Spectral Gradient Orthogonalization Improves Differentially Private Training at Scale

Sabari Shanmugam, Nick Barnes, Kerry Taylor

School of Computing, Australian National University, Canberra, Australia
{sabari.shanmugam, nick.barnes, kerry.taylor}@anu.edu.au

Abstract. Differentially private training adds isotropic Gaussian noise to clipped gradients, corrupting every singular direction equally. In vision models, where spatial correlation concentrates gradient energy into a low-rank subspace, most of this noise falls in directions that carry little signal. Spectral gradient orthogonalization via polar decomposition is introduced as a post-processing step that recovers directional signal from the noisy gradient's low-rank structure at zero additional privacy cost. A phase transition governs the utility of this approach: orthogonalization improves accuracy only when the per-direction spectral signal-to-noise ratio (SNR) suffices for singular vector recovery; in low-SNR regimes, the directional bias of the gradient is replaced by a nearly random orthogonal update, and the transformation is harmful. The recovery threshold is determined by the spectral gap of the gradient and is surpassed at large batch sizes. Empirically, the benefit scales with model capacity: spectral orthogonalization achieves a +20.9% improvement over DP-SGD on WRN-28-10 ($B = 4096$) and +14.9% on ResNet-18, while reducing inter-run variance by a factor of two to three. In the fine-tuning regime, spectral orthogonalization matches the stability of DP-Adam while maintaining a first-order memory footprint. Combining spectral with temporal denoising yields 50.3% on CIFAR-10 ($\varepsilon = 4$), the highest accuracy in any tested configuration. These gains are specific to moderate-to-high-SNR regimes such as large-batch training of higher-capacity models. Small-batch or low-SNR settings are better served by DP-SGD or temporal denoising.

Keywords: Differential privacy · Spectral methods · Gradient orthogonalization · Private training · Polar decomposition

1 Introduction

Computer vision models trained on sensitive data, such as medical images or visual biometrics (facial recognition, iris scans), must protect individual privacy while maintaining utility [19]. Differential privacy (DP) [9] provides a formal guarantee: the model's behavior changes negligibly whether or not any single example is included in the training set, preventing the memorization or recall

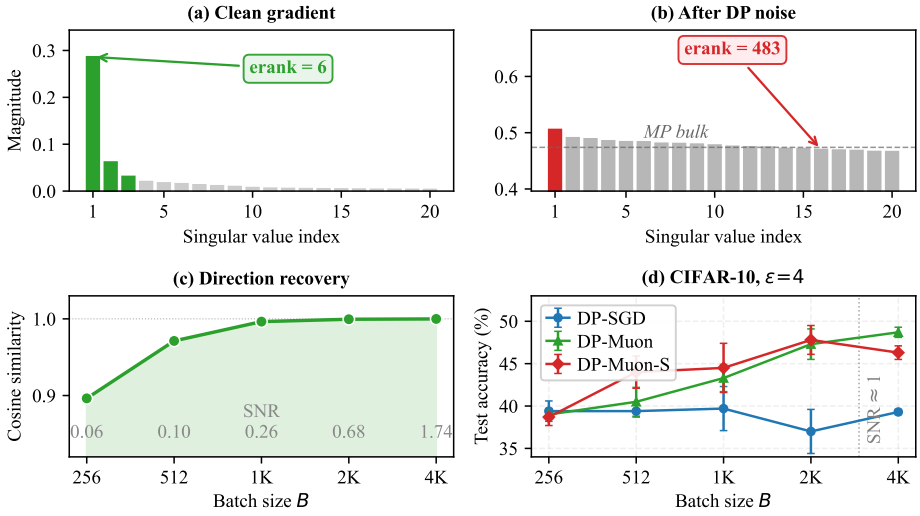

Fig. 1: Spectral gradient orthogonalization overview. *(a)* Gradient singular values from a representative convolutional layer (epoch 1, CIFAR-10): most energy concentrates in a small number of directions (effective rank 6). *(b)* After DP noise addition ($\varepsilon = 4$), isotropic perturbation raises the effective rank to 483: the signal becomes indistinguishable from the noise-dominated singular values (Marchenko–Pastur bulk), making the noisy gradient nearly directionless. *(c)* Direction recovery: cosine similarity between the leading singular vector of the noisy and clean gradients, measured via Monte Carlo simulation (100 trials per batch size) using the measured WRN-16-4 spectrum with noise calibrated to Tab. 1. High similarity indicates that the gradient's directional signal persists through the noise, providing the geometric condition for orthogonalization to succeed. Recovery transitions sharply from 0.90 at $B = 256$ to $\geq$0.99 at $B \geq 1024$. *(d)* Test accuracy on CIFAR-10 ($\varepsilon = 4$) follows the same pattern: DP-SGD stagnates while spectral methods scale with batch size past the recovery threshold.

of individual training samples. The standard mechanism, differentially private stochastic gradient descent (DP-SGD) [1], enforces this by clipping each per-sample gradient to a fixed norm and adding calibrated Gaussian noise before aggregation. The noise ensures that the contribution of any one example is masked, but it degrades the optimization signal, reducing model accuracy. This utility cost is especially severe when training from scratch, where every parameter must be learned under noise. For vision architectures with large convolutional and attention layers, the gradient of each weight matrix is independently perturbed by isotropic noise, regardless of its structure.

Neural network training proceeds by computing gradients of the loss with respect to each weight matrix. Recent work in optimization has shown that these gradients are approximately low-rank [2, 22], a property exploited by optimizers such as Shampoo [14] and Muon [18] for faster convergence. In vision models this structure is especially pronounced: the effective rank of a convolutional

gradient is typically an order of magnitude smaller than the matrix dimension, a property we characterize empirically (Sec. 4). The noise added by DP-SGD, however, is *spectrally indiscriminate*: isotropic Gaussian perturbation corrupts every singular direction equally, regardless of how much signal that direction carries. The result is a spectral mismatch (Fig. 1b): the few directions that carry signal receive the prescribed noise, while the many directions that carry little signal become pure noise, yet contribute equally to the aggregate update.

While spectral orthogonalization is established in non-private optimization [18], its utility under differential privacy is fundamentally regime-dependent. Replacing the noisy gradient with its nearest orthogonal matrix $UV^\top$ via polar decomposition (computed efficiently through Newton-Schulz iteration) is shown to transition from harmful to substantially beneficial at a predictable SNR threshold, providing the first principled criterion for when spectral structure can be recovered from the noise bulk in private training. By the post-processing theorem [10], this operation incurs zero additional privacy cost. Orthogonalization preserves the gradient's singular vectors but sets all singular values to one, suppressing the inflated magnitudes of noise-dominated directions. This is beneficial when the true signal directions are recoverable from the noisy gradient, but harmful otherwise: at low SNR, orthogonalization normalizes the noise bulk to unit norm, producing a nearly random update that is strictly worse than the noisy-but-directionally-biased gradient used by DP-SGD.

This leads to a theoretical prediction: orthogonalization should improve utility when the gradient signal-to-noise ratio (SNR) is high enough for the true singular vectors to be recoverable from the noisy gradient, and be neutral or harmful otherwise. This prediction is verified empirically, revealing a **phase transition** governed by the SNR. At large batch sizes, the noise per element shrinks inversely with B, pushing the SNR past the recovery threshold. The crossover batch size depends on task complexity: it occurs at $B \approx 1024$ on CIFAR-10, shifts to $B \approx 4096$ to 8192 on CIFAR-100 where the gradient signal is $\sim 9\times$ weaker, and is verified across three datasets in Sec. 5.2.

This phenomenon is characterized through a systematic empirical study spanning batch sizes $B \in \{256, 512, \ldots, 8192\}$ and privacy budgets $\varepsilon \in \{2, 3, 4, 6, 8\}$. The contributions of this work are three-fold. First, a recovery condition (Proposition 1) linking batch size, noise multiplier, and spectral gap is derived and verified across CIFAR-10, CIFAR-100, and SVHN, revealing that spectral orthogonalization transitions from harmful to substantially beneficial as the gradient SNR crosses a predictable threshold. Second, spectral gradient orthogonalization via approximate polar decomposition (Newton-Schulz iteration) is introduced as a deterministic post-processing step that preserves the (ε, δ)-DP guarantee at zero additional privacy cost, achieving +20.9% over DP-SGD on WRN-28-10 ($B = 4096$, $\varepsilon = 4$) and reducing inter-run variance by a factor of two to three. Third, the complementarity of spectral and temporal denoising is demonstrated: combining orthogonalization with low-pass filtering (DOPPLER) yields 50.3% on CIFAR-10 ($\varepsilon = 4$), the highest accuracy in any tested configuration.

2 Related Work

Differentially private optimization. DP-SGD [1] provides (ε, δ)-DP guarantees via per-sample gradient clipping and Gaussian noise addition. Subsequent work has improved the utility-privacy trade-off through better accounting [4,26], larger batch training [7], and architectural choices [28]. DP-Adam [20] applies adaptive learning rates to privatized gradients but has shown limited benefit at moderate privacy budgets in our experiments, consistent with findings that adaptive methods require careful tuning under DP [6].

Noise reduction via spectral and temporal filtering. A growing line of work reduces the impact of DP noise through spectral or temporal post-processing. Spectral-DP [12] adds calibrated noise in the frequency domain rather than the spatial domain and applies spectral filtering, achieving the same (ε, δ)-DP guarantee with lower effective perturbation. DOPPLER [41] treats the privatized gradient sequence as a noisy signal and applies a low-pass filter to suppress high-frequency noise components while preserving low-frequency gradient information, yielding three to ten percent accuracy improvements across datasets and models. DiSK [40] extends this direction with a simplified Kalman filter that adapts its gain over training, achieving best-reported from-scratch results on CIFAR-100 and ImageNet-1k using augmentation multiplicity and architecture-specific tuning. Other approaches include frequency-domain gradient computation (GReDP [34]) and joint spectral-temporal filtering (FFTKF [38]).

These methods modify the privacy mechanism or its inputs. Spectral-DP and GReDP shape where noise enters by operating in the frequency domain, while DOPPLER and DiSK filter the privatized gradient across the training trajectory. Anisotropic Gaussian mechanisms [17] take a related route by reshaping the noise covariance at injection, which requires estimating that covariance from public data or a separate privacy budget and modifying the accounting. Spectral orthogonalization differs on the axis that is central to this work. It leaves the standard isotropic mechanism untouched and applies only deterministic post-processing to the already-privatized gradient, recovering directional signal from its *low-rank matrix structure* at zero additional privacy cost, with no covariance estimation and no change to the accounting. Because it targets matrix structure rather than component-wise magnitudes, it is complementary to the frequency-domain and temporal methods above (Sec. 4.3).

Parameter-efficient private fine-tuning. Parameter-efficient methods (e.g., DP-BiTFiT [5], low-rank reparametrization [37], and LoRA [16,23]) reduce trainable parameters to lower per-parameter noise. In this regime, all optimizers achieve similar accuracy, consistent with the finding that low-rank constraints remove the spectral mismatch that orthogonalization addresses in full-parameter training.

Gradient structure and low-rank methods. Neural network gradients exhibit low-rank structure [2, 22]: most gradient energy concentrates in a small subspace whose dimension correlates with task complexity. GaLore [42] projects gradients

into a low-rank subspace for memory efficiency. Our work differs in applying spectral processing as a *post-processing* step on the already-privatized gradient, preserving the DP guarantee by the post-processing theorem [10].

Spectral methods in optimization. Shampoo [14], SOAP [33], and Muon [18] exploit spectral structure for preconditioning or orthogonalization in non-private training. Their interaction with DP noise injection has not been studied; spectral gradient orthogonalization is investigated here in the DP setting, identifying the batch-size regime in which it provides a benefit.

3 Method

3.1 Preliminaries

Differential Privacy. A randomized mechanism $\mathcal{M}$ satisfies (ε, δ)-differential privacy [9] if for all adjacent datasets D, D' differing in a single record and all measurable sets S: $\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta$.

DP-SGD. Each per-sample gradient g_i is clipped to bound its Frobenius norm: $\bar{g}_i = g_i \cdot \min(1, C/\|g_i\|_F)$, where C is the clipping threshold. The clipped gradients are averaged and perturbed:

$$\tilde{G} = \frac{1}{B}\left(\sum_{i=1}^{B} \bar{g}_i + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right), \tag{1}$$

where σ is the noise multiplier calibrated via Rényi DP accounting [26].

Post-Processing Theorem. Any function of the output of a differentially private mechanism preserves the same (ε, δ) guarantee [10]. This is the key property enabling our approach: all spectral processing is applied *after* noise addition.

3.2 Spectral Gradient Orthogonalization

For a 2D parameter tensor, the method proceeds as:

Step 0: DP-SGD Gradient. Per-sample gradients are clipped and noised exactly as in standard DP-SGD (Eq. (1)), producing the privatized gradient $\tilde{G} \in \mathbb{R}^{m \times n}$. All subsequent steps operate on $\tilde{G}$ and are therefore post-processing.

Step 1: Momentum Accumulation. Following the Muon formulation [18], EMA-style Nesterov momentum is applied to the privatized gradient $\tilde{G}$: $M_t = \beta M_{t-1} + (1-\beta)\tilde{G}_t$, $\hat{M}_t = (1-\beta)\tilde{G}_t + \beta M_t$, where β is the momentum coefficient and $\hat{M}_t$ is the Nesterov lookahead. The $(1-\beta)$ scaling produces a weighted average rather than an accumulation, following the original Muon implementation.

Step 2: Orthogonalization via Newton-Schulz. The momentum matrix is orthogonalized using five iterations of a quintic Newton-Schulz scheme [18]:

$$X_0 = \hat{M}_t / \|\hat{M}_t\|_F, \quad X_{k+1} = aX_k + bX_k X_k^\top X_k + cX_k (X_k X_k^\top)^2, \tag{2}$$

where $(a, b, c) = (3.4445, -4.7750, 2.0315)$ are coefficients optimized for convergence speed. Frobenius-norm initialization ensures all singular values of X_0 are at most one (since $\|X_0\|_2 \leq \|X_0\|_F = 1$), which is sufficient for Newton-Schulz convergence. The output $X_K \approx UV^\top$, the closest orthogonal matrix to $\hat{M}_t$. The iterations require only matrix multiplications. The resulting wall-clock and memory overhead is modest and is measured per configuration in supp material.

Step 3: Dimension Scaling. A correction factor $\hat{G} = \sqrt{\max(1, m/n)} \cdot X_K$ accounts for rectangular gradients.[1] The parameter update is $W_{t+1} = W_t - \eta\hat{G}$. For 1D parameters (biases, layer norms), standard DP-SGD with momentum is used.

3.3 Magnitude-Preserving Variant and Privacy

Pure orthogonalization discards all singular value information. At higher SNR the leading singular value carries a meaningful magnitude signal, so a variant rescales the orthogonal update:

$$\hat{G}_{\text{scaled}} = \sigma_1(M_t) \cdot \sqrt{\max\Big(1, \frac{m}{n}\Big)} \cdot X_K, \tag{3}$$

where $M_t = \beta M_{t-1} + \tilde{G}_t$ is a heavy-ball momentum buffer (without Nesterov lookahead), X_K is the Newton-Schulz output applied to M_t, and $\sigma_1(M_t)$ is its spectral norm. Since $\sigma_1(M_t)$ is computed from the noisy buffer, it overestimates the clean magnitude by up to $\|N\|_2$, a bias that diminishes as the SNR increases. Both variants apply only deterministic post-processing to the Gaussian mechanism output (Eq. (1)); by the post-processing theorem, the privacy guarantee is identical to DP-SGD with no additional cost.

In practice, the magnitude-preserving variant is recommended only at moderate-to-high SNR, where the spectral-norm estimate $\sigma_1(M_t)$ is reliable. At low SNR magnitude bias dominates, and pure orthogonalization is preferable. It is also not recommended on SVHN where it shows unstable bimodal convergence.

4 Why Orthogonalization Helps at Large Batch Sizes

4.1 Signal-to-Noise Ratio in DP Gradients

The noisy gradient (1) decomposes as $\tilde{G} = G + N$, where $G = \frac{1}{B}\sum_i \bar{g}_i$ is the clipped average and $N \sim \mathcal{N}(0, \frac{\sigma^2 C^2}{B^2}\mathbf{I})$ is the privacy noise. The noise N is i.i.d.

[1] We cap scaling factor to 4.0 to prevent instability for highly rectangular matrices.

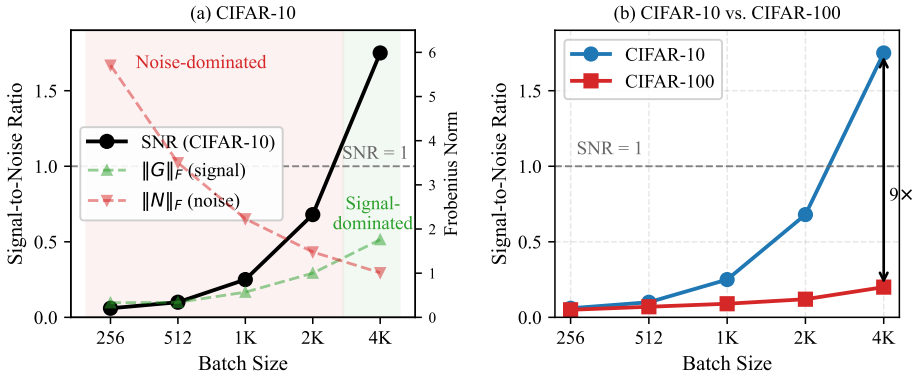

Fig. 2: Gradient SNR vs. batch size at $\varepsilon = 4$. *(a)* CIFAR-10 SNR with signal and noise Frobenius norms; the horizontal dashed line marks SNR = 1, above which orthogonalization recovers useful gradient structure. *(b)* CIFAR-10 vs. CIFAR-100: the SNR gap widens with batch size, reaching $\sim 9\times$ at $B = 4096$.

Gaussian by construction of the Gaussian mechanism [1, 9] and is independent of the data; clipping affects the signal G but not the noise distribution. The per-element noise standard deviation scales as $\sigma C/B$: *larger batches reduce the noise magnitude relative to the signal.*

The signal G is approximately low-rank, with effective rank [29] $\mathrm{erank}(G) = (\sum_i \sigma_i(G))^2 / \sum_i \sigma_i(G)^2 \ll \min(m, n)$. After adding DP noise, the effective rank increases dramatically (*e.g.*, from ~ 11 to ~ 197 for WRN-16-4 at $\varepsilon = 4$), as noise fills all singular directions.

Tab. 1 and Fig. 2 report measured SNR values across batch sizes for WRN-16-4 at $\varepsilon = 4$. On CIFAR-10, the SNR increases from 0.06 at $B = 256$ to 1.75 at $B = 4096$, crossing 1.0 between $B = 2048$ and $B = 4096$. On CIFAR-100, the SNR is substantially lower at every batch size (0.20 vs 1.75 at $B = 4096$; 0.12 vs 0.68 at $B = 2048$), because the gradient signal norm *decreases* with batch size on this harder task rather than growing as on CIFAR-10. The Frobenius SNR is a conservative proxy for the spectral recovery condition: since $\|N\|_F / \|N\|_2 \approx \sqrt{\min(m, n)}$, a Frobenius SNR of 1 implies $\sigma_1(G) > \|N\|_2$ by a factor of $\sqrt{\min(m, n)}$, so the actual recovery threshold is reached at smaller batch sizes than the SNR = 1 line suggests in Fig. 2. In particular, orthogonalization can recover the leading singular direction even when the Frobenius SNR is below 1, provided the gradient signal is concentrated in a spiked leading direction whose magnitude exceeds the noise operator norm $\|N\|_2$.

4.2 Spectral Recovery Threshold

The spectral-gap and SNR measurements used in this section (Tab. 1) are obtained from a separate post-hoc analysis run. They are not released and are not used during training, hyperparameter selection, or model release. Deployed mod-

Table 1: Measured gradient SNR ($\|G\|_F/\|N\|_F$) for WRN-16-4 at $\varepsilon = 4$, averaged over the first epoch. σ is the noise multiplier (identical for both datasets at $N = 50$K). Signal and noise norms are shown for CIFAR-10. As training progresses, gradient structure typically becomes more low-rank, which lowers the recovery threshold.

B	σ	$\|G\|_F$ (signal)	$\|N\|_F$ (noise)	SNR (C-10)	SNR (C-100)
256	0.88	0.33	5.71	0.06	0.05
512	1.08	0.34	3.50	0.10	0.07
1024	1.38	0.57	2.23	0.25	0.09
2048	1.83	1.00	1.48	0.68	0.12
4096	2.49	1.76	1.01	**1.75**	0.20

els are accounted for under standard DP-SGD, with no contribution from these statistics. The recovery threshold below is therefore an explanatory framework, not a private estimator for selecting batch size or enabling orthogonalization.

Orthogonalization replaces $\tilde{G}$ with $UV^\top$, projecting onto the Stiefel manifold [11]. This transformation preserves directions but normalizes magnitudes. The effect depends on the SNR:

Low SNR (small batch, small ε): The noise dominates. $UV^\top$ reflects the singular structure of the noise rather than the signal. Orthogonalization amplifies noisy directions to unit norm, providing no benefit.

High SNR (large batch, large ε): The signal's top singular directions persist through the noise. By the Wedin *sin* θ theorem [35], the angle θ between the clean and noisy leading singular vectors satisfies $\sin\theta \leq \|N\|_2/(\sigma_1(G) - \sigma_2(G))$. When the spectral gap exceeds the noise magnitude, orthogonalization approximately recovers the signal directions.

The observed threshold behavior is conceptually analogous to the phase transitions studied in spiked matrix models [3], where signal recovery occurs only when the dominant singular value exceeds the noise operator norm. The Wedin $\sin\theta$ bound provides a rigorous recovery condition without requiring the distributional assumptions of the spiked model.

Proposition 1 (Recovery condition). *Let $\tilde{G} = G + N \in \mathbb{R}^{m\times n}$ with $m \leq n$, where G has singular values $\sigma_1(G) \geq \sigma_2(G) \geq 0$ and N has i.i.d. entries with variance σ^2C^2/B^2. By the Wedin $\sin\theta$ bound [35], the angle θ between the leading left singular vectors of $\tilde{G}$ and G satisfies:*

$$\sin\theta \leq \frac{\|N\|_2}{\sigma_1(G) - \sigma_2(G)}. \tag{4}$$

Since $\|N\|_2 \leq \sigma C(\sqrt{m} + \sqrt{n})/B$ with high probability [31], orthogonalization recovers the signal subspace when:

$$B > \frac{\sigma C\,(\sqrt{m} + \sqrt{n})}{\sigma_1(G) - \sigma_2(G)}. \tag{5}$$

The recovery condition is derived by applying the Wedin $\sin\theta$ bound [32, 35] to the scaling laws of the Gaussian mechanism; a step-by-step derivation linking each bound to its source theorem is provided in the supplementary material, Section 1. The spectral gap $\sigma_1(G) - \sigma_2(G)$ is measured at epoch 1, the hardest regime for recovery since gradient structure becomes more low-rank as training progresses [22]; the threshold is therefore conservative. Equation (5) yields a minimum batch size that depends on the noise multiplier σ, the clipping threshold C, the layer dimensions m, n, and the spectral gap of the clean gradient. Datasets with weaker gradient signal (*i.e.* smaller spectral gap) require proportionally larger batches, consistent with the $\sim 9\times$ SNR gap between CIFAR-10 and CIFAR-100 (Tab. 1) and the corresponding shift in the crossover batch size observed in Sec. 5.2. Proposition 1 provides a necessary condition for leading singular vector recovery; full subspace recovery requires analogous gaps at each rank, which is verified empirically via the effective rank measurements in Tab. 1.

Numerical validation. Plugging measured spectral gaps from ResNet-18 [15] (epoch 1, CIFAR-10) into Eq. (5) with the noise multipliers from Tab. 1, 62% of layers satisfy the recovery condition at $B = 256$ and 100% by $B = 1024$. The network-wide bottleneck is the deepest convolutional layer (512×4608, spectral gap $\Delta\sigma \approx 0.12$, $\sqrt{m} + \sqrt{n} \approx 90$), which transitions between $B = 512$ and $B = 1024$. This aligns with the empirical crossover for ResNet-18 in Fig. 6(b), where DP-Muon already leads DP-SGD by +5.8% at $B = 256$ (when most layers have recovered) and the gap widens steadily with batch size. For the wider WRN-16-4, increased layer dimensionality raises $\sqrt{m} + \sqrt{n}$ and pushes the predicted B^* higher, consistent with the observed crossover at $B \approx 2048$.

This aligns with the observed phase transition (Tab. 1): at $B \leq 512$ (SNR $<$ 0.1) orthogonalization amplifies noise; at $B = 4096$ (SNR $= 1.75$) the signal dominates and improvements are consistent. DP-SGD stagnates because it uses the noisy gradient as-is, gaining little from improved directional structure; orthogonalization converts that structure into a cleaner update, explaining why spectral methods scale faster with both B and ε (Fig. 4).

4.3 Comparison with Temporal Filtering

Recent methods such as DOPPLER [41] and DiSK [40] reduce DP noise by applying temporal filters to the sequence of privatized gradients, also as post-processing steps incurring no additional privacy cost. The two classes of methods scale differently with batch size: temporal filtering provides roughly constant noise reduction determined by the filter coefficient β, largely independent of batch size. Spectral orthogonalization, by contrast, depends on the per-step SNR and grows more effective with batch size as the spectral gap exceeds the noise operator norm (Sec. 4). At small batch sizes temporal filtering dominates; at large batch sizes spectral orthogonalization catches up and eventually dominates. Since the two target different noise structures (temporal averaging vs. spatial structure), they are complementary and can be combined (Sec. 5.4).

Table 2: Test accuracy (%) on CIFAR-10 at $\varepsilon = 4$. Mean ± std over three seeds. Best per column in **bold**.

Method	$B=256$	$B=512$	$B=1024$	$B=2048$	$B=4096$
DP-SGD	39.4 ± 1.2	39.4 ± 0.7	39.7 ± 2.6	37.0 ± 2.6	39.3 ± 0.3
DP-Adam	$\mathbf{39.9 \pm 2.9}$	41.0 ± 1.1	41.8 ± 1.8	42.0 ± 0.5	38.8 ± 1.7
DP-Muon	39.0 ± 0.5	40.5 ± 1.7	43.3 ± 1.6	47.3 ± 1.8	$\mathbf{48.7 \pm 0.6}$
DP-Muon-S	38.7 ± 1.0	$\mathbf{44.0 \pm 1.9}$	$\mathbf{44.5 \pm 2.9}$	$\mathbf{47.8 \pm 1.7}$	46.3 ± 0.8

5 Experiments

5.1 Experimental Setup

Architecture and Datasets. Wide ResNet (WRN-16-4) [39] is used for all from-scratch experiments. Three datasets are evaluated: CIFAR-10 and CIFAR-100 [21] (50K training samples each) and SVHN [27] (73K training samples). Per-sample Frobenius norm clipping with $C = 1.0$ and $\delta = 10^{-5}$ is used throughout. The noise multiplier σ is calibrated via Rényi DP accounting [26, 36] to achieve the target ε. Training epochs are held fixed across batch sizes; larger batches therefore take fewer gradient steps per epoch.

Methods. Eight methods are compared, all sharing the same per-sample clipping and noise addition: (i) **DP-SGD** (momentum $\beta = 0.9$, LR $= 0.3$); (ii) **DP-Adam** [20] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, LR $= 0.001$); (iii) **DP-Muon** ($UV^\top$ orthogonalization, Nesterov $\beta = 0.95$, LR $= 0.02$); (iv) **DP-Muon-S** ($\sigma_1 \cdot UV^\top$, Eq. (3), heavy-ball $\beta = 0.9$, LR $= 0.3$); (v) **DiSK-SGD** [40] (Kalman filter, $\kappa = 1.5$, on DP-SGD); (vi) **DiSK-Muon** (Kalman filter on DP-Muon); (vii) **DOPPLER-SGD** [41] (EMA low-pass on DP-SGD); (viii) **DOPPLER-Muon** (EMA low-pass on DP-Muon). All methods use cosine learning rate decay with linear warmup. Results report best test accuracy (%) across epochs, averaged over three seeds (42, 43, 44) with standard deviation. DP-Muon uses $\beta = 0.95$ following the original Muon implementation [18]; the orthogonalization step, not the momentum coefficient, drives the gain, as DP-Muon-S uses identical $\beta = 0.9$ heavy-ball momentum to DP-SGD yet achieves comparable improvements.

5.2 Batch Size Sweep

Batch size is varied at a fixed privacy budget of $\varepsilon = 4$. Tabs. 2 and 3 report results on CIFAR-10/100; Fig. 3 visualizes the trends across all three datasets. **CIFAR-10:** At $B = 256$, all methods perform within 1.3% of each other, consistent with the low-SNR regime (SNR = 0.06, Tab. 1). DP-SGD remains flat across batch sizes (37.0 to 39.7%), while spectral methods scale upward: DP-Muon reaches 48.7% at $B = 4096$ (+9.4%). DP-Muon-S peaks at $B = 2048$ (47.8%) and declines at $B = 4096$, suggesting magnitude preservation is less beneficial at high SNR. DP-Adam shows no batch-size scaling [6].

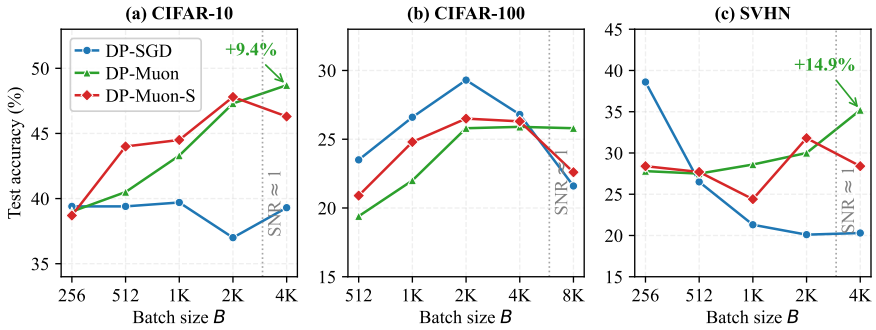

Fig. 3: Test accuracy vs. batch size at $\varepsilon = 4$ across three datasets. *(a)* CIFAR-10: DP-SGD is flat while spectral methods scale upward (+9.4% at $B = 4096$). *(b)* CIFAR-100: the crossover shifts to larger batch sizes, consistent with the lower SNR. *(c)* SVHN: DP-SGD degrades with batch size while DP-Muon improves (+14.9% at $B = 4096$). Vertical dotted lines mark the approximate SNR ≈ 1 threshold. Mean over three seeds.

Table 3: Test accuracy (%) on CIFAR-100 at $\varepsilon = 4$ with tuned learning rates (DP-SGD $\eta = 1.0$, DP-Adam $\eta = 0.01$, DP-Muon $\eta = 0.1$, DP-Muon-S $\eta = 0.5$). Mean $\pm$ std over three seeds. Best per column in **bold**.

Method	$B = 512$	$B = 1024$	$B = 2048$	$B = 4096$	$B = 8192$
DP-SGD	23.5 ± 1.0	26.6 ± 2.6	$\mathbf{29.3 \pm 1.2}$	$\mathbf{26.8 \pm 0.6}$	21.6 ± 0.9
DP-Adam	$\mathbf{25.5 \pm 1.3}$	$\mathbf{27.4 \pm 1.8}$	28.9 ± 0.7	26.0 ± 0.7	18.6 ± 1.2
DP-Muon	19.4 ± 1.7	22.0 ± 1.0	25.8 ± 0.7	25.9 ± 0.5	$\mathbf{25.8 \pm 0.6}$
DP-Muon-S	20.9 ± 1.2	24.8 ± 0.3	26.5 ± 0.7	26.3 ± 0.9	22.6 ± 1.0

CIFAR-100: The gradient SNR is $\sim 9\times$ lower than on CIFAR-10 (Tab. 1), so Proposition 1 predicts a proportionally larger batch is needed. At $B \leq 1024$, DP-Muon trails DP-SGD by 3 to 5%: in this low-SNR regime, orthogonalization replaces a noisy but directionally biased gradient with a nearly random orthogonal update, which is strictly worse (as formalized by the recovery condition of Proposition 1). Direction recovery (Fig. 1c) exceeds 0.99 only at $B = 4096$ (vs. $B = 1024$ on CIFAR-10), yet the accuracy crossover lags because orthogonalization normalizes noise-only directions (the Marchenko–Pastur bulk [25]) to unit norm alongside recovered signal directions. With tuned learning rates, DP-SGD peaks at $B = 2048$ (29.3%) and declines to 21.6% at $B = 8192$, while DP-Muon increases steadily to 25.8%, overtaking by +4.2% at the largest batch size.

SVHN: DP-SGD accuracy decreases monotonically (Fig. 3c) from 38.6% at $B = 256$ to 20.3% at $B = 4096$. DP-Muon increases by 14.9% from 27.8% to 35.2% at the largest batch size. We omit DP-Muon-S due to bimodal convergence.

Cross-Dataset Summary: The crossover batch size shifts with SNR: $B \approx 1024$ on CIFAR-10, $B \approx 4096$ to 8192 on CIFAR-100, and $B \approx 256$ on SVHN, consistent with Proposition 1. DiSK [40] reports 42.0% on CIFAR-100 ($\varepsilon = 8$)

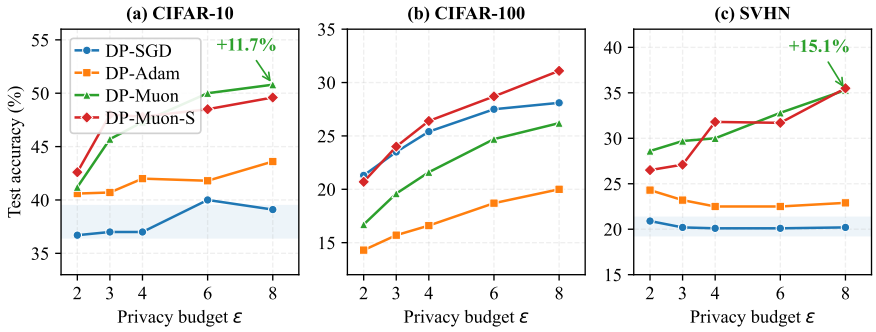

Fig. 4: Test accuracy vs. ε at $B = 2048$ across three datasets. DP-SGD stagnates on CIFAR-10 and SVHN (shaded bands) while spectral methods scale with the privacy budget. DP-Muon-S is strongest on CIFAR-100 at all ε. Mean over three seeds.

using augmentation multiplicity and WRN-28-10; isolating the interaction with spectral processing is deferred to future work.

5.3 Privacy Budget Sweep

Figure 4 shows accuracy at $B = 2048$ across privacy budgets $\varepsilon \in \{2, 3, 4, 6, 8\}$. DP-Muon accuracy scales steadily with the privacy budget (from 41.2% at $\varepsilon = 2$ to 50.8% at $\varepsilon = 8$ on CIFAR-10), while DP-SGD remains largely flat (36.7 to 40.0%). On CIFAR-100, DP-Muon-S outperforms DP-SGD at $\varepsilon \geq 3$, with the advantage growing to +3.0% at $\varepsilon = 8$. On SVHN, DP-SGD is flat at ∼20% regardless of ε, while DP-Muon improves from 28.6% to 35.3%.

5.4 Comparison with Temporal Denoising

Recent temporal denoising methods, DiSK [40] (Kalman filtering) and DOPPLER [41] (low-pass filtering), reduce DP noise by filtering the sequence of privatized gradients across training steps. As discussed in Sec. 4.3, temporal and spectral filtering target different noise reduction mechanisms and are in principle complementary.

Table 4 tests this hypothesis on CIFAR-10 at $\varepsilon = 4$; Fig. 5 visualizes the crossover. At $B \leq 512$, temporal filtering dominates: DiSK-SGD reaches 49.0% at $B = 256$, outperforming all spectral methods by 9%. At $B \geq 2048$, the pattern reverses: DOPPLER-Muon achieves 50.3% at $B = 4096$, the highest accuracy in any configuration. DiSK-SGD declines from 49.0% to 39.9% as fewer steps reduce the filter's averaging window; DiSK-Muon outperforms DiSK-SGD above $B = 512$, confirming that spectral processing improves temporal filtering once SNR is sufficient.

5.5 Fine-Tuning

In the parameter-efficient regime (LoRA, rank 8), all methods achieve 85.6 to 86.8%, a spread of 1.2%, confirming that the spectral benefit is specific to full-

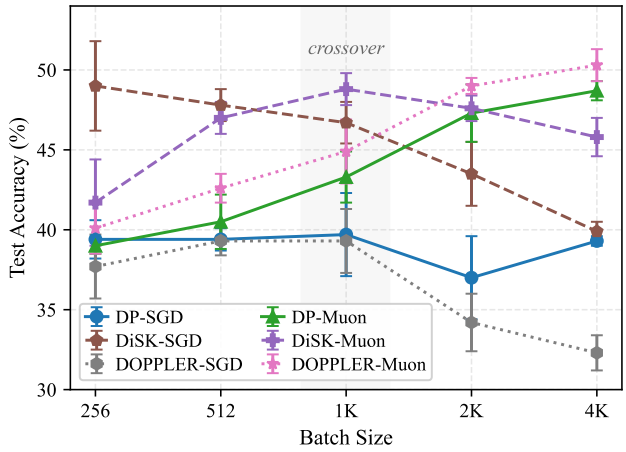

Fig. 5: Temporal vs. spectral denoising on CIFAR-10 at $\varepsilon = 4$. Temporal methods (DiSK-SGD, dashed brown) dominate at small batch sizes but decline as fewer gradient steps reduce the filter's averaging window. Spectral methods (DP-Muon, solid green) improve with batch size as per-step SNR increases. DOPPLER-Muon (dotted pink) combines both mechanisms and reaches the highest accuracy (50.3%) at $B = 4096$. Error bars show ± 1 std over three seeds.

parameter training. To test whether the benefit extends to full fine-tuning, all parameters of ViT-Small [8] (pretrained on ImageNet-21k [30]) are unfrozen and fine-tuned on CIFAR-100 at $\varepsilon = 4$.

DP-Muon matches DP-Adam within 0.4% at both batch sizes, while DP-SGD collapses from 86.9% to 79.5% at $B = 4096$. DP-Muon achieves this parity as a stateless post-processing step with no second-moment buffer, whereas DP-Adam doubles optimizer memory. That DP-SGD collapses while both remain stable indicates gradient processing is essential for fine-tuning at scale.

5.6 Architecture Diversity

To verify that the spectral benefit generalizes beyond WRN-16-4, the batch sweep is repeated on ResNet-18 and WRN-28-10 (CIFAR-10, $\varepsilon = 4$).

On ResNet-18, DP-Muon leads from $B = 256$ (+5.8%), consistent with Eq. (5): narrower layers yield smaller $\sqrt{m} + \sqrt{n}$, crossing the recovery threshold earlier. The gap grows to +14.9% at $B = 4096$. On WRN-28-10, the standard architecture in the DP optimization literature, the gap is even larger: DP-Muon reaches 56.2% at $B = 2048$ (+17.5%) and 54.9% at $B = 4096$ (+20.9% over DP-SGD). DP-SGD accuracy on WRN-28-10 monotonically decreases from 44.1% at $B = 512$ to 34.0% at $B = 4096$, as the spectral mismatch grows with parameter count. DP-Muon improves through $B = 2048$, confirming that orthogonalization is most effective when this mismatch is most severe.

Table 4: Temporal vs. spectral denoising on CIFAR-10 ($\varepsilon = 4$). DiSK-SGD applies the Kalman filter [40] ($\kappa = 1.5$) to DP-SGD. DOPPLER-Muon combines low-pass filtering [41] with spectral orthogonalization. Mean ± std over three seeds.

Denoising	Method	$B=256$	$B=512$	$B=1024$	$B=2048$	$B=4096$
None	DP-SGD	39.4 ± 1.2	39.4 ± 0.7	39.7 ± 2.6	37.0 ± 2.6	39.3 ± 0.3
Temporal	DiSK-SGD	$\mathbf{49.0 \pm 2.8}$	$\mathbf{47.8 \pm 1.0}$	46.7 ± 1.3	43.5 ± 2.0	39.9 ± 0.6
Temporal	DOPPLER-SGD	37.7 ± 2.0	39.3 ± 0.9	39.3 ± 2.0	34.2 ± 1.8	32.3 ± 1.1
Spectral	DP-Muon	39.0 ± 0.5	40.5 ± 1.7	43.3 ± 1.6	47.3 ± 1.8	48.7 ± 0.6
Both	DiSK-Muon	41.7 ± 2.7	47.0 ± 1.0	$\mathbf{48.8 \pm 1.0}$	47.6 ± 0.8	45.8 ± 1.2
Both	DOPPLER-Muon	40.1 ± 1.6	42.6 ± 0.9	44.9 ± 2.0	$\mathbf{49.0 \pm 0.5}$	$\mathbf{50.3 \pm 1.0}$

Table 5: Two-point settings, best per column in **bold**. *(Left)* Full fine-tuning of ViT-Small on CIFAR-100 (ε=4), mean ± std over three seeds. *(Right)* From-scratch NF-ResNet-50 on ImageNet-1k (B=32,768, 120 epochs, K=4, three seeds).

Method	LR	$B=2048$	$B=4096$
DP-SGD	0.01	86.9 ± 1.1	79.5 ± 3.4
DP-Adam	0.0003	$\mathbf{89.4 \pm 0.3}$	$\mathbf{89.7 \pm 0.1}$
DP-Muon	0.002	89.1 ± 0.2	89.3 ± 0.3

Method	ε=4	ε=8
DP-SGD	18.47	24.60
DP-AdamW [24]	10.37	18.64
DP-Muon	**20.97**	**27.32**

5.7 Beyond CIFAR Resolution

Figure 6(a) repeats the batch sweep on Tiny-ImageNet (64×64, 200 classes, 100K images, WRN-16-4, $\varepsilon = 4$). DP-SGD degrades from 21.1% to 17.3% while spectral methods improve (DP-Muon-S: 21.3% at $B = 4096$, +4.0%), with crossover between $B = 1024$ and 2048, consistent with Proposition 1. Table 5 extends the from-scratch evaluation to full ImageNet-1k scale (224×224, 1000 classes) with NF-ResNet-50, under an identical protocol for all optimizers. DP-Muon improves over DP-SGD by +2.50% at $\varepsilon = 4$ and +2.72% at $\varepsilon = 8$, with seven- and three-fold lower across-seed variance. The spectral-recovery benefit therefore holds at standard scale, in the standalone setting without temporal denoising.

6 Conclusion

An SNR-governed phase transition determines the utility of spectral gradient processing in differentially private training. The recovery condition derived from the Wedin bound (Proposition 1) provides a predictive threshold for singular vector recovery from isotropic noise, validated quantitatively against per-layer spectral measurements. Below this threshold, orthogonalization is harmful. Above it, the benefit is consistent and grows with batch size, privacy budget, and model capacity. Empirical results across four architectures (WRN-16-4, ResNet-18, WRN-28-10, ViT-Small) and three resolutions confirm that while DP-SGD stagnates or

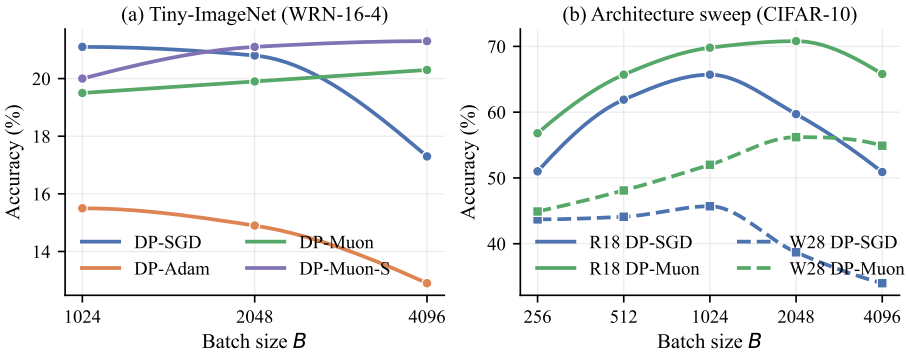

Fig. 6: Batch-size sweeps as smooth curves. *(a)* Tiny-ImageNet (WRN-16-4, ε=4). *(b)* Architecture sweep on CIFAR-10 (ε=4); ResNet-18 (solid), WRN-28-10 (dashed). Markers are measured points; curves are monotone cubic (PCHIP) interpolation.

degrades at scale, spectral orthogonalization enables continued scaling, achieving +20.9% over DP-SGD on WRN-28-10 ($B = 4096$) and +14.9% on ResNet-18. Complementarity with temporal filtering yields 50.3% on CIFAR-10 ($\varepsilon = 4$), the highest accuracy in any tested configuration.

Limitations and future work. DP-Muon-S exhibits bimodal convergence on SVHN due to noise-dominated σ_1 fluctuations (Sec. 3.3). Layer-wise selective orthogonalization and optimal singular value shrinkage [13] are natural extensions.

Acknowledgements This research was funded partially by the Australian Government through the Australian Research Council (ARC) Discovery Project DP250104538. The authors thank Clairva.ai for providing the GPU support used in this work.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: ACM Conf. Comput. Commun. Secur. pp. 308–318 (2016). https://doi.org/10.1145/2976749.2978318
2. Aghajanyan, A., Zettlemoyer, L., Gupta, S.: Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In: Assoc. Comput. Linguist. pp. 7319–7328 (2021)
3. Baik, J., Arous, G.B., Péché, S.: Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. Annals of Probability **33**(5), 1643–1697 (2005). https://doi.org/10.1214/009117905000000233
4. Balle, B., Barthe, G., Gaboardi, M.: Privacy profiles and amplification by subsampling. J. Privacy and Confidentiality **10**(1) (2020). https://doi.org/10.29012/jpc.726
5. Bu, Z., Wang, Y.X., Zha, S., Karypis, G.: Differentially private bias-term fine-tuning of foundation models. arXiv preprint arXiv:2210.00036 (2022)

6. Bu, Z., Wang, Y.X., Zha, S., Karypis, G.: Differentially private optimization on large model at small cost. In: ICML. vol. 202, pp. 3192–3218 (2023)
7. De, S., Berrada, L., Hayes, J., Smith, S.L., Balle, B.: Unlocking high-accuracy differentially private image classification through scale. In: NeurIPS (2022)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
9. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography Conf. pp. 265–284 (2006). https://doi.org/10.1007/11681878_14
10. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(3–4), 211–407 (2014). https://doi.org/10.1561/0400000042
11. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM Journal on Matrix Analysis and Applications **20**(2), 303–353 (1998). https://doi.org/10.1137/S0895479895290954
12. Feng, C., Xu, N., Wen, W., Venkitasubramaniam, P., Ding, C.: Spectral-DP: Differentially private deep learning through spectral perturbation and filtering. In: IEEE Symp. Security and Privacy. pp. 1944–1960 (2023). https://doi.org/10.1109/SP46215.2023.10179457
13. Gavish, M., Donoho, D.L.: The optimal hard threshold for singular values is $4/\sqrt{3}$. IEEE Trans. Inform. Theory **60**(8), 5040–5053 (2014). https://doi.org/10.1109/TIT.2014.2323359
14. Gupta, V., Koren, T., Singer, Y.: Shampoo: Preconditioned stochastic tensor optimization. In: ICML. pp. 1842–1850 (2018)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
17. Ji, T., Li, P.: Less is more: Revisiting the Gaussian mechanism for differential privacy. In: USENIX Security Symp. (2024)
18. Jordan, K.: Muon: An optimizer for hidden layers in neural networks (2024), blog post, https://kellerjordan.github.io/posts/muon/, accessed 2026-06-26
19. Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F.: Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence **2**(6), 305–311 (2020). https://doi.org/10.1038/s42256-020-0186-1
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
21. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009)
22. Li, C., Farkhoor, H., Liu, R., Yosinski, J.: Measuring the intrinsic dimension of objective landscapes. In: ICLR (2018)
23. Li, X., Tramèr, F., Liang, P., Hashimoto, T.: Large language models can be strong differentially private learners. In: ICLR (2022)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
25. Marchenko, V.A., Pastur, L.A.: Distribution of eigenvalues for some sets of random matrices. Matematicheskii Sbornik **114**(4), 507–536 (1967)
26. Mironov, I.: Rényi differential privacy. In: IEEE Comput. Secur. Found. pp. 263–275 (2017). https://doi.org/10.1109/CSF.2017.11

27. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011)
28. Ponomareva, N., Hazimeh, H., Kurakin, A., Xu, Z., Denison, C., McMahan, H.B., Vassilvitskii, S., Chien, S., Thakurta, A.G.: How to DP-fy ML: A practical guide to machine learning with differential privacy. J. Artificial Intelligence Res. **77**, 1113–1201 (2023). https://doi.org/10.1613/jair.1.14649
29. Roy, O., Vetterli, M.: The effective rank: A measure of effective dimensionality. In: Eur. Signal Process. Conf. pp. 606–610 (2007)
30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
31. Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. Cambridge University Press (2012). https://doi.org/10.1017/CBO9780511794308.006
32. Vu, V.Q., Lei, J.: Minimax sparse principal subspace estimation in high dimensions. The Annals of Statistics **41**(6), 2905–2947 (2013), earlier version: arXiv:1211.0373, 2012
33. Vyas, N., Morwani, D., Zhao, R., Shapira, I., Brandfonbrener, D., Janson, L., Kakade, S.: SOAP: Improving and stabilizing Shampoo using Adam. arXiv preprint arXiv:2409.11321 (2024)
34. Wang, H., Jiang, T., Guo, Y., Jia, X., Cai, C., Ding, C.: GReDP: A more robust approach for differential privacy training with gradient-preserving noise reduction. arXiv preprint arXiv:2409.11663 (2024)
35. Wedin, P.Å.: Perturbation bounds in connection with singular value decomposition. BIT Numerical Mathematics **12**(1), 99–111 (1972). https://doi.org/10.1007/BF01932678
36. Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., Mironov, I.: Opacus: User-friendly differential privacy library in PyTorch. arXiv preprint arXiv:2109.12298 (2021)
37. Yu, D., Zhang, H., Chen, W., Yin, J., Liu, T.Y.: Large scale private learning via low-rank reparametrization. In: ICML. pp. 12208–12218 (2021)
38. Yun, J.: FFTKF: Fast Fourier transform-based spectral and temporal gradient filtering for differentially private optimization. arXiv preprint arXiv:2505.04468 (2025)
39. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: BMVC (2016)
40. Zhang, X., Bu, Z., Balle, B., Hong, M., Razaviyayn, M., Mirrokni, V.: DiSK: Differentially private optimizer with simplified Kalman filter for noise reduction. In: ICLR (2025)
41. Zhang, X., Bu, Z., Hong, M., Razaviyayn, M.: DOPPLER: Differentially private optimizers with low-pass filter for privacy noise reduction. In: NeurIPS (2024)
42. Zhao, J., Zhang, Z., Chen, B., Wang, Z., Anandkumar, A., Tian, Y.: GaLore: Memory-efficient LLM training by gradient low-rank projection. arXiv preprint arXiv:2403.03507 (2024)