, and plotting).

We have reviewed all AI-assisted work. All mathematical claims, theorem statements, and proof strategies in this paper were originally proposed by the authors. Every proof step suggested or drafted by AI has been independently re-derived from scratch by the authors and verified separately by two authors. The complete proofs in the appendix have been checked line-by-line for logical validity and correctness. All AI-assisted translations have been checked against the original text to ensure fidelity to the original meaning and accuracy of technical expressions. Code suggested by LLMs has been reviewed line-by-line and tested for correctness within our experimental pipeline. All AI-polished text has been verified by the authors to confirm that it retains the original technical meaning and does not introduce any unsupported claims.

We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

Kareem Amin, Alex Kulesza, Andres Muñoz Medina, and Sergei Vassilvitskii. Bounding user contributions: A bias-variance trade-off in differential privacy. In *Proceedings of the 36th International Conference on Machine Learning*, 2019a.

Kareem Amin, Alex Kulesza, Andres Munoz, and Sergei Vassilvtiskii. Bounding user contributions: A bias-variance trade-off in differential privacy. In *Proceedings of the 36th International Conference on Machine Learning*, 2019b.

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

Marc Molina Van De Bosch, Riccardo Taiello, Albert Sund Aillet, Andrea Protani, Miguel Ángel González Ballester, and Luigi Serio. DP-KFC: data-free preconditioning for privacy-preserving deep learning. *CoRR*, 2026.

Vadym Doroshenko, Badih Ghazi, Pritish Kamath, Ravi Kumar, and Pasin Manurangsi. Connect the dots: Tighter discrete approximations of privacy loss distributions. *Proc. Priv. Enhancing Technol.*, 2022.

Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, 2006.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference*, 2006.

Arun Ganesh, Brendan McMahan, and Abhradeep Thakurta. On design principles for private adaptive optimizers. *CoRR*, 2025.

Matan Gavish and David L. Donoho. Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, 2017.

Atefeh Gilani, Naima Tasnim, Lalitha Sankar, and Oliver Kosut. Geoclip: Geometry-aware clipping for differentially private sgd. In *Advances in Neural Information Processing Systems 38*, 2025.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems 34, NeurIPS 2021*, 2021.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Jiyan He, Xuechen Li, Da Yu, Huishuai Zhang, Janardhan Kulkarni, Yin Tat Lee, Arturs Backurs, Nenghai Yu, and Jiang Bian. Exploring the limits of differentially private deep learning with group-wise clipping. In *International Conference on Learning Representations*, 2023.

Peter Kairouz, Brendan McMahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, 2021.

Ryo Karakida, Shotaro Akaho, and Shun ichi Amari. Universal statistics of fisher information in deep neural networks: Mean field approach. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, 2019.

Ryo Karakida, Shotaro Akaho, and Shun ichi Amari. Pathological spectra of the fisher information metric and its variants in deep neural networks. *Neural Computation*, 2021.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015.

Xiaocan Li, Shuo Wang, and Yinghao Cai. Tutorial: Complexity analysis of singular value decomposition and its variants, 2019.

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

Guanbiao Lin, Hongyang Yan, Guang Kou, Teng Huang, Shiyu Peng, Yingying Zhang, and Changyu Dong. Understanding adaptive gradient clipping in dp-sgd, empirically. *Int. J. Intell. Syst.*, 2022.

H. Brendan McMahan and Galen Andrew. A general approach to adding differential privacy to iterative training procedures. *CoRR*, abs/1812.06210, 2018.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

Ilya Mironov. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium, CSF 2017*, 2017.

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

Qiaoyue Tang, Frederick Shpilevskiy, and Mathias Lécuyer. Dp-adambc: Your dp-adam is actually DP-SGD (unless you apply bias correction). In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*, 2024.

Florian Tramèr and Dan Boneh. Differentially private learning needs better features (or much more data). In *9th International Conference on Learning Representations, ICLR 2021*, 2021.

A Appendix

A.1 A Ky Fan Lemma for the Square-Root Mass

Lemma A.1 is invoked in Step 1 of Appendix A.3; we state and prove it here. It is the square-root analogue of the classical Ky Fan maximum principle, and may be of independent use in trace-constrained spectral allocation.

Lemma A.1 ($\sqrt{}$-Ky Fan)**.** *Let $A \succeq 0$ have eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$. For any orthonormal set $\{v_1, \ldots, v_k\}$,*

$$\sum_{i=1}^{k} \sqrt{v_i^\top A v_i} \;\leq\; \sum_{i=1}^{k} \sqrt{\lambda_i},$$

with equality iff $\operatorname{span}\{v_1, \ldots, v_k\}$ is contained in the top-k eigenspace of A (and equal to it whenever $\lambda_k > \lambda_{k+1}$).

Proof. Let $V = [\, v_1 \;\ldots\; v_k \,]$ and consider the $k \times k$ compression $V^\top A V$ with eigenvalues $\mu_1 \geq \cdots \geq \mu_k \geq 0$; these are also the eigenvalues of the operator PAP restricted to its range, $P = VV^\top$. Two classical facts chain together. First, *interlacing*: the eigenvalues of a compression satisfy $\mu_i \leq \lambda_i$ for all $i \leq k$ (the minimax principle applied to the i-th Rayleigh quotient over subspaces of the range). Second, *Schur majorization*: the diagonal entries $d_i = v_i^\top A v_i$ of the Hermitian matrix $V^\top A V$, sorted as $d_{(1)} \geq \cdots \geq d_{(k)}$, are weakly majorized by its eigenvalues, $\sum_{i \leq j} d_{(i)} \leq \sum_{i \leq j} \mu_i$ for all $j \leq k$. Since $\sqrt{\cdot}$ is concave and increasing, Karamata's inequality gives $\sum_i \sqrt{d_{(i)}} \leq \sum_i \sqrt{\mu_i} \leq \sum_{i \leq k} \sqrt{\lambda_i}$, proving the inequality.

For equality, both chained inequalities must be equalities: $d = \mu$ (the v_i diagonalize the compression) and $\mu_i = \lambda_i$ for $i \leq k$. When $\lambda_k > \lambda_{k+1}$, the equality case of interlacing forces the compression's top eigenvectors to coincide with those of A, hence $\operatorname{span}\{v_i\} = \operatorname{span}\{u_1, \ldots, u_k\}$; the converse is immediate. □

Remark A.2 (Ties are harmless for Proposition 4.2)**.** When $\lambda_k = \lambda_{k+1}$, maximizers of Lemma A.1 are not unique: any k-dimensional subspace of the corresponding eigenspace qualifies. This does not affect Appendix A.3: whichever top-k set the pooled estimator returns, the counting argument of Step 2 applies to it, and the gap lower bound depends only on the spectrum.

A.2 Proof of Theorem 4.1

Lemma A.3 (Per-direction bias)**.** *Let $\tilde{g} = s + n$ with s the clipped signal part and $n = \frac{\sigma}{\sqrt{|B|}} M^{\mathrm{inv}} \varepsilon$. For every unit vector v,*

$$\mathbb{E}\big[\big(v^\top(\tilde{g} - a)\big)^2 \,\big|\, \theta^t\big] \;=\; \mathbb{E}\big[\big(v^\top(s - a)\big)^2 \,\big|\, \theta^t\big] \;+\; \frac{\sigma^2}{|B|}\, v^\top (M^\top M)^{-1} v,$$

where we used $M^{\mathrm{inv}} M^{\mathrm{inv}\top} = (M^\top M)^{-1}$, valid for the full-rank transform of Theorem 4.3 (and on the column space in the low-rank variant).

Proof. Expand $\tilde{g} - a = (s - a) + n$. The two cross terms vanish in expectation since $\mathbb{E}[\varepsilon] = 0$ and ε is independent of s; the remaining term is $\frac{\sigma^2}{|B|} M^{\mathrm{inv}} \mathbb{E}[\varepsilon\varepsilon^\top] M^{\mathrm{inv}\top} = \frac{\sigma^2}{|B|} M^{\mathrm{inv}} M^{\mathrm{inv}\top}$. For $M = \tau^{-1/2} \Lambda_c^{-1/4} V^\top$, $M^{\mathrm{inv}} = \tau^{1/2} V \Lambda_c^{1/4}$, hence $M^{\mathrm{inv}} M^{\mathrm{inv}\top} = \tau V \Lambda_c^{1/2} V^\top = (M^\top M)^{-1}$. □

Step 1: bias of the $|B|$-scaled EMA (claims (i), (ii)). The EMA update computes a weighted average of outer products $|B|(\tilde{g}_t - a_t)(\tilde{g}_t - a_t)^\top$. By Lemma A.3, the noise contributes, per eigendirection u_i,

$$|B| \cdot \frac{\sigma^2}{|B|}\, \tau \sqrt{\lambda_{c,i}} \;=\; \sigma^2 \tau \sqrt{\lambda_{c,i}}.$$

By induction over the EMA recursion the same contamination holds for the stationary estimate, proving (i). Claim (ii) is exact, not asymptotic: it reflects the identity $|B| \cdot \mathrm{Cov}(\text{mean}) = \mathrm{Cov}(\text{single})$; increasing the batch reduces neither the systematic inflation nor the floor it induces below.

Step 2: anisotropy and the failure of isotropic correction. The contamination operator $\sigma^2\tau V\,\mathrm{diag}(\sqrt{\lambda_c})\,V^\top$ has eigenvalues $\sigma^2\tau\sqrt{\lambda_{c,i}}$: proportional to $\sqrt{\lambda_i}$, never uniform. Subtracting any scalar δ per coordinate subtracts δ from every eigenvalue of the estimate; matching the largest bias ($\delta = \sigma^2\tau\sqrt{\lambda_{\max}}$) over-corrects all smaller directions into negative values, while matching any average bias under-corrects each direction by a factor depending on its own $\sqrt{\lambda_i}$. No scalar correction is consistent under the scale invariance $(M, C, \sigma) \equiv (M/C, 1, \sigma/C)$ of Remark 3.2.

Step 3: exact inversion (claim (iii)). With τ frozen, $\widehat{\lambda}_i$ and λ_i obey $\widehat{\lambda}_i = \lambda_i + \sigma^2\tau\sqrt{\lambda_i}$. Setting $x = \sqrt{\lambda_i}$ gives $x^2 + \sigma^2\tau\, x - \widehat{\lambda}_i = 0$, whose nonnegative root yields equation 3. *Self-consistency:* in practice τ is computed from $\{\widehat{\lambda}_i\}$; since $\sqrt{\widehat{\lambda}_i} = \sqrt{\lambda_i} + \frac{\sigma^2\tau}{2} + o(\sigma^2)$, we have $\tau(\widehat{\lambda}) = \tau(\lambda) + O(\sigma^2\tau)$, and substituting into equation 3 leaves a residual of $O(\sigma^4\tau^2)$: the correction is first-order exact. The clamp at h_1 preserves positive semi-definiteness and is applied after correction—before clamping, so that contaminated small eigenvalues are not masked behind h_1; τ is recomputed after clamping, keeping the normalization consistent with the constrained quantity $\mathrm{Tr}(M^\top M\widehat{\Sigma})$.

Step 4: special cases. *Diagonal reduction.* For $M = \mathrm{diag}(b_j^{-1})$, $M^{\mathrm{inv}}M^{\mathrm{inv}\top} = \mathrm{diag}(b_j^2)$ and the contamination of coordinate j is $\sigma^2 b_j^2$ (after $|B|$-scaling): exactly the per-coordinate noise-variance subtraction of AdaCliP (Pichapati et al., 2019) up to the noise convention. Theorem 4.1 extends that correction from diagonal to general geometries, where it becomes anisotropic and scale-coupled through τ. *Low-rank confinement.* In the streaming rank-k variant, $\tilde{\omega}$ and hence n are supported on the k-dimensional principal subspace; the contamination operator is $\sigma^2\tau V_k\mathrm{diag}(\sqrt{\lambda_c})V_k^\top$ on that subspace, so equation 3 acts only on the retained eigenvalues, and the orthogonal complement—never released—remains uncontaminated.

Step 5: claim (iv). Immediate from Lemma A.3: the contamination operator is $\frac{\sigma^2}{|B|}(M^\top M)^{-1}$, and $\mathrm{Tr}[(M^\top M)^{-1}]$ is precisely the noise-injection term of Theorem 4.4. The pollution is, eigenvalue by eigenvalue, the operator whose trace the metric of Theorem 4.3 is chosen to minimize. □

Step 6: residual floor (proof of Theorem 4.1). *Step 6.1 (fluctuation, claim (i)).* Write the stationary EMA as a weighted sum over history, $\widehat{\Sigma} = \sum_s w_s\,|B|(\tilde{g}_s - a_s)(\tilde{g}_s - a_s)^\top$ with $w_s = (1-\beta_2)\beta_2^{T-s}$ and $\sum_s w_s = 1$. The noise parts of the summands are independent across s (the ε_s are independent), so for the corrected estimator $\widetilde{\lambda}_i$,

$$\mathrm{Var}\big[\widetilde{\lambda}_i\big] \;=\; \sum_s w_s^2\,\mathrm{Var}\Big[|B|\,(u_i^\top n_s)^2\Big].$$

By Lemma A.3, $u_i^\top n_s \sim \mathcal{N}\big(0,\, \frac{\sigma^2\tau\sqrt{\lambda_{c,i}}}{|B|}\big)$, and for a centered Gaussian $X \sim \mathcal{N}(0, v)$ the fourth moment gives $\mathrm{Var}[X^2] = 2v^2$; hence

$$\mathrm{Var}\Big[|B|(u_i^\top n_s)^2\Big] \;=\; |B|^2 \cdot 2\Big(\frac{\sigma^2\tau\sqrt{\lambda_{c,i}}}{|B|}\Big)^2 \;=\; 2\,(\sigma^2\tau)^2\,\lambda_{c,i}.$$

The geometric sum of squared weights is $\sum_s w_s^2 = \frac{(1-\beta_2)^2}{1-\beta_2^2} = \frac{1-\beta_2}{1+\beta_2} = B_{\mathrm{eff}}^{-1}$, so

$$\mathrm{Var}\big[\widetilde{\lambda}_i\big] \;=\; \frac{2\,\sigma^4\tau^2\lambda_{c,i}}{B_{\mathrm{eff}}} \;+\; o(\sigma^4),$$

proving (i). Note the fluctuation is proportional to $\lambda_{c,i}$, the *contaminated* scale of the direction, not to its true λ_i: directions are lost relative to their polluted magnitude, which is what makes the floor self-reinforcing at small λ_i.

Step 6.2 (thresholds, claim (ii), first half). A direction carries information iff signal dominates noise at the relevant level. At the systematic level:

$$\lambda \;\geq\; \sigma^2\tau\sqrt{\lambda} \;\iff\; \sqrt{\lambda} \geq \sigma^2\tau \;\iff\; \lambda \geq \lambda^*_{\mathrm{sys}} = \sigma^4\tau^2.$$

At the fluctuation level, the noise standard deviation from Step 6.1 is $\sqrt{2}\,\sigma^2\tau\sqrt{\lambda}/\sqrt{B_{\text{eff}}}$, so

$$\lambda \;\geq\; \frac{\sqrt{2}\,\sigma^2\tau\sqrt{\lambda}}{\sqrt{B_{\text{eff}}}} \;\iff\; \lambda \geq \lambda^*_{\text{corr}} = \frac{2\,\sigma^4\tau^2}{B_{\text{eff}}},$$

the factor 2 again tracing to the Gaussian fourth moment. Since $B_{\text{eff}} \geq 1$, correction lowers the floor from $\sigma^4\tau^2$ to $2\sigma^4\tau^2/B_{\text{eff}}$: the systematic component, once removed, is gone for good.

Step 6.3 (clamp, claim (ii), second half). From Step 3, the Taylor expansion $\sqrt{\widehat{\lambda}_i} = \sqrt{\lambda_i} + \frac{\sigma^2\tau}{2} + o(\sigma^2)$ shows the corrected estimate satisfies

$$\widetilde{\lambda}_i \;\leq\; \lambda_i - \tfrac{1}{2}(\sigma^2\tau)^2 + \rho_i,$$

where the constant $\frac{1}{2}$ is the first-order Taylor coefficient of $\sqrt{\cdot}$ just displayed, and ρ_i is the fluctuation term with $\mathbb{E}[\rho_i] = 0$ and $\mathrm{Var}[\rho_i]$ from Step 6.1. Hence whenever

$$\lambda_i \;\leq\; \tfrac{1}{2}(\sigma^2\tau)^2 \;=\; \tfrac{1}{2}\,\lambda^*_{\text{sys}},$$

Markov's inequality gives $\Pr[\widetilde{\lambda}_i > h_1] \leq \mathrm{Var}[\rho_i]/\big(\frac{1}{2}\lambda^*_{\text{sys}} - \lambda_i\big)^2$, which is $o(1)$ for λ_i bounded away from the threshold as $B_{\text{eff}} \to \infty$ (standard in practice: $\beta_2 \geq 0.99$ over thousands of steps). The clamp at h_1 is therefore the high-probability consequence of sub-floor estimation—not a numerical accident—and it is the honest encoding of "this direction contains no information."

Step 6.4 (monitorability, claim (iii)). Both λ^*_{sys} and λ^*_{corr} depend only on (σ, τ, β_2); τ is maintained by the algorithm itself (Eq. 12). Per layer, the threshold λ^* and observed clamp fraction are computable at no cost per iteration, providing built-in feasibility diagnostics: a layer whose spectrum lies mostly below λ^* has under-estimable geometry under the current privacy budget, and the method degrades it to standard clipping (Sec. 5).

A.3 Proof of Prop 4.2

Step 1: the pooled optimum and claim (a). For a rank-k transform with column-orthonormal range $\{v_1, \ldots, v_k\}$ and scales $m_i > 0$, the same Lagrangian computation as in Appendix A.4, Step 2, gives optimal scales $m_i^4 \propto \mu_i^{-1}$ ($\mu_i = v_i^\top \bar{\Sigma} v_i$) and value $(\sum_i \sqrt{\mu_i})^2/\gamma$. By Lemma A.1,

$$\max_{\text{orthonormal } \{v_i\}} \sum_{i=1}^{k} \sqrt{v_i^\top \bar{\Sigma} v_i} \;=\; \sum_{i=1}^{k} \sqrt{\lambda_{(i)}},$$

attained uniquely at the global top-k eigenspace. Under equation 2, layer b has fewer than k eigenvalues among the pooled top-k: otherwise its k-th would satisfy $\lambda_{b,k} \geq \lambda_{(k)}$. Hence the range of any pooled transform meets the b-coordinate subspace in dimension $k_b^G < k$.

Step 2: the gap, claim (b). Every pooled top-k direction of layer l lies in layer l's own top-k (a direction in the global top-k has at most $k-1$ superiors within its layer). The layered set therefore *contains* the pooled set, and

$$\Delta_A \;=\; \underbrace{\sum_{l} S_l(k)}_{\text{layered } Lk \text{ directions}} \;-\; \underbrace{\sum_{i \leq k} \sqrt{\lambda_{(i)}}}_{\text{pooled } k \text{ directions}} \;\geq\; S_b(k) - \sum_{\substack{i \leq k \\ i \in b}} \sqrt{\lambda_{b,i}} \;\geq\; \sqrt{\lambda_{b,k}} > 0,$$

under the wide-weak condition of Proposition 4.2, layer b has fewer than k eigenvalues among the pooled top-k: otherwise its k-th would satisfy $\lambda_{b,k} \geq \lambda_{(k)}$.

Step 3: layered attains the bound. The layered construction solves L independent subproblems with budgets $\gamma_l \propto S_l(k)$; by the Cauchy–Schwarz step of Appendix A.4, Step 4, proportional budgets leave no cross-layer efficiency on the table (S_l/γ_l equalized). Its $\sqrt{}$-mass is exactly $\sum_l S_l(k)$: equality in Step 2.

Step 4: polynomial spectra, claim (c). For $\lambda_{b,i} = c_b i^{-\beta}$, $S_b(k) = \sqrt{c_b}\sum_{i\le k} i^{-\beta/2} = \sqrt{c_b}\,\Theta\big(k^{1-\beta/2}\big)$ for $\beta < 2$ (and $\Theta(\sqrt{c_b}\log k)$ at $\beta = 2$); the subtracted term counts at most $k-1$ directions and is of no larger order, so $\Delta_A = \Omega(\sqrt{c_b}k^{1-\beta/2})$.

Step 5: the h_1-path analysis, claim (d). The h_1-path analysis. A pooled transform supported on the top-k eigenspace assigns layer b no principal direction; through the orthogonal-complement path retained by the mechanism (Remark 5.1), b's component is scaled by $h_1^{-1/4}/\sqrt{\tau}$ instead of its correct $\lambda_{b,1}^{-1/4}/\sqrt{\tau}$: an over-amplification of factor $(\lambda_{b,1}/h_1)^{1/4}$. Three consequences follow. *(i) Constraint.* b's mass enters the clip proxy at the wrong scale: $\mathrm{Tr}(M^\top M\Sigma_b) \ni \tau^{-1}h_1^{-1/2}\,\mathrm{Tr}(P_{\perp,b}\Sigma_b)$, inflating the clipping probability for the *joint* clip by the same Markov term that the convergence analysis charges to the clipping error. *(ii) Coupling.* Because the pooled baseline normalizes the concatenated ω jointly, the oversized b-block compresses every other layer's signal in the same update — the per-layer distortions controlled above become cross-layer through the shared denominator, and Jensen's inequality carries the compression into the descent term. *(iii) Noise mispricing.* Along the h_1-path the released noise returns to parameter space scaled by $h_1^{1/4}\sqrt{\tau}$ while the signal scales as $h_1^{-1/4}/\sqrt{\tau}$: the signal-to-noise ratio of b's surviving components is off by $h_1^{1/2}/\tau$ relative to its correct pricing. Assembling (i)–(iii), the layer-wise convergence bound retains a clipping excess of order $\beta_b\,(h_1^{-1/2}/\tau)\,\mathrm{Tr}(P_{\perp,b}\Sigma_b)$ per step, absent under layered estimation, which prices b at its own spectrum. Since the excess scales as $h_1^{-1/2}$, decreasing h_1 worsens the pooled failure while leaving the layered construction unaffected except through the complement path it retains by design, yielding a testable prediction. □

A.4 Proof of Theorem 4.3

The argument parallels (Gilani et al., 2025), Appendix B, with two modifications noted below: the debiased–clamped spectrum replaces the raw estimate, and the whitening comparison is added as the equality case of the allocation step.

Step 0: Gram reduction and convexity. For any $A \succ 0$ there is a factor M with $M^\top M = A$ (take $M = A^{1/2}$); objective and constraint depend on M only through A. On $\mathbb{S}_{++}$, $A \mapsto \mathrm{Tr}[A^{-1}]$ is strictly convex and strictly decreasing in the Loewner order, and the constraint is linear: the program is therefore convex, with a unique Gram optimum.

Step 1: exact activity of the constraint. If A is feasible with slack, scaling tA with $t = \gamma/\mathrm{Tr}(A\Sigma) > 1$ remains feasible and strictly decreases $\mathrm{Tr}[A^{-1}]$ by $1/t$; every optimum is active, $\mathrm{Tr}(A^*\Sigma) = \gamma$.

Step 2: stationarity and the multiplier. With $\Sigma \succ 0$ (after clamping $\lambda_i \ge h_1 > 0$) the Slater condition holds. The Lagrangian $\mathcal{L} = \mathrm{Tr}[A^{-1}] + \mu\big(\mathrm{Tr}(A\Sigma) - \gamma\big)$ has derivatives $\nabla_A\,\mathrm{Tr}[A^{-1}] = -A^{-2}$ and $\nabla_A\,\mathrm{Tr}(A\Sigma) = \Sigma$, giving $A^2 = \mu^{-1}\Sigma^{-1}$, i.e. $A = \mu^{-1/2}\Sigma^{-1/2}$. Activity yields $\gamma = \mu^{-1/2}\mathrm{Tr}(\Sigma^{1/2}) = \mu^{-1/2}S$, so $A^* = (\gamma/S)\,\Sigma^{-1/2}$ and $M^* = (\gamma/S)^{1/2}\Lambda^{-1/4}U^\top$ in the eigenbasis. Back-substitution: $\mathrm{Tr}[A^{*-1}] = (S/\gamma)\mathrm{Tr}[\Sigma^{1/2}] = S^2/\gamma$. Uniqueness: strict convexity of $\mathrm{Tr}[A^{-1}]$ on $\mathbb{S}_{++}$ gives a unique A^*; M^* is unique up to left orthogonal factors, under which the mechanism is invariant.

Step 3: transformed covariance and ordering. Direct computation: $M^*\Sigma M^{*\top} = (\gamma/S)\Lambda^{1/2}$: relative variance ordering is preserved—unlike whitening, which maps the spectrum to $(\gamma/k)I$ on a k-dimensional truncation and erases it.

Step 4: the allocation step, reused twice. Cauchy–Schwarz in the form $\big(\sum_i \sqrt{\lambda_i}\big)^2 \le k\sum_i \lambda_i$, with equality iff $\lambda_1 = \cdots = \lambda_k$, serves two purposes in this paper. *(i) Whitening comparison (Theorem 4.3, second half).* Full whitening $M = (\gamma/d)^{1/2}\Lambda^{-1/2}U^\top$ has $M^\top M = (\gamma/d)\Sigma^{-1}$: it is feasible with exact activity, $\mathrm{Tr}(M^\top M\Sigma) = (\gamma/d)\cdot d = \gamma$, and attains value $\mathrm{Tr}[(M^\top M)^{-1}] = (d/\gamma)\sum_i \lambda_i$. The Cauchy–Schwarz inequality $\big(\sum_i \sqrt{\lambda_i}\big)^2 \le d\sum_i \lambda_i$, with equality iff $\lambda_1 = \cdots =$

λ_d, gives $(d/\gamma)\sum_i \lambda_i \geq (\sum_i \sqrt{\lambda_i})^2/\gamma = S^2/\gamma$, strictly in the non-flat case. On a k-dimensional truncation the identical computation yields $(k/\gamma)\sum_{i\leq k} \lambda_i \geq (\sum_{i\leq k} \sqrt{\lambda_i})^2/\gamma$, with equality iff the retained spectrum is flat. *(ii) Cross-layer allocation* (invoked in Appendix A.3, Step 3): minimizing $\sum_l S_l^2/\gamma_l$ s.t. $\sum_l \gamma_l = \gamma$ gives $\gamma_l \propto S_l$ with value $(\sum_l S_l)^2/\gamma$; the same inequality shows any non-proportional split is strictly worse.

Step 5: pipeline order (modification note). In the implementation, the eigendecomposition in Step 2 is computed on the debiased and clamped eigenvalues $\lambda_i \leftarrow \max\{\lambda_i - \sigma^2\tau\sqrt{\lambda_i},\, h_1\}$, clamped above at h_2, and S and τ are recomputed *after* this pipeline. Correction before clamping is necessary (clamping first would mask the contamination of small eigenvalues); recomputing τ after keeps the normalization consistent with the constrained quantity. □

A.5 Proof of Theorem 4.4

Appendix A, applied to the block-diagonal concatenation. We give the full argument for completeness, highlighting the two steps where the block structure enters.

Step 0: concatenated mechanism. Define $M_t = \mathrm{blkdiag}(M_t^{(1)}, \ldots, M_t^{(m)})$, $a_t = \bigoplus_{l=1}^m a_t^{(l)}$, and per-sample noise $n_t = \bigoplus_l n_t^{(l)}$ with $n_t^{(l)} \sim \mathcal{N}(0, \sigma^2 I)$ (per-sample level σ; batch averaging scales all noise terms by $1/|B|$). The update is $\theta_{t+1} = \theta_t - \eta\tilde{g}_t$ with

$$\tilde{g}_t = M_t^{-1}\hat{\omega}_t + a_t, \quad \hat{\omega}_t = \mathrm{clip}_1\left(M_t(g_t - a_t) + n_t\right).$$

Because blocks are coordinate-disjoint, for any concatenated vectors $x = \oplus_l x^{(l)}$, $y = \oplus_l y^{(l)}$:

$$\|x\|^2 = \sum_l \|x^{(l)}\|^2, \qquad \langle x, y\rangle = \sum_l \langle x^{(l)}, y^{(l)}\rangle. \tag{12}$$

Step 1: per-layer clipping error. Let $c_t^{(l)}$ denote the clipped centered part and $\Delta_t^{(l)} = c_t^{(l)} - (g_t^{(l)} - a_t^{(l)})$ the clipping distortion. Exactly as in Gilani et al. (2025), Eq. (52)–(53), $\Delta_t^{(l)} = 0$ when no clipping occurs and $\|\Delta_t^{(l)}\| \leq \|g_t^{(l)} - a_t^{(l)}\| \leq G + \|a_t^{(l)}\|$ otherwise, so with $\theta^t = (\theta_0, \ldots, \theta_t)$,

$$\mathbb{E}\big[\|\Delta_t^{(l)}\| \,\big|\, \theta^t\big] \leq \big(G + \|a_t^{(l)}\|\big)\, \Pr\big[\|M_t^{(l)}(g_t^{(l)} - a_t^{(l)})\| > 1 \,\big|\, \theta^t\big].$$

The per-layer Markov bound: writing $\mu^{(l)} = \mathbb{E}[g_t^{(l)} \mid \theta^t]$,

$$\begin{aligned}\Pr\big[\|M_t^{(l)}(g_t^{(l)} - a_t^{(l)})\| > 1 \,\big|\, \theta^t\big] &\leq \mathbb{E}\big[\|M_t^{(l)}(g_t^{(l)} - a_t^{(l)})\|^2 \,\big|\, \theta^t\big] \\ &= \mathrm{Tr}\big(M_t^{(l)\top} M_t^{(l)} \Sigma_t^{(l)}\big) + \big\|M_t^{(l)}\big(\mu^{(l)} - a_t^{(l)}\big)\big\|^2,\end{aligned}$$

using $\mathbb{E}\|X\|^2 = \mathrm{Tr}(\mathrm{Cov}\, X) + \|\mathbb{E}X\|^2$.

Step 2: smoothness descent. By L-smoothness, $\mathbb{E}[f(\theta_{t+1}) \mid \theta^t] \leq f(\theta_t) - \eta\langle \nabla f(\theta_t), \mathbb{E}[\tilde{g}_t \mid \theta^t]\rangle + \frac{L\eta^2}{2}\mathbb{E}[\|\tilde{g}_t\|^2 \mid \theta^t]$. Decompose $\tilde{g}_t = \Delta_t + g_t + \mathcal{M}_t^{-1} n_t$ with $\Delta_t = \oplus_l \Delta_t^{(l)}$. By equation 12 and independence of n_t ($\mathbb{E}[n_t] = 0$, $\mathbb{E}[n_t n_t^\top] = \sigma^2 I$),

$$\mathbb{E}\|\tilde{g}_t\|^2 = \mathbb{E}\|\Delta_t + g_t\|^2 + \sigma^2\, \mathrm{Tr}\big[(\mathcal{M}_t^\top \mathcal{M}_t)^{-1}\big], \qquad \mathrm{Tr}\big[(\mathcal{M}_t^\top \mathcal{M}_t)^{-1}\big] = \sum_l \mathrm{Tr}\big[(M_t^{(l)\top} M_t^{(l)})^{-1}\big],$$

where the second equality uses block diagonality of $\mathcal{M}_t^\top \mathcal{M}_t$. The inner product term decomposes by equation 12 as well:

$$\langle \nabla f, \mathbb{E}[\Delta_t + g_t]\rangle = \|\nabla f(\theta_t)\|^2 + \sum_l \big\langle \nabla_l f(\theta_t),\, \mathbb{E}[\Delta_t^{(l)} \mid \theta^t]\big\rangle,$$

using $\mathbb{E}[g_t^{(l)} \mid \theta^t] = \nabla_l f(\theta_t)$.

Step 3: Cauchy–Schwarz and Jensen, per layer. For each layer, $\langle \nabla_l f, \mathbb{E}[\Delta_t^{(l)} \mid \theta^t] \rangle \le \|\nabla_l f\| \, \mathbb{E}[\|\Delta_t^{(l)}\| \mid \theta^t]$, and

$$\mathbb{E}\|\Delta_t + g_t\|^2 \le 3 \sum_l \Big(\mathbb{E}\|g_t^{(l)} - \nabla_l f\|^2 + \|\nabla_l f\|^2 + \mathbb{E}\|\Delta_t^{(l)}\|^2 \Big) = 3\Big(\sigma_g^2 + \|\nabla f\|^2 + \sum_l \mathbb{E}\|\Delta_t^{(l)}\|^2 \Big),$$

by equation 12 and $\sum_l \|\nabla_l f\|^2 = \|\nabla f\|^2$. Substituting the Step-1 bounds and $\|\nabla_l f\| \le G$ (Jensen) combines the two clipping contributions into $\eta\, \beta_l(a_t^{(l)}) \big(\mathrm{Tr}(M_t^{(l)\top} M_t^{(l)} \Sigma_t^{(l)}) + \|M_t^{(l)}(\mu^{(l)} - a_t^{(l)})\|^2 \big)$ per layer.

Step 4: telescoping. Collecting terms,

$$\Big(\eta - \frac{3L\eta^2}{2} \Big) \mathbb{E}\|\nabla f(\theta_t)\|^2 \le \mathbb{E} f(\theta_t) - \mathbb{E} f(\theta_{t+1}) + \frac{3L\eta^2}{2} \sigma_g^2 + \frac{L\eta^2 \sigma^2}{2} \sum_l \mathbb{E}\, \mathrm{Tr}\big[(M_t^{(l)\top} M_t^{(l)})^{-1} \big] + \eta \sum_l \mathbb{E}\big[\beta_l(a_t^{(l)}) \big(\cdots \big) \big],$$

summing over t, telescoping, using $f(\theta_T) \ge f(\theta^*)$, and dividing by $T(\eta - \frac{3L\eta^2}{2})$ (positive since $\eta < \frac{2}{3L}$) yields equation 6. □

Proof of the allocation claim. With $a_t^{(l)} = \mathbb{E}[g_t^{(l)} \mid \theta^t]$ the bias term vanishes; with $M_t^{(l)} = M_t^{(l)*}$, the half-whitening optimum gives $\mathrm{Tr}[(M_t^{(l)*\top} M_t^{(l)*})^{-1}] = S_l^2/\gamma_l$ (noise term), exact constraint activity $\mathrm{Tr}(M_t^{(l)*\top} M_t^{(l)*} \widehat{\Sigma}_t^{(l)}) = \gamma_l$ (clipping term), and $\beta_l \le 2G^2(1 + 3\mathcal{L}\eta)$ by Jensen, $G + \|a\| \le 2G$. Minimizing

$$\frac{\mathcal{L}\eta\sigma^2}{|B|} \sum_l \frac{S_l^2}{\gamma_l} \;+\; 4G^2(1 + 3\mathcal{L}\eta) \sum_l \gamma_l$$

over $\{\gamma_l\}$ gives $\gamma_l \propto S_l$ by a standard Cauchy–Schwarz argument. □

A.6 PRIVACY: THE ADAPTIVE-COMPOSITION STEP

We formalize the induction sketched in Proposition 5.2.

Base case. At $t = 0$ the transforms $(M_0^{(l)}, a_0^{(l)})$ are data-independent (Algorithm 1), so the release $\tilde{g}_0$ is exactly the subsampled Gaussian mechanism: Poisson sampling at rate q followed by a Gaussian mechanism with multiplier σ_{mult} on the clipped mean (sensitivity $2C/|B|$, per-coordinate noise $\sigma/\sqrt{|B|}$). Its Rényi divergence of order λ is bounded by $q^2\lambda/(2\sigma_{\mathrm{mult}}^2)$.

Inductive step. Assume $\{\tilde{g}_\tau\}_{\tau<t}$ satisfies $(\varepsilon_{t-1}, \delta_{t-1})$-DP. The statistics $(M_t^{(l)}, a_t^{(l)})$ are deterministic functions of $\{\tilde{g}_\tau\}_{\tau<t}$, hence post-processing; conditioned on them, iteration t is the same subsampled Gaussian mechanism as the base case, with unchanged sensitivity and noise. By adaptive composition of Rényi divergence, the T-fold composition satisfies (ε, δ)-DP for exactly the (ε, δ) that the accountant outputs from $(\sigma_{\mathrm{mult}}, q, T)$ — identically to standard DP-SGD.

B ADDITIONAL EXPERIMENTS

Table 5: Test accuracy (%) on IMDB (LSTM and TF) under varying privacy budgets ($\delta = 10^{-5}$).

Model	**Method**	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 8$
LSTM@IMDB	DP-SGD	78.2±4.3	83.1±2.9	86.0±0.8
	Quantile	79.0±4.0	84.0±2.9	87.0±0.8
	AdaClip	80.1±3.9	85.2±2.7	88.1±0.8
	GeoClip	81.2±3.6	86.1±2.6	89.0±0.7
	SP-GeoClip	**83.3**±3.5	**88.2**±2.4	**90.6**±0.7
TF@IMDB	DP-SGD	84.5±3.9	88.0±2.8	90.1±0.7
	Quantile	85.1±3.9	88.6±2.7	90.6±0.7
	AdaClip	85.8±3.8	89.2±2.7	91.2±0.6
	GeoClip	87.1±3.8	90.3±2.5	92.0±0.6
	SP-GeoClip	**88.4**±3.7	**91.2**±2.5	**93.0**±0.5

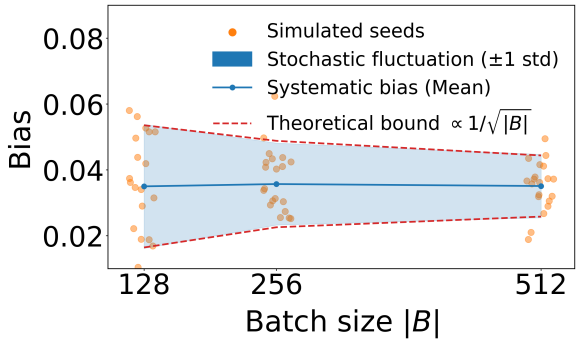

Figure 3: Batch-size invariance of spectral estimation.

Table 6: Test accuracy (%) across nine dataset-model combinations at $\epsilon = 3$ ($\delta = 10^{-5}$). SP-GeoClip uniformly outperforms all baselines.

Model	DP-SGD	Quantile	AdaClip	GeoClip	**SP-GeoClip**
MLP@MNIST	94.1±1.6	94.8±1.6	95.3±1.3	95.9±1.5	**96.4**±1.1
CNN@MNIST	93.1±2.3	93.7±1.9	94.0±1.6	94.4±2.0	**95.3**±1.3
CNN@CIFAR10	68.3±2.0	70.1±2.0	72.2±1.9	75.1±1.9	**77.3**±1.8
EmbedAvg@AGN	87.1±2.8	88.0±2.8	89.0±2.7	90.1±2.5	**91.3**±2.5
LSTM@AGN	81.0±2.0	81.9±1.6	82.6±2.3	84.0±1.9	**86.4**±1.4
TF@AGN	76.8±1.5	77.6±2.2	78.4±1.8	79.8±2.4	**81.9**±1.7
EmbedAvg@IMDB	80.5±2.1	81.4±1.4	82.2±1.9	83.8±2.3	**86.2**±1.6
LSTM@IMDB	83.1±2.9	84.0±2.9	85.2±2.7	86.1±2.6	**88.2**±2.4
TF@IMDB	88.0±2.8	88.6±2.7	89.2±2.7	90.3±2.5	**91.2**±2.5

Table 7: Half-whitening vs. full whitening. Gap Δ and bound ratio R confirm strict superiority.

Model	**Metric**	$\varepsilon = 1$	$\varepsilon = 3$	$\varepsilon = 8$	R
MLP@MNIST	Half-whitening	90.9	93.9	95.2	—
	Full whitening	89.8	92.8	94.6	~ 4.7
	Δ	-1.1	-1.1	-0.6	
LSTM@IMDB	Half-whitening	82.2	86.5	89.0	—
	Full whitening	78.1	83.0	86.9	~ 25.8
	Δ	-4.1	-3.5	-2.1	