Gradient Clipping Beyond Vector Norms: A Spectral Approach for Matrix-Valued Parameters

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient clipping is a standard safeguard for training neural networks under noisy, heavy-tailed stochastic gradients; yet, most clipping rules treat all parameters as vectors and ignore the matrix structure of modern architectures. We show empirically that data outliers often amplify only a small number of leading singular values in layer-wise gradient matrices, while the rest of the spectrum remains largely unchanged. Motivated by this phenomenon, we propose *spectral clipping*, which stabilizes training by clamping singular values that exceed a threshold while preserving the singular directions. This framework generalizes classical gradient norm clipping and can be easily integrated into existing optimizers. We provide a convergence analysis for non-convex optimization with spectrally clipped SGD, yielding an $\mathcal{O}\left(K^{1-\alpha/3\alpha-2}\right)$ rate for the minimum expected gradient norm under heavy-tailed noise. We further prove that, on a heavy-tailed matrix quadratic, spectral clipping can attain a strictly smaller asymptotic loss constant than norm clipping. To minimize hyperparameter tuning, we introduce layer-wise adaptive thresholds based on moving averages or sliding-window quantiles of the top singular values. Finally, we develop efficient implementations that clip only the top r singular values via randomized truncated SVD, avoiding full decompositions for large layers. We demonstrate competitive performance across synthetic heavy-tailed settings and neural network training tasks.

1 INTRODUCTION

Optimization algorithms lie at the core of deep learning (LeCun et al., 2015; Choudhury et al., 2024; Mishchenko & Defazio, 2023; Shi et al., 2023). As the cost of training large-scale models (Achiam et al., 2023; Berahas et al., 2016) continues to grow, the demand for more effective and efficient optimization methods has become increasingly critical (Guo et al., 2025; Nguyen et al., 2017; Schaipp et al., 2023; Takáč et al., 2013; Li et al., 2022).

The parameters of modern deep learning models are predominantly matrix-valued (Vaswani et al., 2017; LeCun et al., 2015; He et al., 2016). Despite this, most existing optimization algorithms operate by implicitly vectorizing all parameters, thereby discarding intrinsic algebraic and geometric structures. Consequently, widely used algorithms such as Stochastic Gradient Descent (SGD) Robbins & Monro (1951), RMSProp (Tieleman, 2012), AdaDelta Zeiler (2012), Adam (Kingma, 2014), and AdamW (Loshchilov & Hutter, 2017; Bjorck et al., 2021) are fundamentally designed for vector-valued parameters. Recently, Jordan et al. (2024) advocated the use of different optimizers for matrix-valued parameters, enabling the exploitation of structural properties that are lost under vectorization. This line of work has revealed a rich and largely unexplored design space for optimization algorithms in deep learning (Shen et al., 2025; Chang et al., 2025; An et al., 2025; Pethick et al., 2025a; Shah et al., 2025; He et al., 2017; Choudhury et al., 2026).

Matrix-aware optimizers use parameter structure to determine update directions (Jordan et al., 2024; An et al., 2025; Gupta et al., 2018; Pethick et al., 2025a; Vyas et al., 2024). Standard gradient norm clipping (Mikolov et al., 2012; Pascanu et al., 2013) rescales the entire gradient by one factor. We study spectral clipping, which instead clamps singular values individually, and analyze its convergence and its possible advantage over norm clipping under heavy-tailed noise.

Specifically, we provide empirical evidence showing that the leading singular values of gradient matrices tend to increase in the presence of data noise. Motivated by this observation, we propose

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.

clipping the top singular values of matrix-valued gradients when they exceed a predefined threshold, thereby improving training stability.

1.1 Background and Related Works

Most deep learning tasks reduce to solving the minimization problem $\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x})$ and SGD (Robbins & Monro, 1951) is the most common approach for solving it. The update rule of SGD is given by $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \boldsymbol{g}_k$, where $\eta_k > 0$ is the step size and $\boldsymbol{g}_k$ is an unbiased estimator of the gradient of f at $\boldsymbol{x}_k$, i.e., $\mathbb{E}\left[\boldsymbol{g}_k \mid \boldsymbol{x}_k\right] = \nabla f(\boldsymbol{x}_k)$. There are many studies showing the convergence of SGD for solving convex and non-convex problems (Gorbunov et al., 2020b; Ghadimi & Lan, 2013). In practice, SGD works well when the distribution of gradient noise is light-tailed. To address rare but potentially very large stochastic gradients, one often replaces $\boldsymbol{g}_k$ by $\mathrm{clip}(\boldsymbol{g}_k, \tau) := \min\left\{1, \frac{\tau}{\|\boldsymbol{g}_k\|}\right\}\boldsymbol{g}_k$ (Koloskova et al., 2023), which yields the update

$$\boldsymbol{x}_{k+1} \quad = \quad \boldsymbol{x}_k - \eta_k \mathrm{clip}(\boldsymbol{g}_k, \tau). \tag{1}$$

Here $\tau > 0$ is the clipping threshold, which ensures that $\|\mathrm{clip}(\boldsymbol{g}_k, \tau)\| \leq \tau$ for all k. Gradient clipping was established as a practical remedy for exploding gradients in recurrent neural networks (Mikolov et al., 2012; Pascanu et al., 2013) and remains a standard stabilization tool in modern training pipelines. Beyond this classical motivation, a recent theoretical literature has clarified when clipping helps and what its limitations are. Koloskova et al. (2023) showed that clipped SGD introduces a threshold-dependent stochastic bias and provided tight convergence guarantees, while Zhang et al. (2020a); Qian et al. (2021) analyzed clipping-based methods under relaxed smoothness. Very recent work also extends norm clipping beyond the Euclidean setting: Pethick et al. (2025b) studies generalized gradient norm clipping under non-Euclidean (L_0, L_1)-smoothness and instantiates this viewpoint in Clipped Scion.

Concurrently, Jiang et al. (2026) propose SPECTRA, which mainly clips optimizer updates; spectral clipping of raw gradients is optional and off by default in their LLM experiments. As they also observe, noise inflates the top singular values of the gradient itself (Figure 1), so we clip the gradient before it reaches the optimizer (Appendix J). Their analyses assume finite variance and compare with global clipping via upper bounds (their Theorems 4.6 and 4.7); ours allow infinite variance, and Theorem 3.6 proves a strictly smaller asymptotic loss constant for spectral than for Frobenius clipping on a matrix quadratic with heavy-tailed spikes.

Additional related works about heavy tailed gradient noise and spectral computation could be found in Appendix A.

1.2 Main Contributions

The main contributions of this work are summarized below.

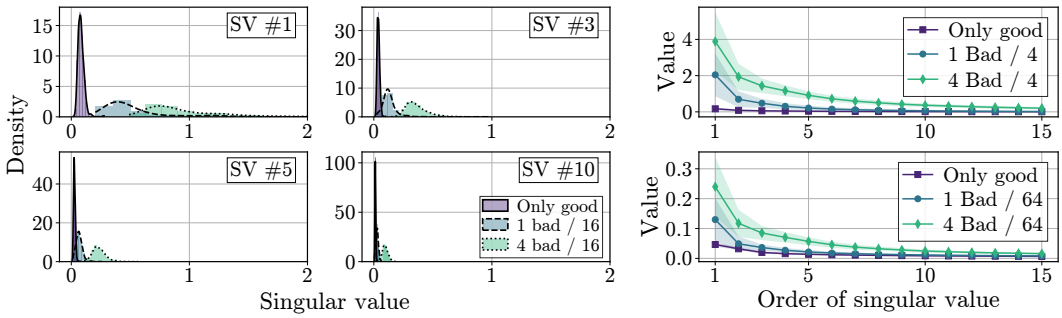
• **Spectral Gradient Clipping.** We study clipping individual singular values of gradient matrices (Section 2.2). The operator preserves singular directions and recovers standard norm clipping for column-vector gradients (Section 2.3).

• **Convergence Analysis.** Under a bounded centered α-moment assumption, Theorem 3.3 bounds nonconvex stationarity in terms of the clipping bias. With the specified horizon-dependent parameters, Theorem 3.4 gives a rate of $\mathcal{O}(K^{-2(\alpha-1)/(3\alpha-2)})$ for the stationarity potential and $\mathcal{O}(K^{-(\alpha-1)/(3\alpha-2)})$ for the minimum expected gradient norm. These guarantees apply to exact spectral clipping.

• **A Strict Comparison with Norm Clipping.** On a matrix quadratic with heavy-tailed rank-r noise in random directions, Theorem 3.6 shows that spectral clipping attains a strictly smaller asymptotic constant in the expected excess loss than Frobenius clipping.

• **Adaptive Clipping Strategies.** In Section 4, we provide adaptive strategies for selecting clipping thresholds τ_k (2) based on exponential moving averages (10) or quantiles (11) of the top singular values of gradient matrices. This allows us to tackle noise and adapt to gradient scales during training. We show the effectiveness of these strategies in experiments.

• **Efficient Implementation of Spectral Clipping.** For the efficient implementation of our proposed method to train large neural networks, we discuss two important aspects: (i) **Faster Implementation.** In Section 5.2, we describe how truncated SVD can be used to implement spectral clipping of gradient



(a) Singular value distributions for selected SVs

(b) Top-15 singular values

Figure 1: Spectral profiles of the first feed-forward matrix $W_{\text{MLP},1}$ in layer 5 of GPT-2 Small when one or four tokens of a clean batch ($\tau_{\text{good}} = 0.95$) are replaced by bad samples ($\tau_{\text{bad}} = 10^{-7}$). (a) Distributions of selected singular values at batch size 16. (b) Top-15 singular values at batch sizes 4 and 64. Bad samples shift the distribution and make it broader.

matrices. We introduce Algorithm 2, which is more efficient compared to implementing spectral clipping with full SVD and achieves similar final test accuracy in our CV experiment (Figure 7b, Appendix F). (ii) **Layerwise Adaptive Clipping Scheme.** In Section 5.3, we propose the layerwise adaptive clipping scheme, where we maintain different clipping thresholds for different layers of the neural network.

• **Numerical Experiments.** In Section 6, we present several numerical experiments that demonstrate the effectiveness of the proposed spectral clipping technique. In particular, we show that (i) SGDM with spectral clipping outperforms SGDM with norm clipping on a CV task with real label noise in Section 6.1; (ii) on the language modeling task in Section 6.2, spectral clipping outperforms norm clipping for SGDM, Adam and Muon; and (iii) SGDM with spectral clipping outperforms Adam in Section 6.3.

2 Gradient Clipping beyond Vector Norms

In this section, we first conduct an experiment in which we artificially add noise and check the influence of this noise on the spectrum of gradient matrices. Based on our observations in this experiment, we suggest a novel clipping technique in the second part. Finally, we show that our method is more general than the norm clipping method discussed in (1).

2.1 Singular Values of Gradient Matrices

Standard norm clipping treats gradients as vectors. To examine the relevance of matrix structure for clipping, we study the singular values of layer-wise gradient matrices. Here, we consider the pretrained GPT-2 model $p_\theta(\cdot \mid \cdot)$ (Radford et al., 2019). This model is already trained, and given a sequence of tokens $s_{[1:t-1]}$, it assigns a high probability to a *good* token; i.e., $p_\theta(s_t \mid s_{[1:t-1]})$ is high if s_t is good, and a lower probability to a *bad* token; i.e., $p_\theta(s_t \mid s_{[1:t-1]})$ is low if s_t is a bad token Sutskever et al. (2014). Now we implement backpropagation on p_θ to compute the gradients of the parameters for both the good and bad sets of tokens.

In Figure 1, we plot the distribution of the largest singular values of gradient matrices computed from mini-batches containing different numbers of bad tokens. When all tokens in the mini-batch are good, the singular values of the gradient matrix remain small. As the number of bad tokens in the mini-batch increases, the top singular values increase significantly.

We perform this analysis for each layer of GPT-2 and observe similar trends across all layers (see Appendix I for additional details). These observations lead us to consider a clipping technique that truncates the singular values of gradient matrices in order to address noise in modern deep learning.

2.2 Spectral Clipping of Gradient Matrices

Motivated by the empirical behavior observed in the previous section, we treat model parameters as matrices and clip the singular values of the corresponding stochastic gradient matrices. To elaborate on our technique, we consider the matrix-parameter optimization problem $\min_{\boldsymbol{X} \in \mathbb{R}^{m \times n}} f(\boldsymbol{X})$, where

$f : \mathbb{R}^{m\times n} \to \mathbb{R}$. To solve this problem, we propose the update rule

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \eta_k \mathcal{C}_{\tau_k}(\boldsymbol{G}_k). \tag{2}$$

where $\mathcal{C}_{\tau_k}$ is an operator that clamps the singular values of $\boldsymbol{G}_k$ to a threshold $\tau_k > 0$. For example, if the SVD of the gradient matrix is $\boldsymbol{G}_k = \boldsymbol{U}_k \mathrm{diag}(\boldsymbol{\sigma}_k)\boldsymbol{V}_k^\top$ (where $d := \min\{m, n\}, \boldsymbol{U}_k \in \mathbb{R}^{m\times d}, \boldsymbol{V}_k \in \mathbb{R}^{n\times d}$ and $\boldsymbol{\sigma}_k \in \mathbb{R}^d_+$ contain the singular values) then $\mathcal{C}_{\tau_k}(\boldsymbol{G}_k) := \boldsymbol{U}_k \mathrm{diag}(\mathrm{clamp}_\tau(\boldsymbol{\sigma}_k))\boldsymbol{V}_k^\top$, where the coordinates of the vector $\mathrm{clamp}_\tau(\boldsymbol{\sigma}_k)$ are given by

$$\mathrm{clamp}_{\tau_k}(\boldsymbol{\sigma}_k)[i] = \begin{cases} \boldsymbol{\sigma}_k[i] & \text{if } \boldsymbol{\sigma}_k[i] \le \tau_k \\ \tau_k & \text{otherwise.} \end{cases} \tag{3}$$

We call this technique spectral clipping. The operator $\mathcal{C}_{\tau_k}$ leaves the singular directions $\boldsymbol{U}_k, \boldsymbol{V}_k$ unchanged and ensures $\sigma_{\max}(\mathcal{C}_{\tau_k}(\boldsymbol{G}_k)) \le \tau_k$ and $\|\mathcal{C}_{\tau_k}(\boldsymbol{G}_k)\|_F \le \sqrt{d}\tau_k$. As we observe, the top singular values of the gradient matrices spike in the presence of noise; this clipping technique can help in practice.

2.3 Gradient Norm Clipping as a Special Case

We next show that standard gradient norm clipping is a special case of spectral clipping. Norm clipping treats parameters as vectors and updates using (1). Observe that any vector $\boldsymbol{g}_k \in \mathbb{R}^d$ can be viewed as a matrix in $\mathbb{R}^{d\times 1}$, and its SVD is $\boldsymbol{g}_k = \boldsymbol{U}_k \mathrm{diag}(\boldsymbol{\sigma}_k)\boldsymbol{V}_k^\top$, with $\boldsymbol{U}_k = \frac{\boldsymbol{g}_k}{\|\boldsymbol{g}_k\|} \in \mathbb{R}^d$, $\boldsymbol{\sigma}_k = \|\boldsymbol{g}_k\| \in \mathbb{R}$, and $\boldsymbol{V}_k = 1 \in \mathbb{R}$. Then from (3), we have $\mathrm{clamp}_{\tau_k}(\|\boldsymbol{g}_k\|) = \min\{\|\boldsymbol{g}_k\|, \tau_k\}$ and thus $\mathcal{C}_{\tau_k}(\boldsymbol{g}_k) = \frac{\boldsymbol{g}_k}{\|\boldsymbol{g}_k\|}\min\{\|\boldsymbol{g}_k\|, \tau_k\}$. Plugging this into (2) gives $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \min\{1, \tau_k/\|\boldsymbol{g}_k\|\}\boldsymbol{g}_k$, which coincides with the norm-clipping update (1) (with $\tau_k = \tau$). Therefore, spectral clipping generalizes gradient norm clipping, and *any analysis of our method captures that of gradient norm clipping as a special case*.

3 Convergence Guarantees

In this section, we establish convergence guarantees for SGD with spectral clipping (2). For this section, we consider only the constant spectral clipping threshold, i.e., $\tau_k = \tau$ in (2). We make some standard assumptions on the function f and the stochastic gradient matrix $\boldsymbol{G}_k$.

Assumption 3.1. We assume f is L-smooth, i.e., $f(\boldsymbol{X}) \le f(\boldsymbol{Y}) + \langle \nabla f(\boldsymbol{Y}), \boldsymbol{X} - \boldsymbol{Y}\rangle + L/2\|\boldsymbol{X} - \boldsymbol{Y}\|_F^2$ for all $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{m\times n}$.

This is a standard assumption in the optimization literature for the analysis of optimization algorithms (Nesterov, 2004; Gorbunov et al., 2020a; Koloskova et al., 2023). Moreover, we assume that the stochastic gradient estimates satisfy the bounded α-moment assumption (Zhang et al., 2020b; Hübler et al., 2024; Chezhegov et al., 2024; Choudhury et al., 2026), also called p-BCM.

Assumption 3.2. We assume the gradient $\boldsymbol{G}_k$ computed at each iteration k is unbiased i.e. $\mathbb{E}[\boldsymbol{G}_k \mid \boldsymbol{X}_k] = \nabla f(\boldsymbol{X}_k)$ and has bounded α-moment i.e. $\mathbb{E}[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F^\alpha \mid \boldsymbol{X}_k] \le \sigma^\alpha$ for some $\alpha \in (1, 2]$ and $\sigma > 0$.

For $\alpha = 2$, Assumption 3.2 reduces to the classical bounded-variance assumption (Lan, 2012; Dekel et al., 2012), and it is extensively used in optimization to study different stochastic algorithms. However, the bounded variance assumption can be restrictive in modern applications. Many recent empirical studies suggest that the stochastic gradient noise in image classification (Simsekli et al., 2019; Battash et al., 2024), large language models (Zhang et al., 2020b; Ahn et al., 2023), and reinforcement (Garg et al., 2021) tasks follows a heavy-tailed distribution. Assumption 3.2 accommodates such heavy-tailed behavior with $\alpha < 2$ (Zhang et al., 2020b; Mohammadi et al., 2015). Recent work has also started to analyze matrix-structured and sign-based optimizers under heavy-tailed noises. In particular, Yu et al. (2026) establish convergence guarantees for SignSGD (Bernstein et al., 2018), Lion (Chen et al., 2023) and Muon under a generalized heavy-tailed noise model. Under this assumption, we can show that $\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2 \mid \boldsymbol{X}_k\right] \le 2\|\nabla f(\boldsymbol{X}_k)\|_F^2 + 2^{4-\alpha} d^{1-\alpha/2}\tau^{2-\alpha}\sigma^\alpha$ (see Lemma C.1, Appendix C) for $d = \min\{m, n\}$ and use this to prove the following bound on the iterates of SGD with spectral clipping (2).

Theorem 3.3. Suppose Assumptions 3.1 and 3.2 hold, and let $f_* := \inf_{\boldsymbol{X}} f(\boldsymbol{X}) > -\infty$. Then for any $K \geq 1$, SGD with spectral clipping (2) and a constant step size $\eta_k = \eta := \frac{1}{4L\sqrt{K}}$ satisfies

$$\min_{0\leq k\leq K-1} \mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F^2\right] \leq \frac{16L(f(\boldsymbol{X}_0)-f_*)}{\sqrt{K}} + \frac{C_{\alpha,d}\tau^{2-\alpha}\sigma^{\alpha}}{2\sqrt{K}} + \frac{2}{K}\sum_{k=0}^{K-1}\Delta_k \tag{4}$$

where $\Delta_k := \mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k) - \mathbb{E}\left[\mathcal{C}_{\tau_k}(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\|_F^2\right]$ and $C_{\alpha,d} := 2^{4-\alpha}d^{1-\alpha/2}$ with $d = \min\{m,n\}$.

The first term in (4) is the standard optimization error for L-smooth objectives. The second term depends on $\tau^{2-\alpha}\sigma^{\alpha}$ and captures the effect of stochastic noise after clipping. For $\alpha < 2$, this term reflects how clipping controls the contribution of heavy-tailed gradient noise to the convergence rate. The last term is dependent on the clipping threshold τ and measures the bias introduced by the nonlinear operator $\mathcal{C}_{\tau_k}$. Using conditional unbiasedness, we can rewrite the bias as $\nabla f(\boldsymbol{X}_k) - \mathbb{E}[\mathcal{C}_{\tau_k}(\boldsymbol{G}_k) \mid \boldsymbol{X}_k] = \mathbb{E}[\boldsymbol{G}_k - \mathcal{C}_{\tau_k}(\boldsymbol{G}_k) \mid \boldsymbol{X}_k]$, so Δ_k is small whenever clipping is rarely active (i.e., $\mathcal{C}_{\tau_k}(\boldsymbol{G}_k) \approx \boldsymbol{G}_k$ with high probability). A similar phenomenon of bias is observed in SGD with gradient norm clipping Koloskova et al. (2023). When clipping is disabled by taking $\tau_k = \infty$, we have $\mathcal{C}_{\tau_k}(\boldsymbol{G}_k) = \boldsymbol{G}_k$ and hence $\Delta_k = \mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k) - \mathbb{E}\left[\boldsymbol{G}_k \mid \boldsymbol{X}_k\right]\|_F^2\right] = 0$.

For $\alpha = 2$, note that the second term on the right hand side of (4) becomes independent of the clipping threshold τ. Since clipping can introduce bias through Δ_k but does not improve the variance dependent second term of (4) when $\alpha = 2$, the best bound is obtained by disabling clipping ($\tau_k = \infty$), thereby recovering the optimal $\mathcal{O}(1/\sqrt{K})$ convergence rate for non-convex minimization with SGD (Robbins & Monro, 1951; Ghadimi & Lan, 2013). However, when gradients have heavy-tailed noise, i.e., $\alpha < 2$, we need clipping to control both the second and third terms of (4) and achieve a better rate of convergence. We provide a detailed proof of Theorem 3.3 in Appendix C.1.

The bias due to a finite clipping threshold in (4), i.e., $\frac{2}{K}\sum_{k=0}^{K-1}\Delta_k$, is unavoidable. However, we can choose the threshold τ depending on the noise level σ to prove convergence.

Theorem 3.4. Suppose Assumptions 3.1 and 3.2 hold, and let $f_* := \inf_{\boldsymbol{X}} f(\boldsymbol{X}) > -\infty$. Then, for any $K \geq 1$, SGD with spectral clipping (2), a constant step size $\eta_k = \eta \leq 1/4dL$ and clipping threshold $\tau \geq \max\{2, 8\sigma\}$ satisfy the following

$$\min_{0\leq k\leq K-1} \mathbb{E}\left[\phi\left(\|\nabla f(\boldsymbol{X}_k)\|_F\right)\right] \leq \frac{4(f(\boldsymbol{X}_0)-f_*)}{\eta K} + 32\tau^{-2(\alpha-1)}\sigma^{2\alpha} + 16\eta dL\tau^{2-\alpha}\sigma^{\alpha} \tag{5}$$

where $\phi(t) := \min\{t, t^2\}$ and $d = \min\{m,n\}$. Moreover, choose $\tau = \max\{2, 8\sigma, \sigma K^{1/(3\alpha-2)}\}$ and $\eta = \min\{1/4dL, 1/L\tau^{\alpha}\}$. For fixed $\alpha, \sigma > 0$, d, L, and $f(\boldsymbol{X}_0) - f_*$, we have

$$\min_{0\leq k\leq K-1} \mathbb{E}\left[\phi\left(\|\nabla f(\boldsymbol{X}_k)\|_F\right)\right] = \mathcal{O}\left(K^{\frac{-2\alpha+2}{3\alpha-2}}\right). \tag{6}$$

Since $\mathbb{E}[Z] \leq \sqrt{\mathbb{E}[\phi(Z)]} + \mathbb{E}[\phi(Z)]$ for any nonnegative random variable Z (split on $Z \leq 1$ and apply Cauchy–Schwarz), (6) implies $\min_{0\leq k<K} \mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] = \mathcal{O}\big(K^{-(\alpha-1)/(3\alpha-2)}\big)$, which matches the exponent of the lower bound of Zhang et al. (2020b, Theorem 6).

3.1 Spectral versus Norm Clipping under Low-Rank Noise

Theorems 3.3 and 3.4 hold for any noise satisfying Assumption 3.2 and also cover norm clipping (Section 2.3). Since such noise need not have any spectral structure, these worst-case bounds cannot show when spectral clipping is preferable. In Figure 1, however, corrupted samples inflate only a few leading singular values. To show the theoretical advantages of spectral clipping in this regime, we consider the following problem:

$$\min_{\boldsymbol{X}\in\mathbb{R}^{d\times d}} f(\boldsymbol{X}), \qquad f(\boldsymbol{X}) := f_* + \tfrac{1}{2}\left\langle \boldsymbol{X} - \boldsymbol{X}_*, \boldsymbol{H}(\boldsymbol{X} - \boldsymbol{X}_*)\right\rangle, \tag{7}$$

where $\boldsymbol{H}$ is symmetric positive definite with eigenvalues $h_1, \ldots, h_d$.

Assumption 3.5. Let $1 \leq r < d$, $a_0 > 0$ and $\kappa \in (1, 2)$. The stochastic gradient at iteration k is

$$\boldsymbol{G}_k = \nabla f(\boldsymbol{X}_k) + \boldsymbol{\Xi}_k, \qquad \boldsymbol{\Xi}_k := A_k \boldsymbol{U}_k \boldsymbol{V}_k^\top, \tag{8}$$

where $\boldsymbol{U}_k, \boldsymbol{V}_k \in \mathbb{R}^{d\times r}$ are Haar distributed with orthonormal columns, A_k is Pareto with shape κ and scale a_0, and $A_k, \boldsymbol{U}_k, \boldsymbol{V}_k$ are mutually independent and independent across iterations.

Theorem 3.6. Consider problem (7) under Assumption 3.5, and let each method use the step sizes $\eta_k = \frac{c}{k+k_0}$ with $c, k_0 > 0$. Let $\tau_{\text{sp}} > 0$ satisfy $\tau_{\text{sp}} \leq (1 - r/d)^2 a_0$, or $\tau_{\text{sp}} \leq a_0/2$ if $r \leq d/2$. Set

$$\lambda_{\text{sp}} := \left(1 - \tfrac{r}{d}\right)^2 + \tfrac{r(4d-3r-1)\,\kappa\tau_{\text{sp}}}{2(\kappa+1)\,a_0 d^2}, \qquad \Psi_{\boldsymbol{H}}(t) := \tfrac{t^2}{2d} \textstyle\sum_{j=1}^{d} \tfrac{h_j}{2th_j - 1} \quad \text{for } t > \tfrac{1}{2h_{\min}},$$

and let $t_{\boldsymbol{H}}$ be the unique minimizer of $\Psi_{\boldsymbol{H}}$. Let $\boldsymbol{X}_k^{\text{sp}}$ follow (2) with $\tau_k = \tau_{\text{sp}}$ and $c = t_{\boldsymbol{H}}/\lambda_{\text{sp}}$, and let $\boldsymbol{X}_k^{\text{F}}$ follow (1) with the Frobenius norm, any threshold $\tau_{\text{F}} > 0$ and any $c > 0$. Then

$$\lim_{k\to\infty} k\left(\mathbb{E}\left[f(\boldsymbol{X}_k^{\text{sp}})\right] - f_*\right) = \tfrac{r\tau_{\text{sp}}^2\,\Psi_{\boldsymbol{H}}(t_{\boldsymbol{H}})}{\lambda_{\text{sp}}^2} < \tfrac{r(\kappa+2)a_0^2\,\Psi_{\boldsymbol{H}}(t_{\boldsymbol{H}})}{\kappa} \leq \lim_{k\to\infty} k\left(\mathbb{E}\left[f(\boldsymbol{X}_k^{\text{F}})\right] - f_*\right). \tag{9}$$

Theorem 3.6 shows that, in this quadratic setting, norm clipping has a strictly larger constant in the $1/k$ rate than spectral clipping: *norm clipping divides the whole gradient by a large norm, while spectral clipping clips only the spike*. The proof is provided in Appendix D. Figure 4 (Appendix D) confirms this numerically: the constant of spectral clipping is about 4.4 times smaller than that of the best-tuned norm clipping, and all measured constants agree with the theoretical values.

4 Adaptive Clipping Strategies

As discussed in Theorem 3.3, clipping introduces bias; therefore, we need to choose the threshold τ to handle the noise σ and the bias $\frac{1}{K}\sum_{k=0}^{K-1}\Delta_k$. In Theorem 3.4, we show that one can choose τ depending on the noise level σ to ensure convergence. However, in practice, we do not have any knowledge of σ. Moreover, Zhang et al. (2020b) shows that the gradient noise distribution changes during the course of training. Therefore, we propose adaptive strategies for choosing the clipping thresholds. We describe our strategies below.

EMA of Singular Values. The clipping threshold τ_k can be chosen using an exponential moving average of the top singular values of the gradient matrix $\boldsymbol{G}_k$, i.e.,

$$\begin{aligned} \hat{\tau}_k &= \theta\hat{\tau}_{k-1} + (1-\theta)\sigma_{\max}(\boldsymbol{G}_k), \\ \tau_k &= \hat{\tau}_k/(1-\theta^{k+1}) \end{aligned} \tag{10}$$

for some $\theta \in (0, 1)$. In all our experiments, we use $\theta = 0.9$, which places more weight on previous estimates and less on the current gradient. As discussed in the previous section, the top singular values of gradient matrices increase in the presence of noise. Therefore, when there is sudden noise in the gradients, it will not significantly distort our clipping threshold. Similar adaptive ideas were proposed by (Zhang et al., 2020b) for a norm based clipping method.

Quantile of Singular Values. We also propose a quantile-based adaptive threshold computed over a sliding window of w top singular values, i.e.,

$$\tau_k = \mathcal{Q}_q(\mathcal{S}_k), \tag{11}$$

$$\mathcal{S}_k = \left(\mathcal{S}_{k-1} \setminus \sigma_{\max}\left(\boldsymbol{G}_{k-w}\right)\right) \cup \sigma_{\max}\left(\boldsymbol{G}_k\right) \tag{12}$$

for some $q \in (0, 1]$. Here $\mathcal{S}_k$ is the set of the last w top singular values, and $\mathcal{Q}_q$ denotes the quantile operator at level q. Thus, (11) sets the clipping threshold based on the q-th quantile of the sliding window set $\mathcal{S}_k$, and (12) updates the window by discarding the outdated singular value $\sigma_{\max}\left(\boldsymbol{G}_{k-w}\right)$ and adding the new one $\sigma_{\max}\left(\boldsymbol{G}_k\right)$. In both rules, the threshold is updated after $\boldsymbol{G}_k$ is observed and applied at the next iteration: $\boldsymbol{G}_k$ is clipped with $\mathcal{C}_{\tau_{k-1}}$, where $\tau_{-1} = +\infty$ (Algorithms 1 and 3). Note that our EMA based clipping strategy can still be sensitive to occasional extremely large singular values. However, this quantile-based rule can mitigate this issue by selecting the threshold based on the empirical distribution of the top singular values. We use $w = 100$ and, in the trace regression experiment, $q = 0.85$; for the MLP, q is tuned (Appendix G).

In Figure 7a (Appendix F), we compare spectral clipping with different choices of threshold τ_k for the trace regression problem (Kadri et al., 2020)

$$\min_{\boldsymbol{X}\in\mathbb{R}^{n\times n}} f(\boldsymbol{X}) := \frac{1}{N}\sum_{i=1}^{N}\left(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle - \boldsymbol{b}_i\right)^2 \tag{13}$$

and plot the average gradient bias $\frac{1}{K}\sum_{k=0}^{K-1}\Delta_k$ on the y-axis. In this example, we observe that $\frac{1}{K}\sum_{k=0}^{K-1}\Delta_k$ approaches 0 when the threshold τ_k is a sufficiently large constant, but is non-zero when τ_k is a small constant (i.e. ≤ 0.1). This example shows that we need to tune the clipping threshold τ_k for better accuracy of the spectral clipping algorithm. Moreover, we also implement our EMA based (10) and quantile-based (12) clipping strategy. We find that these two adaptive clipping strategies are independent of tuning τ_k and work well in Figure 7a.

5 Implementation of Spectral Clipping

In this section, we discuss three aspects to make spectral clipping efficient for training deep neural networks. First, we discuss how spectral clipping can be combined with other optimization algorithms. Then, we propose an efficient clipping procedure using a truncated SVD approximation. Finally, in the third subsection, we suggest using an adaptive threshold for each layer of the deep neural network.

5.1 Spectral Clipping beyond SGD

Algorithm 1 EMA Spectral Clipping

Require: $\boldsymbol{X}_0$, $\boldsymbol{B}_{-1} = 0$, β, θ, $\hat{\tau}_{-1} = 0$, $\tau_{-1} = +\infty$

1: **for** $k = 0, \ldots, K-1$ **do**
2: Compute $\boldsymbol{G}_k$
3: $\boldsymbol{B}_k = \beta\boldsymbol{B}_{k-1} + (1-\beta)\mathcal{C}_{\tau_{k-1}}(\boldsymbol{G}_k)$
4: $\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \eta\boldsymbol{B}_k$
5: $\hat{\tau}_k = \theta\hat{\tau}_{k-1} + (1-\theta)\sigma_{\max}(\boldsymbol{G}_k)$
6: $\tau_k = \hat{\tau}_k / 1-\theta^{k+1}$
7: **end for**

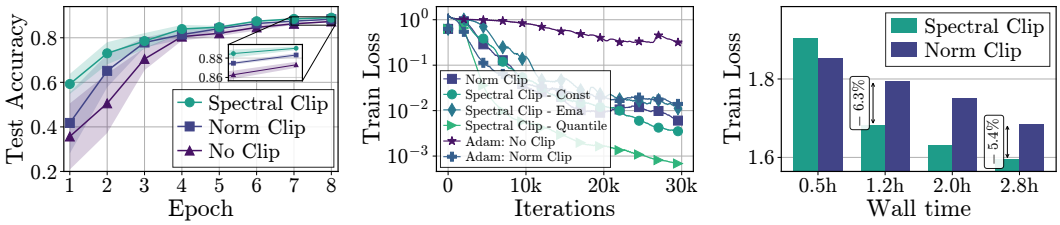
In algorithm (2), we use spectral clipping for the simple SGD update. However, this idea of spectral clipping can be extended to other algorithms like SGD with momentum (SGDM) (Polyak, 1964; Liu et al., 2020) or Adam (Kingma, 2014). For training modern deep neural networks, we often use gradient norm clipping along with SGDM or Adam. Therefore, we can replace gradient norm clipping with spectral clipping for these algorithms. We provide the pseudocode for SGDM with EMA based and quantile based spectral gradient clipping in Algorithm 1 and Algorithm 3 (Appendix E).

5.2 Spectral Gradient Clipping with Truncated SVD

Here, we discuss a practical implementation of the spectral clipping method. Recall that the update rule in (2) requires applying spectral clipping to the gradient matrix $\boldsymbol{G}_k$ at each iteration, which involves knowledge of singular value decomposition (SVD) (Stewart, 1993). For the very large matrices that arise in modern language models, computing a complete SVD at every step is prohibitively expensive (Dongarra et al., 2018) and therefore impractical.

Recently, Zhao et al. (2024) developed a memory-efficient optimization algorithm called GaLore (Gradient Low-Rank Projection) for large model training. Specifically, GaLore projects the gradients into a low-rank subspace and performs updates only in that space. To obtain this low-rank projection, Su et al. (2025) computes a rank-r truncated SVD using a fast randomized SVD algorithm (Halko et al., 2011). This motivates an analogous strategy for spectral clipping: instead of computing the entire SVD of the gradient matrix $\boldsymbol{G}$, we only compute a rank-r truncated SVD using a randomized method from Halko et al. (2011) and then clamp the top r singular values of $\boldsymbol{G}$.

Rank-r Truncated SVD. For a matrix $\boldsymbol{G} \in \mathbb{R}^{m\times n}$ with a full SVD given by $\boldsymbol{G} = \sum_{i=1}^{\min\{m,n\}}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^\top = \boldsymbol{U}\mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{V}^\top$, the rank-$r$ truncated SVD retains only the leading r singular values and singular vectors: $\boldsymbol{G}_r := \sum_{i=1}^{r}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^\top = \boldsymbol{U}[:r]\mathrm{diag}(\boldsymbol{\sigma}[:r])\boldsymbol{V}[:r]^\top$, where $\boldsymbol{U}[:r] \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V}[:r] \in \mathbb{R}^{n\times r}$ contain the first r columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, and $\boldsymbol{\sigma}[:r] \in \mathbb{R}^r$ stacks the top r singular values. The matrix $\boldsymbol{G}_r$ is the best rank-r approximation to $\boldsymbol{G}$ (Eckart & Young, 1936). Randomized SVD (Halko et al., 2011) efficiently approximates this truncated SVD by constructing a low-dimensional subspace that captures the dominant singular directions and then performing an SVD in that subspace. PyTorch implements this algorithm of Halko et al. (2011) via the `torch.svd_lowrank` function and computes the truncated SVD approximation. Its cost scales as $\mathcal{O}(mnr + (m+n)r^2)$, which is substantially cheaper than a full SVD when $r \ll \min\{m,n\}$. We utilize `torch.svd_lowrank` in Algorithm 2 for spectral clipping of matrices.



(a) CifarNet on CIFAR-10N Aggregate

(b) MLP with heavy-tailed noise

(c) Time comparison

Figure 2: In Figure 2a, we plot test accuracy of SGDM with no clipping, norm clipping, and spectral clipping on CIFAR-10N Aggregate. In Figure 2b, we compare the performance of SGDM and Adam for different clipping strategies for MLP problem. In Figure 2c, we compare wall-clock time in NanoGPT experiment.

Algorithm 2 Clipping via Truncated SVD

Require: Gradient matrix: $\boldsymbol{G} \in \mathbb{R}^{m\times n}$, threshold $\tau > 0$, rank $r \in \mathbb{N}$

1: $\overline{\boldsymbol{U}}, \overline{\boldsymbol{\sigma}}, \overline{\boldsymbol{V}} = \texttt{torch.svd_lowrank}(\boldsymbol{G}, r)$

2: $\overline{\boldsymbol{\sigma}}_{\text{EXTRA},\tau} = \overline{\boldsymbol{\sigma}} - \text{clamp}_\tau(\overline{\boldsymbol{\sigma}})$

3: **return** $\boldsymbol{G} - \overline{\boldsymbol{U}}\,\text{diag}(\overline{\boldsymbol{\sigma}}_{\text{EXTRA},\tau})\,\overline{\boldsymbol{V}}^\top$

In Algorithm 2, `torch.svd_lowrank` first approximates the top r singular values and associated singular vectors of $\boldsymbol{G}$ using a randomized SVD routine. Next, we compute the vector $\overline{\boldsymbol{\sigma}}_{\text{EXTRA},\tau}$ that captures how much each of the top r singular values exceeds the clipping threshold τ. Finally, we subtract the corresponding rank-r matrix $\overline{\boldsymbol{U}}\,\text{diag}(\overline{\boldsymbol{\sigma}}_{\text{EXTRA},\tau})\,\overline{\boldsymbol{V}}^\top$ from $\boldsymbol{G}$, which effectively clamps the top r singular values at τ, while leaving smaller singular values and all directions outside the rank-r subspace unchanged. With exact leading singular components, Algorithm 2 coincides with exact spectral clipping whenever $\sigma_{r+1}(\boldsymbol{G}) \le \tau$; the randomized approximation introduces an additional error, which we evaluate empirically.

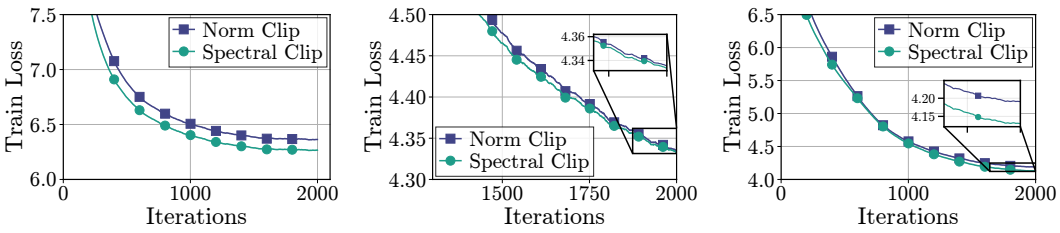
In practice, we choose r to be small ($r \le 64$ in our experiments), so that Algorithm 2 computes and clips only the few most dominant singular directions of the gradient matrix, providing substantial computational savings compared to a full SVD while retaining the main benefits of spectral clipping. Figure 7b (Appendix F) shows that this approximation costs little in accuracy: on our CV task (Section 6.1), randomized truncated SVD with $r = 10$ matches the final test accuracy of full SVD over 20 epochs with identical hyperparameters, at a fraction of the cost. The memory overhead of Algorithm 2 is also negligible: for GPT-2 models, spectral clipping changes the peak memory of a training step by at most 0.15% (Appendix H).

5.3 Layer-wise Adaptive Clipping

In deep neural networks, different layers operate at different scales. Thus, using a single clipping threshold τ_k at iteration k for the entire network can be suboptimal (Brock et al., 2021; Zhang et al., 2020b). This can result in one of two scenarios: (1) over-clipping layers with small singular values, which results in slow optimization, or (2) under-clipping layers with large singular values, thus harming the stability of the algorithm.

To address this, we propose a layer-wise adaptive spectral clipping scheme in which each layer maintains its own adaptive clipping threshold. Let $\boldsymbol{G}_k^{(\ell)}$ denote the gradient of layer $\ell \in \{1, \cdots, L\}$ at iteration k. Then, for each layer ℓ, we have an EMA-based clipping threshold $\tau_k^{(\ell)}$ given by $\hat{\tau}_k^{(\ell)} = \theta\hat{\tau}_{k-1}^{(\ell)} + (1-\theta)\sigma_{\max}\left(\boldsymbol{G}_k^{(\ell)}\right), \tau_k^{(\ell)} = \hat{\tau}_k^{(\ell)}/(1-\theta^{k+1})$ or a quantile-based clipping threshold given by $\tau_k^{(\ell)} = \mathcal{Q}_q\left(\mathcal{S}_k^{(\ell)}\right)$ and $\mathcal{S}_k^{(\ell)} = \left(\mathcal{S}_{k-1}^{(\ell)} \setminus \sigma_{\max}\left(\boldsymbol{G}_{k-w}^{(\ell)}\right)\right) \cup \sigma_{\max}\left(\boldsymbol{G}_k^{(\ell)}\right)$. These per-layer thresholds $\{\tau_k^{(\ell)}\}_{\ell=1}^L$ are then used for spectral clipping of layer-wise gradient matrices $\{\boldsymbol{G}_{k+1}^{(\ell)}\}_{\ell=1}^L$ using Algorithm 2.

This layer-wise clipping scheme can have several practical advantages. First, it adapts the clipping threshold to each layer of the neural network, allowing aggressive clipping in layers with large singular values while avoiding unnecessary clipping in layers with small singular values. Secondly, the EMA-based and quantile-based updates of the clipping threshold can adapt during training (e.g., phase transitions), thus automatically increasing or decreasing $\tau_k^{(\ell)}$ as the spectrum of $\boldsymbol{G}_k^{(\ell)}$ evolves.



(a) SGDM (b) Muon (c) Adam

Figure 3: Training loss of GPT-2 on the FineWeb dataset for different optimizers.

6 Numerical Experiments

We compare spectral clipping with existing gradient clipping strategies on a CV task with real human label noise (Section 6.1), on language modeling (Section 6.2), and on a synthetic regression task with injected low-rank, heavy-tailed gradient noise (Section 6.3). Hyperparameter tuning details for all experiments are provided in Appendix G.

6.1 Training CifarNet on CIFAR-10N

We train AirBench CifarNet (Jordan, 2024) on CIFAR-10N (Wei et al., 2022) with three sets of human labels, Aggregate, Random-1 and Worst, with 9.06%, 17.18% and 40.14% label noise in our training subset, respectively. *CIFAR-10N is a natural setting to evaluate spectral clipping: its label noise comes from real human annotators, so mislabeled samples act as genuine data outliers*, and gradient clipping is used in this setting to limit the influence of mislabeled samples (Ye et al., 2025); we show that spectral clipping achieves the highest final test accuracy on all three label sets. We train each model with SGDM for 8 epochs in the AirBench setup, using no clipping, norm clipping, or spectral clipping with a constant threshold.

As shown in Figure 2a, spectral clipping with a constant threshold achieves the highest test accuracy on CIFAR-10N Aggregate after eight epochs and converges much faster: after the first epoch it is 17 points ahead of norm clipping and 24 points ahead of no clipping, and it stays ahead at every epoch. Results for Random-1 and Worst labels are shown in Figure 6 (Appendix F), where spectral clipping again leads throughout training.

6.2 Training Language Models

Spectral clipping outperforms norm clipping with SGDM and Muon when training NanoGPT (Karpathy, 2022) on the Shakespeare character-level dataset (Figures 8a and 8b, Appendix F). Despite the higher per-iteration cost of the truncated SVD, after 1.2 hours of training SGDM with spectral clipping reaches a training loss that is 6.3% lower than with norm clipping on this task (Figure 2c). Next, we evaluate our approach on the GPT-2 124M model (Radford et al., 2019), trained from scratch on the FineWeb dataset (Penedo et al., 2024). Our results, shown in Figure 3, demonstrate that spectral clipping consistently outperforms standard norm-based clipping when used with SGDM, Muon, and Adam. In Figure 8c (Appendix F), we compared wall-clock time with SGDM showing that after 27 minutes of training, SGDM with spectral clipping achieves a training loss that is 2.6% lower than with norm clipping.

6.3 Training MLP with Heavy-Tailed Noise

To evaluate algorithms under heavy-tailed gradient noise, we train a 4-layer MLP, whose first three layers are 100×100 matrices and whose last layer outputs a scalar, to predict $y = x_1 x_2 x_3$ from $\boldsymbol{x} \in \mathbb{R}^{100}$, and add rank-one two-sided Pareto perturbations to its gradients with probability 10% at each step, using shape $\alpha = 1.1$ and scale $x_m = 5$ ($x_m = 2$ in Figure 5) as in Chezhegov et al. (2024).

We compare Adam and SGDM, each with no clipping, norm clipping, and spectral clipping with a constant, EMA-based or quantile-based threshold. As shown in Figure 2b, SGDM with spectral clipping consistently outperforms SGDM with norm clipping and Adam; quantile-based spectral clipping performs best, and EMA-based spectral clipping is comparable to carefully tuned constant-threshold spectral and norm clipping. Over a grid of thresholds and step sizes, spectral clipping with SGDM also reaches lower training loss than norm clipping across a much wider range of hyperparameters (Figure 5, Appendix F), so it requires less manual tuning.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing, literature checks, and mathematical auditing, including checking algebra and assumptions and revising theorem statements, proofs, and their interpretation. The authors are responsible for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our theoretical results rely only on the assumptions stated in Section 3; complete proofs are given in Appendices C and D, and the latter also reports a numerical check of Theorem 3.6. Spectral clipping is implemented as in Algorithm 2 in the neural network experiments and with the exact SVD in the MLP and trace regression experiments (Appendix G). For every experiment, we describe the dataset, the model, the training budget, the hyperparameter search space and tuning protocol, and the selected hyperparameters in Section 6 and Appendix G; the memory measurements are described in Appendix H. Anonymized source code for all experiments is available at https://anonymous.4open.science/r/spectral-clipping-76B4.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). *arXiv preprint arXiv:2310.01082*, 2023.

Noah Amsel, David Persson, Christopher Musco, and Robert M Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. *arXiv preprint arXiv:2505.16932*, 2025.

Kang An, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, and Tong Zhang. Asgo: Adaptive structured gradient optimization. *arXiv preprint arXiv:2503.20762*, 2025.

Barak Battash, Lior Wolf, and Ofir Lindenbaum. Revisiting the noise model of stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 4780–4788. PMLR, 2024.

Albert S Berahas, Jorge Nocedal, and Martin Takáč. A multi-batch L-BFGS method for machine learning. In *The Thirtieth Annual Conference on Neural Information Processing Systems (NIPS)*, 2016.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd: Compressed optimisation for non-convex problems, 2018. URL https://arxiv.org/abs/1802.04434.

Lukas Biewald. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Johan Bjorck, Kilian Q Weinberger, and Carla Gomes. Understanding decoupled and early weight decay. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6777–6785, 2021.

Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *International conference on machine learning*, pp. 1059–1071. PMLR, 2021.

Da Chang, Yongxiang Liu, and Ganzhao Yuan. On the convergence of muon and beyond. *arXiv preprint arXiv:2509.15816*, 2025.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms, 2023. URL https://arxiv.org/abs/2302.06675.

Savelii Chezhegov, Yaroslav Klyukin, Andrei Semenov, Aleksandr Beznosikov, Alexander Gasnikov, Samuel Horváth, Martin Takáč, and Eduard Gorbunov. Clipping improves adam-norm and adagrad-norm when the noise is heavy-tailed. *arXiv preprint arXiv:2406.04443*, 2024.

Sayantan Choudhury, Nazarii Tupitsa, Nicolas Loizou, Samuel Horvath, Martin akáč, and Eduard Gorbunov. Remove that square root: A new efficient scale-invariant version of adagrad. In *NeurIPS*, 2024.

Sayantan Choudhury, Xiaoran Cheng, Martin Takáč, Sen Na, and Mladen Kolar. Muon with nesterov momentum: Heavy-tailed noise and (randomized) inexact polar decomposition, 2026. URL https://arxiv.org/abs/2605.06884.

Kai Lai Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25(3):463–483, 1954.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *The Journal of Machine Learning Research*, 13(1):165–202, 2012.

Jack Dongarra, Mark Gates, Azzam Haidar, Jakub Kurzak, Piotr Luszczek, Stanimire Tomov, and Ichitaro Yamazaki. The singular value decomposition: Anatomy of optimizing an algorithm for extreme scale. *SIAM review*, 60(4):808–865, 2018.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3): 211–218, 1936.

Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, Zico Kolter, Zachary Lipton, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients. In *International Conference on Machine Learning*, pp. 3610–3619. PMLR, 2021.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020a.

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 680–690. PMLR, 2020b.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Xi He, Dheevatsa Mudigere, Mikhail Smelyanskiy, and Martin Takáč. Distributed hessian-free optimization for deep neural network. In *AAAI Workshops*, 2017.

Nicholas J Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986.

Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed sgd. *arXiv preprint arXiv:2410.13849*, 2024.

Xiaowen Jiang, Andrei Semenov, and Sebastian U. Stich. Enhancing LLM training via spectral clipping. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Keller Jordan. 94% on CIFAR-10 in 3.29 Seconds on a Single GPU. *arXiv preprint arXiv:2404.00498*, 2024.

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

Hachem Kadri, Stéphane Ayache, Riikka Huusari, Alain Rakotomamonjy, and Ralaivola Liva. Partial trace regression and low-rank kraus decomposition. In *International Conference on Machine Learning*, pp. 5031–5041. PMLR, 2020.

Andrej Karpathy. NanoGPT. https://github.com/karpathy/nanoGPT, 2022.

Tosio Kato. *Perturbation theory for linear operators*. Classics in Mathematics. Springer, 2 edition, 1995.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pp. 17343–17363. PMLR, 2023.

Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1): 365–397, 2012.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Shuang Li, William J Swartworth, Martin Takáč, Deanna Needell, and Robert M Gower. SP2: a second order stochastic polyak method. *ICLR 2023*, 2022.

Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems*, 33:18261–18271, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Tomáš Mikolov et al. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, 80(26), 2012.

Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. *arXiv preprint arXiv:2306.06101*, 2023.

Mohammad Mohammadi, Adel Mohammadpour, and Hiroaki Ogata. On estimating the tail index and the spectral measure of multivariate α-stable distributions. *Metrika*, 78(5):549–561, 2015.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Springer, 2004.

Lam Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: a novel method for machine learning problems using stochastic recursive gradient. In *In 34th International Conference on Machine Learning, ICML 2017*, 2017.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318. Pmlr, 2013.

Guilherme Penedo, Hynek Kydlíček, Leandro von Werra, Thomas Wolf, et al. The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos. *arXiv preprint arXiv:2502.07529*, 2025a.

Thomas Pethick, Wanyun Xie, Mete Erdogan, Kimon Antonakopoulos, Tony Silveti-Falls, and Volkan Cevher. Generalized gradient norm clipping & non-euclidean (l_0, l_1)-smoothness. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

Jiang Qian, Yuren Wu, Bojin Zhuang, Shaojun Wang, and Jing Xiao. Understanding gradient clipping in incremental gradient methods. In *International Conference on Artificial Intelligence and Statistics*, pp. 1504–1512. PMLR, 2021.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Fabian Schaipp, Ruben Ohana, Michael Eickenberg, Aaron Defazio, and Robert M Gower. Momo: Momentum models for adaptive learning rates. *arXiv preprint arXiv:2305.07583*, 2023.

Ishaan Shah, Anthony M Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, et al. Practical efficiency of muon for pretraining. *arXiv preprint arXiv:2505.02222*, 2025.

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

Zheng Shi, Abdurakhmon Sadiev, Nicolas Loizou, Peter Richtárik, and Martin Takáč. AI-SARAH: adaptive and implicit stochastic recursive gradient methods. *Transactions on Machine Learning Research*, 2023.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pp. 5827–5837. PMLR, 2019.

Gilbert W Stewart. On the early history of the singular value decomposition. *SIAM review*, 35(4):551–566, 1993.

DiJia Su, Andrew Gu, Jane Xu, Yuandong Tian, and Jiawei Zhao. Galore 2: Large-scale llm pre-training by gradient low-rank projection. *arXiv preprint arXiv:2504.20437*, 2025.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Martin Takáč, Avleen Bijral, Peter Richtárik, and Nathan Srebro. Mini-batch primal and dual methods for SVMs. In *In 30th International Conference on Machine Learning, ICML 2013*, 2013.

Tijmen Tieleman. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26, 2012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321*, 2024.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TBWA6PLJZQm.

Xichen Ye, Yifan Wu, Weizhong Zhang, Xiaoqiang Li, Yifan Chen, and Cheng Jin. Optimized gradient clipping for noisy label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9463–9471, 2025.

Dingzhi Yu, Hongyi Tao, Yuanyu Wan, Luo Luo, and Lijun Zhang. Sign-based optimizers are effective under heavy-tailed noise, 2026. URL https://arxiv.org/abs/2602.07425.

Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521, 2020a.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models?, 2020b. URL https://arxiv.org/abs/1912.03194.

Zhihua Zhang. The singular value decomposition, applications and beyond. *arXiv preprint arXiv:1510.08532*, 2015.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*, 2024.

Supplementary Material

CONTENTS

1 Introduction **1**

1.1 Background and Related Works . . . 2

1.2 Main Contributions . . . 2

2 Gradient Clipping beyond Vector Norms **3**

2.1 Singular Values of Gradient Matrices . . . 3

2.2 Spectral Clipping of Gradient Matrices . . . 3

2.3 Gradient Norm Clipping as a Special Case . . . 4

3 Convergence Guarantees **4**

3.1 Spectral versus Norm Clipping under Low-Rank Noise . . . 5

4 Adaptive Clipping Strategies **6**

5 Implementation of Spectral Clipping **7**

5.1 Spectral Clipping beyond SGD . . . 7

5.2 Spectral Gradient Clipping with Truncated SVD . . . 7

5.3 Layer-wise Adaptive Clipping . . . 8

6 Numerical Experiments **9**

6.1 Training CifarNet on CIFAR-10N . . . 9

6.2 Training Language Models . . . 9

6.3 Training MLP with Heavy-Tailed Noise . . . 9

A Further Related Works **17**

B Technical Lemmas **17**

C Convergence Analysis **21**

C.1 Proof of Theorem 3.3 . . . 23

C.2 Proof of Theorem 3.4 . . . 24

D Proof of Theorem 3.6 **29**

E Missing Pseudocode of Algorithm 3 **40**

F Missing Figures **40**

G Hyperparameter Tuning Details **41**

H Memory Overhead **42**

I Gradient Spectral Profiles under Controlled Token Replacement **43**

J Comparison of Spectral Clipping Before and After Momentum 43

A Further Related Works

Heavy-Tailed Gradient Noise. A growing body of empirical evidence suggests that stochastic gradient noise in modern deep learning is often heavy-tailed (Simsekli et al., 2019; Zhang et al., 2020b) rather than in settings captured by the classical bounded variance assumption (Assumption 3.2 with $\alpha = 2$). This observation has motivated clipping-based algorithms, adaptive clipping rules, and related normalization analyses under bounded α-moment or high-probability assumptions (Gorbunov et al., 2020a; Zhang et al., 2020b; Hübler et al., 2024; Chezhegov et al., 2024). In particular, Zhang et al. (2020b) proposed ACClip, an adaptive coordinate-wise clipping rule based on moving-average thresholds, and Chezhegov et al. (2024) showed that clipping can substantially improve AdaGrad and Adam-type methods in heavy-tailed regimes.

Matrix Structure and Spectral Computations. For a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with rank r, the singular value decomposition (SVD) of $\boldsymbol{A}$ is given by $\boldsymbol{U}\mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times r}$ have orthonormal columns and $\boldsymbol{\sigma} \in \mathbb{R}^r_+$ contains the singular values. The SVD plays an important role in many applications Zhang (2015). Recent optimizer design has increasingly sought to preserve matrix structure rather than flatten parameters into vectors. Matrix-aware preconditioners such as Shampoo and SOAP use tensor structure or eigenspace-adapted second moments to construct richer updates (Gupta et al., 2018; Vyas et al., 2024). Orthogonalization or norm-based methods such as Muon, Scion, and ASGO also operate directly on matrix-valued parameters, but they modify the update direction at every step via orthogonalization, norm-constrained LMOs, or structured preconditioning (Jordan et al., 2024; Pethick et al., 2025a; An et al., 2025; Shah et al., 2025). The computation of the full decomposition is expensive and requires $\mathcal{O}(mn^2)$ work. Over the years, researchers have developed several techniques to reduce this cost for specific applications. Randomized low-rank decompositions provide a standard route to scalable partial SVD computation (Halko et al., 2011), and related low-rank ideas have recently been used in optimization systems such as GaLore (Zhao et al., 2024). Orthogonalization-based optimizers such as Muon typically avoid exact decompositions via Newton-Schulz (Higham, 1986; Jordan et al., 2024), while PolarExpress (Amsel et al., 2025) develops faster GPU-friendly polar decomposition routines for this setting. In this work, we utilize randomized truncated SVD (Halko et al., 2011) to clamp the singular values of the gradient matrix beyond a certain threshold.

B Technical Lemmas

In this section, we present some technical lemmas, which will be used repeatedly to prove the main results of the work in subsequent sections.

Lemma B.1 (Young's Inequality)**.** For any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$ and $\omega > 0$ we have

$$\|\boldsymbol{A}+\boldsymbol{B}\|_F^2 \quad \leq \quad (1+\omega)\|\boldsymbol{A}\|_F^2 + (1+1/\omega)\|\boldsymbol{B}\|_F^2. \tag{14}$$

In particular, for $\omega = 1$, we have

$$\|\boldsymbol{A}+\boldsymbol{B}\|_F^2 \quad \leq \quad 2\|\boldsymbol{A}\|_F^2 + 2\|\boldsymbol{B}\|_F^2. \tag{15}$$

Lemma B.2 (Cauchy-Schwarz Inequality)**.** For any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$, we have

$$\langle \boldsymbol{A}, \boldsymbol{B}\rangle \leq \|\boldsymbol{A}\|_F \|\boldsymbol{B}\|_F. \tag{16}$$

Lemma B.3. For any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$ and $\omega > 0$, we have

$$\langle \boldsymbol{A}, \boldsymbol{B}\rangle \quad \leq \quad \frac{\omega}{2}\|\boldsymbol{A}\|_F^2 + \frac{1}{2\omega}\|\boldsymbol{B}\|_F^2. \tag{17}$$

Lemma B.4 (Jensen's Inequality)**.** For a random variable $\mathbf{a} \in \mathbb{R}$ and a function $g : \mathbb{R}^{m\times n} \to \mathbb{R}$, we have the following:

- if g is convex, then

$$g\left(\mathbb{E}\left[\mathbf{a}\right]\right) \quad \leq \quad \mathbb{E}\left[g(\mathbf{a})\right], \tag{18}$$

- and if g is concave, then

$$g\left(\mathbb{E}\left[\mathbf{a}\right]\right) \quad \geq \quad \mathbb{E}\left[g(\mathbf{a})\right]. \tag{19}$$

Lemma B.5 (Hölder's Inequality)**.** Suppose $p, q \in (1, \infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$ and $\mathbf{a}, \mathbf{b} \in \mathbb{R}$ are random variables. Then we have

$$\mathbb{E}\left[|\mathbf{ab}|\right] \quad \leq \quad \mathbb{E}\left[|\mathbf{a}|^p\right]^{\frac{1}{p}} \mathbb{E}\left[|\mathbf{b}|^q\right]^{\frac{1}{q}}. \tag{20}$$

Lemma B.6 (Markov's Inequality)**.** For the random variable $\mathbf{a} \in \mathbb{R}$ and a positive number $\varepsilon > 0$, we have

$$\Pr[|\mathbf{a}| \geq \varepsilon] \quad \leq \quad \frac{\mathbb{E}\left[|\mathbf{a}|\right]}{\varepsilon}. \tag{21}$$

Lemma B.7 (Mirsky's Inequality)**.** For any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$, we have

$$\sum_{i=1}^{\min\{m,n\}} \left(\sigma_i(\boldsymbol{A}) - \sigma_i(\boldsymbol{B})\right)^2 \quad \leq \quad \|\boldsymbol{A} - \boldsymbol{B}\|_F^2, \tag{22}$$

where $\sigma_1(\cdot) \geq \sigma_2(\cdot) \geq \ldots$ denote the singular values in nonincreasing order.

Lemma B.8 (Weyl's Inequality)**.** For any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$ and $i \leq \min\{m, n\}$, we have

$$|\sigma_i(\boldsymbol{A} + \boldsymbol{B}) - \sigma_i(\boldsymbol{A})| \quad \leq \quad \sigma_1(\boldsymbol{B}) \leq \|\boldsymbol{B}\|_F. \tag{23}$$

Lemma B.9 (Projection onto a Convex Set)**.** Let $\mathcal{K} \subset \mathbb{R}^{m \times n}$ be nonempty, closed and convex, and let $P_{\mathcal{K}}$ be the projection onto $\mathcal{K}$ in the Frobenius norm. Then for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$, we have

$$\|P_{\mathcal{K}}\boldsymbol{A} - P_{\mathcal{K}}\boldsymbol{B}\|_F^2 \quad \leq \quad \langle \boldsymbol{A} - \boldsymbol{B}, P_{\mathcal{K}}\boldsymbol{A} - P_{\mathcal{K}}\boldsymbol{B} \rangle. \tag{24}$$

In particular, $P_{\mathcal{K}}$ is 1-Lipschitz and monotone.

Lemma B.10 (Azuma–Hoeffding Inequality)**.** Let $(S_t)_{t \geq t_0}$ be a martingale with $S_{t_0} = 0$ and $|S_{t+1} - S_t| \leq c_t$ almost surely. Then for any $T > t_0$ and $x > 0$, we have

$$\Pr\left[\max_{t_0 \leq t \leq T} |S_t| \geq x\right] \quad \leq \quad 2\exp\left\{-\frac{x^2}{2\sum_{t=t_0}^{T-1} c_t^2}\right\}. \tag{25}$$

Lemma B.11 below is a variant of Lemmas 1 and 2 of Chung (1954) with $p = 1$ in their notation. It differs from the original statement in three respects: the coefficient is $1 - \frac{\gamma}{k+k_0}$ with an arbitrary real $k_0 > 0$ instead of $1 - \frac{\gamma}{k}$, which is the form produced by the step size $\eta_k = \frac{c}{k+k_0}$; only bounds on $\limsup_{k\to\infty} k\, y_k$ and $\liminf_{k\to\infty} k\, y_k$ are asserted, rather than an asymptotic expansion of y_k; and part (b) also covers $\gamma < 1$, where $k\, y_k$ diverges. Since these modifications are not covered by the original statement, we include a short self-contained proof.

Lemma B.11 (Chung's Lemma)**.** Let $\gamma, \zeta, k_0 > 0$ and let $(y_k)_{k \geq 0}$ be a nonnegative sequence.

(a) If $\gamma > 1$ and $y_{k+1} \leq \left(1 - \frac{\gamma}{k+k_0}\right) y_k + \frac{\zeta}{(k+k_0)^2}$ for all sufficiently large k, then $\limsup_{k\to\infty} k\, y_k \leq \frac{\zeta}{\gamma - 1}$.

(b) If $y_{k+1} \geq \left(1 - \frac{\gamma}{k+k_0}\right) y_k + \frac{\zeta}{(k+k_0)^2}$ for all sufficiently large k, then $\liminf_{k\to\infty} k\, y_k \geq \frac{\zeta}{\gamma - 1}$ when $\gamma > 1$, and $k\, y_k \to \infty$ when $\gamma < 1$.

Proof. Fix $k' \geq 0$ such that the assumed inequality holds for all $k \geq k'$ and $k' + k_0 > \gamma$. Then, for all $k \geq k'$, we have $0 < 1 - \frac{\gamma}{k+k_0} < 1$. Moreover, for all $k' \leq k'' \leq k$, we have

$$\begin{aligned}\prod_{i=k''}^{k-1}\Big(1-\frac{\gamma}{i+k_0}\Big) &\leq \exp\Big\{-\gamma\sum_{i=k''}^{k-1}\frac{1}{i+k_0}\Big\}\\ &\leq \exp\Big\{-\gamma\int_{k''}^{k}\frac{dx}{x+k_0}\Big\}\\ &= \Big(\frac{k''+k_0}{k+k_0}\Big)^{\gamma},\end{aligned} \tag{26}$$

where the first line follows from $1 - x \leq e^{-x}$, the second line follows from $\frac{1}{i+k_0} \geq \int_i^{i+1}\frac{dx}{x+k_0}$, since $x \mapsto \frac{1}{x+k_0}$ is decreasing, and the last line follows from $\int_{k''}^{k}\frac{dx}{x+k_0} = \log\frac{k+k_0}{k''+k_0}$.

Now, for $\bar{y} > 0$, let us define $z_k := \frac{\bar{y}}{k+k_0}$. Then, for all $k \geq 0$, we have

$$\begin{aligned}z_{k+1} - \Big(1-\frac{\gamma}{k+k_0}\Big)z_k - \frac{\zeta}{(k+k_0)^2} &= \frac{\bar{y}}{k+k_0+1} - \frac{\bar{y}}{k+k_0} + \frac{\gamma\bar{y}-\zeta}{(k+k_0)^2}\\ &= -\frac{\bar{y}}{(k+k_0)(k+k_0+1)} + \frac{\gamma\bar{y}-\zeta}{(k+k_0)^2}\\ &= \frac{(\gamma-1)\bar{y} - \zeta + \bar{y}/(k+k_0+1)}{(k+k_0)^2},\end{aligned} \tag{27}$$

where the first line follows from the definition of z_k, the second line follows from $\frac{1}{k+k_0+1} - \frac{1}{k+k_0} = -\frac{1}{(k+k_0)(k+k_0+1)}$, and the last line follows from $\frac{1}{(k+k_0)(k+k_0+1)} = \frac{1}{(k+k_0)^2} - \frac{1}{(k+k_0)^2(k+k_0+1)}$.

Finally, let $m \geq k'$, and let $(w_k)_{k\geq m}$ be a sequence with $w_{k+1} \leq \big(1-\frac{\gamma}{k+k_0}\big)w_k$ for all $k \geq m$. Then, for all $k \geq m$, we have

$$\begin{aligned}\max\{w_{k+1}, 0\} &\leq \max\Big\{\Big(1-\frac{\gamma}{k+k_0}\Big)w_k, 0\Big\}\\ &= \Big(1-\frac{\gamma}{k+k_0}\Big)\max\{w_k, 0\},\end{aligned}$$

where the first line follows from the assumed inequality and the monotonicity of $x \mapsto \max\{x, 0\}$, and the last line follows from $1 - \frac{\gamma}{k+k_0} > 0$. Iterating this inequality from m to $k-1$ and applying (26) with $k'' = m$, we obtain

$$\max\{w_k, 0\} \leq \max\{w_m, 0\}\Big(\frac{m+k_0}{k+k_0}\Big)^{\gamma} \qquad \text{for all } k \geq m. \tag{28}$$

(a) Let $\bar{y} > \frac{\zeta}{\gamma-1}$, so that $(\gamma-1)\bar{y} - \zeta > 0$. Then the right-hand side of (27) is positive for every $k \geq 0$, since also $\frac{\bar{y}}{k+k_0+1} > 0$. Now, let $w_k := y_k - z_k$. Then, for all $k \geq k'$, we have

$$\begin{aligned}w_{k+1} &= y_{k+1} - z_{k+1}\\ &\leq \Big(1-\frac{\gamma}{k+k_0}\Big)y_k + \frac{\zeta}{(k+k_0)^2} - z_{k+1}\\ &\leq \Big(1-\frac{\gamma}{k+k_0}\Big)y_k + \frac{\zeta}{(k+k_0)^2} - \Big(1-\frac{\gamma}{k+k_0}\Big)z_k - \frac{\zeta}{(k+k_0)^2}\\ &= \Big(1-\frac{\gamma}{k+k_0}\Big)w_k,\end{aligned}$$

where the second line follows from the assumption of part (a), the third line follows from (27), whose right-hand side is positive, and the last line follows from the definition of w_k. Hence, (28) applies with $m = k'$, and for all $k \geq k'$ we get

$$\begin{aligned}k\,y_k &= k\,z_k + k\,w_k\\ &\leq \frac{k\bar{y}}{k+k_0} + k\max\{w_{k'}, 0\}\Big(\frac{k'+k_0}{k+k_0}\Big)^{\gamma},\end{aligned}$$

where the first line follows from $y_k = z_k + w_k$, and the last line follows from the definition of z_k, $w_k \leq \max\{w_k, 0\}$ and (28). As $k \to \infty$, the first term tends to $\bar{y}$, and the second term tends to 0 because $\gamma > 1$. Therefore, $\limsup_{k\to\infty} k\,y_k \leq \bar{y}$. Since $\bar{y} > \frac{\zeta}{\gamma-1}$ is arbitrary, letting $\bar{y} \downarrow \frac{\zeta}{\gamma-1}$ proves part (a).

(b) For further analysis, we will consider two cases: either $\gamma > 1$ or $\gamma < 1$.

Case 1. Suppose $\gamma > 1$. Let $0 < \bar{y} < \frac{\zeta}{\gamma - 1}$, so that $(\gamma - 1)\bar{y} - \zeta < 0$. Since $\frac{\bar{y}}{k+k_0+1} \to 0$, there is $m \geq k'$ such that the right-hand side of (27) is negative for all $k \geq m$. Now, let $w_k := z_k - y_k$. Then, for all $k \geq m$, we have

$$\begin{aligned} w_{k+1} &= z_{k+1} - y_{k+1} \\ &\leq \Big(1 - \frac{\gamma}{k+k_0}\Big) z_k + \frac{\zeta}{(k+k_0)^2} - y_{k+1} \\ &\leq \Big(1 - \frac{\gamma}{k+k_0}\Big) z_k + \frac{\zeta}{(k+k_0)^2} - \Big(1 - \frac{\gamma}{k+k_0}\Big) y_k - \frac{\zeta}{(k+k_0)^2} \\ &= \Big(1 - \frac{\gamma}{k+k_0}\Big) w_k, \end{aligned}$$

where the second line follows from (27), whose right-hand side is negative for $k \geq m$, the third line follows from the assumption of part (b), and the last line follows from the definition of w_k. Hence, (28) applies, and for all $k \geq m$ we get

$$\begin{aligned} k\, y_k &= k\, z_k - k\, w_k \\ &\geq \frac{k\bar{y}}{k+k_0} - k \max\{w_m, 0\} \Big(\frac{m+k_0}{k+k_0}\Big)^{\gamma}, \end{aligned}$$

where the first line follows from $y_k = z_k - w_k$, and the last line follows from the definition of z_k, $w_k \leq \max\{w_k, 0\}$ and (28). As $k \to \infty$, the first term tends to $\bar{y}$, and the second term tends to 0 because $\gamma > 1$. Therefore, $\liminf_{k\to\infty} k\, y_k \geq \bar{y}$, and letting $\bar{y} \uparrow \frac{\zeta}{\gamma - 1}$ gives the claim.

Case 2. Now, consider $\gamma < 1$. We have

$$\begin{aligned} y_{k'+1} &\geq \Big(1 - \frac{\gamma}{k'+k_0}\Big) y_{k'} + \frac{\zeta}{(k'+k_0)^2} \\ &\geq \frac{\zeta}{(k'+k_0)^2} \\ &> 0, \end{aligned}$$

where the first line follows from the assumption of part (b), the second line follows from $y_{k'} \geq 0$ and $1 - \frac{\gamma}{k'+k_0} > 0$, and the last line follows from $\zeta > 0$. Moreover, for all $k \geq k' + 1$, we have

$$\begin{aligned} y_{k+1} &\geq \Big(1 - \frac{\gamma}{k+k_0}\Big) y_k + \frac{\zeta}{(k+k_0)^2} \\ &\geq \Big(1 - \frac{\gamma}{k+k_0}\Big) y_k, \end{aligned}$$

where the first line follows from the assumption of part (b), and the last line follows from $\zeta > 0$. Iterating this inequality, we obtain, for all $k \geq k' + 1$,

$$y_k \geq y_{k'+1} \prod_{i=k'+1}^{k-1} \Big(1 - \frac{\gamma}{i+k_0}\Big). \tag{29}$$

It remains to bound the product from below. Let $m \geq k' + 1$ be such that $\frac{\gamma}{m+k_0} \leq 1/2$. For $0 \leq x \leq 1/2$, we have

$$\begin{aligned} \log(1-x) &= -\sum_{j=1}^{\infty} \frac{x^j}{j} \\ &\geq -x - \frac{x^2}{2} \sum_{j=0}^{\infty} x^j \\ &= -x - \frac{x^2}{2(1-x)} \\ &\geq -x - x^2, \end{aligned}$$

where the first line is the Taylor series of $\log(1-x)$, the second line follows from $\frac{1}{j} \le \frac{1}{2}$ for $j \ge 2$, the third line follows from the geometric series, and the last line follows from $1 - x \ge 1/2$. Then, for all $k > m$, we have

$$\begin{aligned}
\prod_{i=m}^{k-1}\Big(1-\frac{\gamma}{i+k_0}\Big) &= \exp\Big\{\sum_{i=m}^{k-1}\log\Big(1-\frac{\gamma}{i+k_0}\Big)\Big\} \\
&\ge \exp\Big\{-\gamma\sum_{i=m}^{k-1}\frac{1}{i+k_0} - \gamma^2\sum_{i=m}^{k-1}\frac{1}{(i+k_0)^2}\Big\} \\
&\ge \exp\Big\{-\gamma\int_{m-1}^{k-1}\frac{dx}{x+k_0} - \gamma^2\sum_{i=m}^{\infty}\frac{1}{(i+k_0)^2}\Big\} \\
&= C_1\Big(\frac{m-1+k_0}{k-1+k_0}\Big)^{\gamma}, \qquad (30)
\end{aligned}$$

where the first line follows from the positivity of all factors, the second line follows from $\log(1-x) \ge -x - x^2$ with $x = \frac{\gamma}{i+k_0} \le 1/2$, the third line follows from $\frac{1}{i+k_0} \le \int_{i-1}^{i}\frac{dx}{x+k_0}$, since $x \mapsto \frac{1}{x+k_0}$ is decreasing, and from adding the nonnegative terms with $i \ge k$ to the second sum, and the last line follows from $\int_{m-1}^{k-1}\frac{dx}{x+k_0} = \log\frac{k-1+k_0}{m-1+k_0}$ with $C_1 := \exp\{-\gamma^2\sum_{i=m}^{\infty}(i+k_0)^{-2}\}$, which is positive because the series converges. Thus, combining (29) and (30), for all $k > m$ we have shown

$$\begin{aligned}
k\,y_k &\ge k\,y_{k'+1}\prod_{i=k'+1}^{m-1}\Big(1-\frac{\gamma}{i+k_0}\Big)\prod_{i=m}^{k-1}\Big(1-\frac{\gamma}{i+k_0}\Big) \\
&\ge C_2\,k\,\Big(\frac{m-1+k_0}{k-1+k_0}\Big)^{\gamma} \\
&= C_2\,(m-1+k_0)^{\gamma}\,\frac{k}{(k-1+k_0)^{\gamma}},
\end{aligned}$$

where the first line follows from (29) and splitting the product at $i = m$, the second line follows from (30) with $C_2 := C_1\,y_{k'+1}\prod_{i=k'+1}^{m-1}\big(1-\frac{\gamma}{i+k_0}\big)$, which is positive because $y_{k'+1} > 0$ and all factors are positive, and the last line is a rearrangement. Since $\gamma < 1$, the right-hand side tends to $+\infty$ as $k \to \infty$, and hence $k\,y_k \to \infty$. This completes the proof of the lemma. □

C Convergence Analysis

Lemma C.1. Under Assumption 3.2, we have

$$\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2 \mid \boldsymbol{X}_k\right] \le 2\|\nabla f(\boldsymbol{X}_k)\|_F^2 + 2^{4-\alpha}d^{1-\alpha/2}\tau^{2-\alpha}\sigma^{\alpha}.$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2 \mid \boldsymbol{X}_k\right] &\le 2\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k) - \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\|_F^2 \mid \boldsymbol{X}_k\right] + 2\mathbb{E}\left[\|\mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\|_F^2 \mid \boldsymbol{X}_k\right] \\
&\le 2\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k) - \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\|_F^2 \mid \boldsymbol{X}_k\right] + 2\|\nabla f(\boldsymbol{X}_k)\|_F^2 \\
&= 2\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k) - \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\|_F^2\mathbf{1}\left\{\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\| \ge 2\sqrt{d}\tau\right\} \mid \boldsymbol{X}_k\right] \\
&\quad +2\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k) - \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\|_F^2\mathbf{1}\left\{\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\| < 2\sqrt{d}\tau\right\} \mid \boldsymbol{X}_k\right] \\
&\quad +2\|\nabla f(\boldsymbol{X}_k)\|_F^2 \\
&\le 8d\tau^2\mathbb{E}\left[\mathbf{1}\left\{\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\| \ge 2\sqrt{d}\tau\right\} \mid \boldsymbol{X}_k\right] \\
&\quad +2\mathbb{E}\left[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F^2\mathbf{1}\left\{\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\| < 2\sqrt{d}\tau\right\} \mid \boldsymbol{X}_k\right] \\
&\quad +2\|\nabla f(\boldsymbol{X}_k)\|_F^2
\end{aligned}$$

where the last line follows from

$$\|\mathcal{C}_\tau(\boldsymbol{G}_k) - \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\|_F \le \min\left\{\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F, 2\sqrt{d}\tau\right\}.$$

Now we bound the first and second terms separately as follows. Note that

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F^2 \, \mathbf{1}\{\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F \le 2\sqrt{d}\tau\} \mid \boldsymbol{X}_k\right] \\
\le \quad & \left(2\sqrt{d}\tau\right)^{2-\alpha} \mathbb{E}\left[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F^{\alpha} \mid \boldsymbol{X}_k\right] \\
\le \quad & \left(2\sqrt{d}\tau\right)^{2-\alpha} \sigma^{\alpha}.
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}\left\{\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\| \ge 2\sqrt{d}\tau\right\} \mid \boldsymbol{X}_k\right] &= \Pr[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\| \ge 2\sqrt{d}\tau \mid \boldsymbol{X}_k] \\
&= \Pr[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|^{\alpha} \ge \left(2\sqrt{d}\tau\right)^{\alpha} \mid \boldsymbol{X}_k] \\
&\le \frac{\mathbb{E}\left[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|^{\alpha}\right]}{\left(2\sqrt{d}\tau\right)^{\alpha}} \\
&\le \left(2\sqrt{d}\tau\right)^{-\alpha} \sigma^{\alpha}.
\end{aligned}$$

Thus, combining the bounds, we have

$$\begin{aligned}
\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2 \mid \boldsymbol{X}_k\right] &\le 4 \cdot \left(2\sqrt{d}\tau\right)^{2-\alpha} \sigma^{\alpha} + 2\|\nabla f(\boldsymbol{X}_k)\|_F^2 \\
&= 2^{4-\alpha} d^{1-\alpha/2} \tau^{2-\alpha} \sigma^{\alpha} + 2\|\nabla f(\boldsymbol{X}_k)\|_F^2.
\end{aligned}$$

This completes the proof of the lemma.

□

C.1 Proof of Theorem 3.3

Theorem C.2. Suppose Assumptions 3.1 and 3.2 hold, and let $f_* := \inf_{\boldsymbol{X}} f(\boldsymbol{X}) > -\infty$. Then for any $K \geq 1$, SGD with spectral clipping (2) and a constant step size $\eta_k = \eta := \frac{1}{4L\sqrt{K}}$ satisfies

$$\min_{0\leq k\leq K-1} \mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F^2\right] \quad \leq \quad \frac{16L(f(\boldsymbol{X}_0)-f_*)}{\sqrt{K}} + \frac{C_{\alpha,d}\tau^{2-\alpha}\sigma^\alpha}{2\sqrt{K}} + \frac{2}{K}\sum_{k=0}^{K-1}\Delta_k$$

where $\Delta_k := \mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k) - \mathbb{E}\left[\mathcal{C}_{\tau_k}(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\|_F^2\right]$ and $C_{\alpha,d} := 2^{4-\alpha}d^{1-\alpha/2}$ with $d = \min\{n, m\}$.

Proof. For simplicity, let us denote the clipped gradient estimate by $\mathcal{C}_\tau(\boldsymbol{G}_k)$. i.e., given $\boldsymbol{G}_k = \boldsymbol{U}_k \mathrm{diag}(\boldsymbol{\sigma}_k)\boldsymbol{V}_k^\top$, the clipped gradient with threshold τ is defined as $\mathcal{C}_\tau(\boldsymbol{G}_k) := \boldsymbol{U}_k \mathrm{diag}(\mathrm{clamp}_\tau(\boldsymbol{\sigma}_k))\boldsymbol{V}_k^\top$, where the coordinates of the vector $\mathrm{clamp}_\tau(\boldsymbol{\sigma})$ are given by:

$$\mathrm{clamp}_\tau(\boldsymbol{\sigma})_i = \begin{cases} \boldsymbol{\sigma}_i & \text{if } \boldsymbol{\sigma}_i \leq \tau \\ \tau & \text{otherwise.} \end{cases}$$

Using this notation, the update rule can be written as

$$\boldsymbol{X}_{k+1} \quad = \quad \boldsymbol{X}_k - \eta_k \mathcal{C}_\tau(\boldsymbol{G}_k).$$

Then, from the smoothness property of the function f, we have

$$\begin{aligned} f(\boldsymbol{X}_{k+1}) \quad &\leq \quad f(\boldsymbol{X}_k) + \langle \nabla f(\boldsymbol{X}_k), \boldsymbol{X}_{k+1} - \boldsymbol{X}_k\rangle + \frac{L}{2}\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|_F^2 \\ &= \quad f(\boldsymbol{X}_k) - \eta_k \langle \nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\boldsymbol{G}_k)\rangle + \frac{\eta_k^2 L}{2}\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2 \\ &= \quad f(\boldsymbol{X}_k) - \eta_k \langle \nabla f(\boldsymbol{X}_k), \boldsymbol{G}_k\rangle - \eta_k \langle \nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\boldsymbol{G}_k) - \boldsymbol{G}_k\rangle + \frac{\eta_k^2 L}{2}\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2. \end{aligned}$$

Then, taking the expectation conditioned on $\boldsymbol{X}_k$, we get

$$\begin{aligned} \mathbb{E}\left[f(\boldsymbol{X}_{k+1}) \mid \boldsymbol{X}_k\right] \quad &\leq \quad f(\boldsymbol{X}_k) - \eta_k\|\nabla f(\boldsymbol{X}_k)\|_F^2 + \eta_k \langle \nabla f(\boldsymbol{X}_k), \mathbb{E}\left[\boldsymbol{G}_k - \mathcal{C}_\tau(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\rangle \\ &\quad + \frac{\eta_k^2 L}{2}\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2 \mid \boldsymbol{X}_k\right] \\ &\leq \quad f(\boldsymbol{X}_k) - \eta_k\|\nabla f(\boldsymbol{X}_k)\|_F^2 + \frac{\eta_k}{2}\|\nabla f(\boldsymbol{X}_k)\|_F^2 + \frac{\eta_k}{2}\left\|\mathbb{E}\left[\boldsymbol{G}_k - \mathcal{C}_\tau(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\right\|_F^2 \\ &\quad + \frac{\eta_k^2 L}{2}\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2 \mid \boldsymbol{X}_k\right] \\ &= \quad f(\boldsymbol{X}_k) - \frac{\eta_k}{2}\|\nabla f(\boldsymbol{X}_k)\|_F^2 + \frac{\eta_k}{2}\left\|\nabla f(\boldsymbol{X}_k) - \mathbb{E}\left[\mathcal{C}_\tau(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\right\|_F^2 \\ &\quad + \frac{\eta_k^2 L}{2}\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2 \mid \boldsymbol{X}_k\right]. \end{aligned}$$

From the Lemma, we have

$$\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2 \mid \boldsymbol{X}_k\right] \quad \leq \quad 2\|\nabla f(\boldsymbol{X}_k)\|_F^2 + C_{\alpha,d}\tau^{2-\alpha}\sigma^\alpha$$

where $C_{\alpha,d} := 2^{4-\alpha}d^{1-\alpha/2}$. Thus, the above inequality provides us

$$\begin{aligned} \mathbb{E}\left[f(\boldsymbol{X}_{k+1}) \mid \boldsymbol{X}_k\right] \quad &\leq \quad f(\boldsymbol{X}_k) - \frac{\eta_k}{2}\|\nabla f(\boldsymbol{X}_k)\|_F^2 + \frac{\eta_k}{2}\left\|\nabla f(\boldsymbol{X}_k) - \mathbb{E}\left[\mathcal{C}_\tau(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\right\|_F^2 \\ &\quad + \frac{\eta_k^2 L}{2}\left(2\|\nabla f(\boldsymbol{X}_k)\|_F^2 + C_{\alpha,d}\tau^{2-\alpha}\sigma^\alpha\right) \\ &= \quad f(\boldsymbol{X}_k) - \frac{\eta_k}{2}\left(1 - 2\eta_k L\right)\|\nabla f(\boldsymbol{X}_k)\|_F^2 + \frac{\eta_k}{2}\left\|\nabla f(\boldsymbol{X}_k) - \mathbb{E}\left[\mathcal{C}_\tau(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\right\|_F^2 \\ &\quad + \frac{\eta_k^2 L C_{\alpha,d}\tau^{2-\alpha}\sigma^\alpha}{2}. \end{aligned}$$

Now, let us define $\Delta_k := \mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k) - \mathbb{E}\left[\mathcal{C}_\tau(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\|_F^2\right]$. Then, by taking the total expectations on both sides, we obtain

$$\mathbb{E}\left[f(\boldsymbol{X}_{k+1})\right] \quad \leq \quad \mathbb{E}\left[f(\boldsymbol{X}_k)\right] - \frac{\eta_k}{2}\left(1 - 2\eta_k L\right)\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F^2\right] + \frac{\eta_k}{2}\Delta_k + \frac{\eta_k^2 L C_{\alpha,d}\tau^{2-\alpha}\sigma^\alpha}{2}.$$

Then, rearranging the terms, we have

$$\frac{\eta_k}{2}\left(1-2\eta_k L\right)\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F^2\right] \quad \leq \quad \mathbb{E}\left[f(\boldsymbol{X}_k)\right]-\mathbb{E}\left[f(\boldsymbol{X}_{k+1})\right]+\frac{\eta_k}{2}\Delta_k+\frac{\eta_k^2 L C_{\alpha,d}\tau^{2-\alpha}\sigma^{\alpha}}{2}.$$

Now, if we choose $\eta_k \leq \frac{1}{4L}$ for all k, then we can lower bound the left hand side with $\frac{\eta_k}{4}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F^2\right] \leq \frac{\eta_k}{2}\left(1-2\eta_k L\right)\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F^2\right]$. Hence, we obtain

$$\frac{\eta_k}{4}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F^2\right] \quad \leq \quad \mathbb{E}\left[f(\boldsymbol{X}_k)\right]-\mathbb{E}\left[f(\boldsymbol{X}_{k+1})\right]+\frac{\eta_k}{2}\Delta_k+\frac{\eta_k^2 L C_{\alpha,d}\tau^{2-\alpha}\sigma^{\alpha}}{2}.$$

Now, summing up both sides for $k=0,1,\cdots K-1$, we get

$$\begin{aligned}\sum_{k=0}^{K-1}\frac{\eta_k}{4}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F^2\right] &\leq \mathbb{E}\left[f(\boldsymbol{X}_0)\right]-\mathbb{E}\left[f(\boldsymbol{X}_K)\right]+\frac{1}{2}\sum_{k=0}^{K-1}\eta_k\Delta_k+\frac{LC_{\alpha,d}\tau^{2-\alpha}\sigma^{\alpha}}{2}\sum_{k=0}^{K-1}\eta_k^2\\ &\leq f(\boldsymbol{X}_0)-f_*+\frac{1}{2}\sum_{k=0}^{K-1}\eta_k\Delta_k+\frac{LC_{\alpha,d}\tau^{2-\alpha}\sigma^{\alpha}}{2}\sum_{k=0}^{K-1}\eta_k^2.\end{aligned}$$

The last line follows from $\mathbb{E}\left[f(\boldsymbol{X}_K)\right] \geq f_*$. For constant step size $\eta_k=\eta$, we obtain.

$$\frac{\eta}{4}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F^2\right] \quad \leq \quad f(\boldsymbol{X}_0)-f_*+\frac{\eta}{2}\sum_{k=0}^{K-1}\Delta_k+\frac{\eta^2 LC_{\alpha,d}\tau^{2-\alpha}\sigma^{\alpha}K}{2}.$$

Now divide both sides by $\frac{\eta K}{4}$ to get

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F^2\right] \quad \leq \quad \frac{4(f(\boldsymbol{X}_0)-f_*)}{\eta K}+\frac{2}{K}\sum_{k=0}^{K-1}\Delta_k+2\eta LC_{\alpha,d}\tau^{2-\alpha}\sigma^{\alpha}.$$

Therefore, by lower bounding the left hand side, we obtain

$$\min_{0\leq k\leq K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F^2\right] \quad \leq \quad \frac{4(f(\boldsymbol{X}_0)-f_*)}{\eta K}+\frac{2}{K}\sum_{k=0}^{K-1}\Delta_k+2\eta LC_{\alpha,d}\tau^{2-\alpha}\sigma^{\alpha}.$$

For $\eta=\frac{1}{4L\sqrt{K}}$, we obtain

$$\min_{0\leq k\leq K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F^2\right] \quad \leq \quad \frac{16L(f(\boldsymbol{X}_0)-f_*)}{\sqrt{K}}+\frac{2}{K}\sum_{k=0}^{K-1}\Delta_k+\frac{C_{\alpha,d}\tau^{2-\alpha}\sigma^{\alpha}}{2\sqrt{K}}.$$

This completes the proof of the theorem. □

C.2 Proof of Theorem 3.4

Theorem C.3. Suppose Assumptions 3.1 and 3.2 hold, and let $f_* := \inf_{\boldsymbol{X}} f(\boldsymbol{X}) > -\infty$. Then, for any $K \geq 1$, SGD with spectral clipping (2), a constant step size $\eta_k = \eta \leq 1/4dL$ and clipping threshold $\tau \geq \max\{2, 8\sigma\}$ satisfies the following

$$\min_{0\leq k\leq K-1}\mathbb{E}\left[\min\left\{\|\nabla f(\boldsymbol{X}_k)\|_F, \|\nabla f(\boldsymbol{X}_k)\|_F^2\right\}\right] \leq \frac{4(f(\boldsymbol{X}_0)-f_*)}{\eta K}+32\tau^{-2(\alpha-1)}\sigma^{2\alpha}+16\eta dL\tau^{2-\alpha}\sigma^{\alpha}.$$

Proof. For simplicity, let us denote the clipped gradient estimate with $\mathcal{C}_\tau(\boldsymbol{G}_k)$. i.e., given $\boldsymbol{G}_k = \boldsymbol{U}_k\mathrm{diag}(\boldsymbol{\sigma}_k)\boldsymbol{V}_k^\top$, the clipped gradient with threshold τ is defined as $\mathcal{C}_\tau(\boldsymbol{G}_k) := \boldsymbol{U}_k\mathrm{diag}(\mathrm{clamp}_\tau(\boldsymbol{\sigma}_k))\boldsymbol{V}_k^\top$, where the coordinates of the vector $\mathrm{clamp}_\tau(\boldsymbol{\sigma})$ are given by:

$$\mathrm{clamp}_\tau(\boldsymbol{\sigma})_i = \begin{cases}\boldsymbol{\sigma}_i & \text{if } \boldsymbol{\sigma}_i \leq \tau\\ \tau & \text{otherwise.}\end{cases}$$

Using this notation, the update rule can be written as

$$\boldsymbol{X}_{k+1} \quad = \quad \boldsymbol{X}_k-\eta_k\mathcal{C}_\tau(\boldsymbol{G}_k).$$

Then, from the smoothness property of the function f, we have

$$\begin{aligned}f(\boldsymbol{X}_{k+1}) &\leq f(\boldsymbol{X}_k)+\langle\nabla f(\boldsymbol{X}_k), \boldsymbol{X}_{k+1}-\boldsymbol{X}_k\rangle+\frac{L}{2}\|\boldsymbol{X}_{k+1}-\boldsymbol{X}_k\|_F^2\\ &= f(\boldsymbol{X}_k)-\eta_k\langle\nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\boldsymbol{G}_k)\rangle+\frac{\eta_k^2 L}{2}\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2\\ &= f(\boldsymbol{X}_k)-\eta_k\langle\nabla f(\boldsymbol{X}_k), \boldsymbol{G}_k\rangle-\eta_k\langle\nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\boldsymbol{G}_k)-\boldsymbol{G}_k\rangle+\frac{\eta_k^2 L}{2}\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2.\end{aligned}$$

Then, taking the expectation conditioned on $\boldsymbol{X}_k$, we obtain

$$\begin{aligned}
\mathbb{E}\left[f(\boldsymbol{X}_{k+1}) \mid \boldsymbol{X}_k\right] &\leq f(\boldsymbol{X}_k) - \eta_k\|\nabla f(\boldsymbol{X}_k)\|_F^2 + \eta_k\left\langle\nabla f(\boldsymbol{X}_k), \mathbb{E}\left[\boldsymbol{G}_k - \mathcal{C}_\tau(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\right\rangle \\
&\quad + \frac{\eta_k^2 L}{2}\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2 \mid \boldsymbol{X}_k\right] \\
&\leq f(\boldsymbol{X}_k) - \eta_k\|\nabla f(\boldsymbol{X}_k)\|_F^2 + \frac{\eta_k}{2}\|\nabla f(\boldsymbol{X}_k)\|_F^2 + \frac{\eta_k}{2}\left\|\mathbb{E}\left[\boldsymbol{G}_k - \mathcal{C}_\tau(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\right\|_F^2 \\
&\quad + \frac{\eta_k^2 L}{2}\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2 \mid \boldsymbol{X}_k\right] \\
&= f(\boldsymbol{X}_k) - \frac{\eta_k}{2}\|\nabla f(\boldsymbol{X}_k)\|_F^2 + \frac{\eta_k}{2}\left\|\nabla f(\boldsymbol{X}_k) - \mathbb{E}\left[\mathcal{C}_\tau(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\right\|_F^2 \\
&\quad + \frac{\eta_k^2 L}{2}\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2 \mid \boldsymbol{X}_k\right].
\end{aligned}$$

For further analysis, we will consider two cases: either $\|\nabla f(\boldsymbol{X}_k)\|_F \leq \tau/2$ or $\|\nabla f(\boldsymbol{X}_k)\|_F > \tau/2$.

Case 1. Suppose $\|\nabla f(\boldsymbol{X}_k)\|_F \leq \tau/2$. Then we have the following bound

$$\begin{aligned}
\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2 &= \|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2\mathbf{1}\left\{\|\boldsymbol{G}_k\|_F \leq \tau\right\} + \|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2\mathbf{1}\left\{\|\boldsymbol{G}_k\|_F > \tau\right\} \\
&\leq \|\boldsymbol{G}_k\|_F^2\mathbf{1}\left\{\|\boldsymbol{G}_k\|_F \leq \tau\right\} + d\tau^2\mathbf{1}\left\{\|\boldsymbol{G}_k\|_F > \tau\right\} \\
&\leq 2\|\nabla f(\boldsymbol{X}_k)\|_F^2 + 2\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F^2\mathbf{1}\left\{\|\boldsymbol{G}_k\|_F \leq \tau\right\} + d\tau^2\mathbf{1}\left\{\|\boldsymbol{G}_k\|_F > \tau\right\} \\
&\leq 2\|\nabla f(\boldsymbol{X}_k)\|_F^2 + 2\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F^2\mathbf{1}\left\{\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F \leq 3\tau/2\right\} + d\tau^2\mathbf{1}\left\{\|\boldsymbol{G}_k\|_F > \tau\right\} \\
&\leq 2\|\nabla f(\boldsymbol{X}_k)\|_F^2 + 2(3\tau/2)^{2-\alpha}\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F^\alpha + d\tau^2\mathbf{1}\left\{\|\boldsymbol{G}_k\|_F > \tau\right\}
\end{aligned}$$

Then, taking the expectation, we have

$$\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2 \mid \boldsymbol{X}_k\right] \leq 2\|\nabla f(\boldsymbol{X}_k)\|_F^2 + 2(3\tau/2)^{2-\alpha}\sigma^\alpha + d\tau^2\mathbb{E}\left[\mathbf{1}\left\{\|\boldsymbol{G}_k\|_F > \tau\right\} \mid \boldsymbol{X}_k\right]$$

Moreover, for $\|\nabla f(\boldsymbol{X}_k)\|_F \leq \tau/2$, we also have $\mathbf{1}\left\{\|\boldsymbol{G}_k\|_F > \tau\right\} \leq \mathbf{1}\left\{\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F > \tau/2\right\}$. Thus, using Markov's Inequality, we obtain

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}\left\{\|\boldsymbol{G}_k\|_F > \tau\right\} \mid \boldsymbol{X}_k\right] &\leq \mathbb{E}\left[\mathbf{1}\left\{\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F > \tau/2\right\} \mid \boldsymbol{X}_k\right] \\
&= \Pr[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F > \tau/2 \mid \boldsymbol{X}_k] \\
&= \Pr[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F^\alpha > (\tau/2)^\alpha \mid \boldsymbol{X}_k] \\
&\leq \frac{\mathbb{E}\left[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F^\alpha \mid \boldsymbol{X}_k\right]}{(\tau/2)^\alpha} \\
&\leq 2^\alpha\tau^{-\alpha}\sigma^\alpha.
\end{aligned}$$

Thus, combining the bounds, we have shown

$$\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2 \mid \boldsymbol{X}_k\right] \leq 2\|\nabla f(\boldsymbol{X}_k)\|_F^2 + C_{\alpha,1}\tau^{2-\alpha}\sigma^\alpha$$

where $C_{\alpha,1} := 2^\alpha d + 3\cdot(2/3)^{\alpha-1} \leq 8d$. Similarly, we want to bound $\|\nabla f(\boldsymbol{X}_k) - \mathbb{E}\left[\mathcal{C}_\tau(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\|_F^2$. Note that

$$\begin{aligned}
\|\nabla f(\boldsymbol{X}_k) - \mathbb{E}\left[\mathcal{C}_\tau(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\|_F^2 &= \left\|\mathbb{E}\left[\boldsymbol{G}_k - \mathcal{C}_\tau(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\right\|_F^2 \\
&\leq \mathbb{E}\left[\|\boldsymbol{G}_k - \mathcal{C}_\tau(\boldsymbol{G}_k)\|_F \mid \boldsymbol{X}_k\right]^2 \\
&= \Big(\mathbb{E}\left[\|\boldsymbol{G}_k - \mathcal{C}_\tau(\boldsymbol{G}_k)\|_F\mathbf{1}\left\{\|\boldsymbol{G}_k\|_F \leq \tau\right\} \mid \boldsymbol{X}_k\right] \\
&\quad + \mathbb{E}\left[\|\boldsymbol{G}_k - \mathcal{C}_\tau(\boldsymbol{G}_k)\|_F\mathbf{1}\left\{\|\boldsymbol{G}_k\|_F > \tau\right\} \mid \boldsymbol{X}_k\right]\Big)^2 \\
&= \mathbb{E}\left[\|\boldsymbol{G}_k - \mathcal{C}_\tau(\boldsymbol{G}_k)\|_F\mathbf{1}\left\{\|\boldsymbol{G}_k\|_F > \tau\right\} \mid \boldsymbol{X}_k\right]^2 \\
&\leq \mathbb{E}\left[\|\boldsymbol{G}_k\|_F\mathbf{1}\left\{\|\boldsymbol{G}_k\|_F > \tau\right\} \mid \boldsymbol{X}_k\right]^2 \\
&\leq \mathbb{E}\left[\|\boldsymbol{G}_k\|_F\mathbf{1}\left\{\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F > \tau/2\right\} \mid \boldsymbol{X}_k\right]^2
\end{aligned}$$

where the last inequality follows from the observation that, for $\|\nabla f(\boldsymbol{X}_k)\|_F \leq \tau/2$, we have $\mathbf{1}\left\{\|\boldsymbol{G}_k\|_F > \tau\right\} \leq \mathbf{1}\left\{\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F > \tau/2\right\}$. Then

$$\begin{aligned}
&\|\nabla f(\boldsymbol{X}_k) - \mathbb{E}\left[\mathcal{C}_\tau(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\|_F^2 \\
&\leq \mathbb{E}\left[\|\boldsymbol{G}_k\|_F\mathbf{1}\left\{\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F > \tau/2\right\} \mid \boldsymbol{X}_k\right]^2 \\
&\leq \mathbb{E}\left[\left(\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F + \|\nabla f(\boldsymbol{X}_k)\|_F\right)\mathbf{1}\left\{\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F > \tau/2\right\} \mid \boldsymbol{X}_k\right]^2 \\
&\leq 4\mathbb{E}\left[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F\mathbf{1}\left\{\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F > \tau/2\right\} \mid \boldsymbol{X}_k\right]^2
\end{aligned}$$

where the last line follows from $\|\nabla f(\boldsymbol{X}_k)\|_F \leq \tau/2$ Then we apply Holder's Inequality to get

$$\begin{aligned}
& \mathbb{E}\left[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F \mathbf{1}\left\{\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F > \tau/2\right\} \mid \boldsymbol{X}_k\right] \\
\leq & \left(\mathbb{E}\left[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F^\alpha \mid \boldsymbol{X}_k\right]\right)^{1/\alpha} \cdot \left(\Pr[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F > \tau/2 \mid \boldsymbol{X}_k]\right)^{1-1/\alpha} \\
\leq & (\sigma^\alpha)^{1/\alpha} \cdot \left(\frac{\mathbb{E}\left[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F^\alpha \mid \boldsymbol{X}_k\right]}{(\tau/2)^\alpha}\right)^{\alpha-1/\alpha} \\
= & \sigma \cdot \left(\frac{2\sigma}{\tau}\right)^{\alpha-1}.
\end{aligned}$$

Thus, by combining the bounds, we obtain.

$$\|\nabla f(\boldsymbol{X}_k) - \mathbb{E}\left[\mathcal{C}_\tau(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\|_F^2 \leq C_{\alpha,2}\tau^{-2(\alpha-1)}\sigma^{2\alpha}$$

where $C_{\alpha,2} := 2^{2\alpha} \leq 16$. Hence, for the $\|\nabla f(\boldsymbol{X}_k)\|_F \leq \tau/2$ case, we have

$$\begin{aligned}
\mathbb{E}\left[f(\boldsymbol{X}_{k+1}) \mid \boldsymbol{X}_k\right] \leq & f(\boldsymbol{X}_k) - \frac{\eta_k}{2}\left(1 - 2\eta_k L\right)\|\nabla f(\boldsymbol{X}_k)\|_F^2 + \frac{\eta_k}{2}C_{\alpha,2}\tau^{-2(\alpha-1)}\sigma^{2\alpha} \\
& + \frac{\eta_k^2 L}{2}C_{\alpha,1}\tau^{2-\alpha}\sigma^\alpha.
\end{aligned}$$

Then for $\eta_k = \eta \leq \frac{1}{4L}$, we obtain

$$\begin{aligned}
\mathbb{E}\left[f(\boldsymbol{X}_{k+1}) \mid \boldsymbol{X}_k\right] & \leq f(\boldsymbol{X}_k) - \frac{\eta_k}{4}\|\nabla f(\boldsymbol{X}_k)\|_F^2 + \frac{\eta_k}{2}C_{\alpha,2}\tau^{-2(\alpha-1)}\sigma^{2\alpha} + \frac{\eta_k^2 L}{2}C_{\alpha,1}\tau^{2-\alpha}\sigma^\alpha \\
& \leq f(\boldsymbol{X}_k) - \frac{\eta_k}{4}\min\left\{\|\nabla f(\boldsymbol{X}_k)\|_F, \|\nabla f(\boldsymbol{X}_k)\|_F^2\right\} + \frac{\eta_k}{2}C_{\alpha,2}\tau^{-2(\alpha-1)}\sigma^{2\alpha} \\
& \quad + \frac{\eta_k^2 L}{2}C_{\alpha,1}\tau^{2-\alpha}\sigma^\alpha.
\end{aligned}$$

Case 2. Now, consider $\|\nabla f(\boldsymbol{X}_k)\|_F > \tau/2$. Similar to the previous case, we also have the following bound

$$\begin{aligned}
f(\boldsymbol{X}_{k+1}) & \leq f(\boldsymbol{X}_k) + \langle \nabla f(\boldsymbol{X}_k), \boldsymbol{X}_{k+1} - \boldsymbol{X}_k\rangle + \frac{L}{2}\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|_F^2 \\
& = f(\boldsymbol{X}_k) - \eta_k\left\langle \nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\boldsymbol{G}_k)\right\rangle + \frac{\eta_k^2 L}{2}\|\mathcal{C}_\tau(\boldsymbol{G}_k)\|_F^2 \\
& \leq f(\boldsymbol{X}_k) - \eta_k\left\langle \nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\boldsymbol{G}_k)\right\rangle + \frac{\eta_k^2 dL\tau^2}{2}
\end{aligned}$$

Here, we take the expectation conditioned on $\boldsymbol{X}_k$ to obtain

$$\begin{aligned}
\mathbb{E}\left[f(\boldsymbol{X}_{k+1}) \mid \boldsymbol{X}_k\right] & \leq f(\boldsymbol{X}_k) - \eta_k\langle \nabla f(\boldsymbol{X}_k), \mathbb{E}\left[\mathcal{C}_\tau(\boldsymbol{G}_k) \mid \boldsymbol{X}_k\right]\rangle + \frac{\eta_k^2 dL\tau^2}{2} \\
& = f(\boldsymbol{X}_k) - \eta_k\langle \nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\rangle \\
& \quad - \eta_k\langle \nabla f(\boldsymbol{X}_k), \mathbb{E}\left[\mathcal{C}_\tau(\boldsymbol{G}_k) - \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k)) \mid \boldsymbol{X}_k\right]\rangle + \frac{\eta_k^2 dL\tau^2}{2} \\
& = f(\boldsymbol{X}_k) - \eta_k\langle \nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\rangle \\
& \quad + \eta_k\|\nabla f(\boldsymbol{X}_k)\|_F\|\mathbb{E}\left[\mathcal{C}_\tau(\boldsymbol{G}_k) - \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k)) \mid \boldsymbol{X}_k\right]\|_F + \frac{\eta_k^2 dL\tau^2}{2} \\
& \leq f(\boldsymbol{X}_k) - \eta_k\langle \nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\rangle \\
& \quad + \eta_k\|\nabla f(\boldsymbol{X}_k)\|_F\mathbb{E}\left[\|\mathcal{C}_\tau(\boldsymbol{G}_k) - \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\|_F \mid \boldsymbol{X}_k\right] + \frac{\eta_k^2 dL\tau^2}{2} \\
& \leq f(\boldsymbol{X}_k) - \eta_k\langle \nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\rangle \\
& \quad + \eta_k\|\nabla f(\boldsymbol{X}_k)\|_F\mathbb{E}\left[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F \mid \boldsymbol{X}_k\right] + \frac{\eta_k^2 dL\tau^2}{2} \\
& \leq f(\boldsymbol{X}_k) - \eta_k\langle \nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\rangle \\
& \quad + \eta_k\|\nabla f(\boldsymbol{X}_k)\|_F\mathbb{E}\left[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F^\alpha \mid \boldsymbol{X}_k\right]^{1/\alpha} + \frac{\eta_k^2 dL\tau^2}{2} \\
& \leq f(\boldsymbol{X}_k) - \eta_k\langle \nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\rangle + \eta_k\sigma\|\nabla f(\boldsymbol{X}_k)\|_F + \frac{\eta_k^2 dL\tau^2}{2}.
\end{aligned}$$

where the second-to-last inequality follows from Jensen's Inequality with $\alpha \in (1, 2]$. Now we want to lower bound the term $\langle \nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\rangle$ when $\|\nabla f(\boldsymbol{X}_k)\|_F > \tau/2$. Note that

$$\langle \nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\rangle = \sum_{i=1}^d \sigma_i \cdot \min\left\{\sigma_i, \tau\right\}$$

where $\{\sigma_i\}_{i=1}^d$ denotes the singular values of $\nabla f(\boldsymbol{X}_k)$. Now we consider two cases:

Case 1. $\tau/2 < \|\nabla f(\boldsymbol{X}_k)\|_F \leq \tau$. In this case, as $\|\nabla f(\boldsymbol{X}_k)\|_F \leq \tau$, all the singular values of $\nabla f(\boldsymbol{X}_k)$ are less than τ. Therefore

$$\begin{aligned}\langle \nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\rangle &= \sum_{i=1}^d \sigma_i^2 \\ &= \|\nabla f(\boldsymbol{X}_k)\|_F^2 \\ &\geq \frac{\tau}{2}\|\nabla f(\boldsymbol{X}_k)\|_F.\end{aligned}$$

Case 2. $\|\nabla f(\boldsymbol{X}_k)\|_F > \tau$. In this case, For each singular value σ_i, we have

$$\min\{\sigma_i, \tau\} \geq \frac{\sigma_i \tau}{\|\nabla f(\boldsymbol{X}_k)\|_F}.$$

Therefore,

$$\begin{aligned}\langle \nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\rangle &= \sum_{i=1}^d \sigma_i \cdot \min\{\sigma_i, \tau\} \\ &\geq \frac{\tau}{\|\nabla f(\boldsymbol{X}_k)\|_F}\sum_{i=1}^d \sigma_i^2 \\ &= \tau\|\nabla f(\boldsymbol{X}_k)\|_F \\ &\geq \frac{\tau}{2}\|\nabla f(\boldsymbol{X}_k)\|_F.\end{aligned}$$

Therefore, we have shown $\langle \nabla f(\boldsymbol{X}_k), \mathcal{C}_\tau(\nabla f(\boldsymbol{X}_k))\rangle \geq \frac{\tau}{2}\|\nabla f(\boldsymbol{X}_k)\|_F$ when $\|\nabla f(\boldsymbol{X}_k)\|_F > \tau/2$. Finally, combining the bounds, we get

$$\begin{aligned}\mathbb{E}\left[f(\boldsymbol{X}_{k+1}) \mid \boldsymbol{X}_k\right] &\leq f(\boldsymbol{X}_k) - \frac{\eta_k\tau}{2}\|\nabla f(\boldsymbol{X}_k)\|_F + \eta_k\sigma\|\nabla f(\boldsymbol{X}_k)\|_F + \frac{\eta_k^2 dL\tau^2}{2} \\ &\leq f(\boldsymbol{X}_k) - \frac{\eta_k\tau}{2}\|\nabla f(\boldsymbol{X}_k)\|_F + \eta_k\sigma\|\nabla f(\boldsymbol{X}_k)\|_F + \eta_k^2 dL\tau\|\nabla f(\boldsymbol{X}_k)\|_F \\ &\leq f(\boldsymbol{X}_k) - \frac{\eta_k}{2}\left(\tau(1-2\eta_k dL) - 2\sigma\right)\|\nabla f(\boldsymbol{X}_k)\|_F.\end{aligned}$$

Here, we choose $\eta_k \leq 1/4dL$ to obtain

$$\mathbb{E}\left[f(\boldsymbol{X}_{k+1}) \mid \boldsymbol{X}_k\right] \leq f(\boldsymbol{X}_k) - \frac{\eta_k}{2}\left(\frac{\tau}{2} - 2\sigma\right)\|\nabla f(\boldsymbol{X}_k)\|_F.$$

For $\tau \geq 8\sigma$, we can further bound it as shown below

$$\begin{aligned}\mathbb{E}\left[f(\boldsymbol{X}_{k+1}) \mid \boldsymbol{X}_k\right] &\leq f(\boldsymbol{X}_k) - \frac{\eta_k\tau}{8}\|\nabla f(\boldsymbol{X}_k)\|_F \\ &\leq f(\boldsymbol{X}_k) - \frac{\eta_k\tau}{8}\|\nabla f(\boldsymbol{X}_k)\|_F + \frac{\eta_k}{2}C_{\alpha,2}\tau^{-2(\alpha-1)}\sigma^{2\alpha} + \frac{\eta_k^2 L}{2}C_{\alpha,1}\tau^{2-\alpha}\sigma^\alpha.\end{aligned}$$

For $\tau \geq 2$, we have

$$\begin{aligned}\mathbb{E}\left[f(\boldsymbol{X}_{k+1}) \mid \boldsymbol{X}_k\right] &\leq f(\boldsymbol{X}_k) - \frac{\eta_k}{4}\|\nabla f(\boldsymbol{X}_k)\|_F + \frac{\eta_k}{2}C_{\alpha,2}\tau^{-2(\alpha-1)}\sigma^{2\alpha} + \frac{\eta_k^2 L}{2}C_{\alpha,1}\tau^{2-\alpha}\sigma^\alpha \\ &\leq f(\boldsymbol{X}_k) - \frac{\eta_k}{4}\min\left\{\|\nabla f(\boldsymbol{X}_k)\|_F, \|\nabla f(\boldsymbol{X}_k)\|_F^2\right\} + \frac{\eta_k}{2}C_{\alpha,2}\tau^{-2(\alpha-1)}\sigma^{2\alpha} \\ &\quad + \frac{\eta_k^2 L}{2}C_{\alpha,1}\tau^{2-\alpha}\sigma^\alpha.\end{aligned}$$

Finally, taking the total expectation, rearranging, summing up the inequality for $k = 0, 1, \cdots, K-1$, and dividing by K, we have

$$\begin{aligned}\frac{\eta}{4K}\sum_{k=0}^{K-1}\mathbb{E}\left[\min\left\{\|\nabla f(\boldsymbol{X}_k)\|_F, \|\nabla f(\boldsymbol{X}_k)\|_F^2\right\}\right] &\leq \frac{f(\boldsymbol{X}_0) - f_*}{K} + \frac{\eta}{2}C_{\alpha,2}\tau^{-2(\alpha-1)}\sigma^{2\alpha} \\ &\quad + \frac{\eta^2 L}{2}C_{\alpha,1}\tau^{2-\alpha}\sigma^\alpha.\end{aligned}$$

Dividing both sides by $\eta/4$, we obtain

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\min\left\{\|\nabla f(\boldsymbol{X}_k)\|_F, \|\nabla f(\boldsymbol{X}_k)\|_F^2\right\}\right] \leq \frac{4(f(\boldsymbol{X}_0)-f_*)}{\eta K} + 2C_{\alpha,2}\tau^{-2(\alpha-1)}\sigma^{2\alpha} + 2\eta L C_{\alpha,1}\tau^{2-\alpha}\sigma^{\alpha}.$$

This implies

$$\begin{aligned}\min_{0\leq k\leq K-1}\mathbb{E}\left[\min\left\{\|\nabla f(\boldsymbol{X}_k)\|_F, \|\nabla f(\boldsymbol{X}_k)\|_F^2\right\}\right] &\leq \frac{4(f(\boldsymbol{X}_0)-f_*)}{\eta K} + 2C_{\alpha,2}\tau^{-2(\alpha-1)}\sigma^{2\alpha} \\ &\quad + 2\eta L C_{\alpha,1}\tau^{2-\alpha}\sigma^{\alpha} \\ &\leq \frac{4(f(\boldsymbol{X}_0)-f_*)}{\eta K} + 32\tau^{-2(\alpha-1)}\sigma^{2\alpha} \\ &\quad + 16\eta d L\tau^{2-\alpha}\sigma^{\alpha}.\end{aligned}$$

For the parameter choices in Theorem 3.4, with α, $\sigma > 0$, d, and L fixed,

$$\tau = \Theta\left(K^{1/(3\alpha-2)}\right),$$
$$\eta = \Theta\left(K^{-\alpha/(3\alpha-2)}\right).$$

The three terms in the finite-time bound therefore satisfy

$$\frac{4(f(\boldsymbol{X}_0)-f_*)}{\eta K} = \mathcal{O}\left(K^{-2(\alpha-1)/(3\alpha-2)}\right),$$
$$32\tau^{-2(\alpha-1)}\sigma^{2\alpha} = \mathcal{O}\left(K^{-2(\alpha-1)/(3\alpha-2)}\right),$$
$$16\eta d L\tau^{2-\alpha}\sigma^{\alpha} = \mathcal{O}\left(K^{-2(\alpha-1)/(3\alpha-2)}\right).$$

Substituting these bounds proves (6). □

D PROOF OF THEOREM 3.6

In this section, we prove Theorem 3.6. The argument has three parts. First, we ask how each clipping map acts on a small gradient added to a large noise sample. To first order, this is described by the derivative $D\Pi(\boldsymbol{\Xi})$ of the map at the noise sample, averaged over the noise. Since Π commutes with rotations (Lemma D.2) and the law of $\boldsymbol{\Xi}$ is invariant under rotations (Lemma D.1), this averaged derivative commutes with rotations as well, and a linear map on matrices with this property is a scalar (Lemma D.3). In other words, on average, clipping simply multiplies a small gradient by a number λ (Lemma D.6). This number equals the trace of $D\Pi(\boldsymbol{\Xi})$ divided by d^2, so it remains to compute the trace, which we do explicitly for both maps (Lemma D.4).

Second, we track the expected squared error of the iterates row by row. Its recursion involves the clipped update only through two quantities: the mean map $\boldsymbol{M}$, which is linear to first order with slope λ (Lemma D.7), and the energy q_0 of the clipped noise at the solution (Lemma D.5). Since the iterates stay close to the solution with overwhelming probability (Lemma D.8), Chung's lemma (Lemma B.11) turns this recursion into the exact limit of $k\big(\mathbb{E}\left[f(\boldsymbol{X}_k)\right] - f_*\big)$, which equals $\frac{q_0}{\lambda^2}\Psi_{\boldsymbol{H}}(c\lambda)$ for either method (Proposition D.9).

Third, we compare the ratio q_0/λ^2 of the two methods: for norm clipping, it is bounded from below uniformly in τ_{F} (Lemma D.10), while for spectral clipping, it is strictly smaller (Lemma D.11), which proves the theorem. Remark D.12 discusses two extensions of this setting: independent amplitudes for the r noise directions, and the case $r = d$.

Setting and Notation. Both methods of Theorem 3.6 start from a deterministic $\boldsymbol{X}_0$ and follow

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \eta_k \Pi(\boldsymbol{G}_k), \qquad \eta_k = \frac{c}{k + k_0}, \qquad c, k_0 > 0, \tag{31}$$

where $\boldsymbol{G}_k = \nabla f(\boldsymbol{X}_k) + \boldsymbol{\Xi}_k$ is the stochastic gradient of Assumption 3.5, and Π is the clipping map: $\Pi = \mathcal{C}_{\tau_{\mathrm{sp}}}$ for spectral clipping, and $\Pi = \mathrm{clip}(\cdot, \tau_{\mathrm{F}})$ with the Frobenius norm for norm clipping. Under either condition on τ_{sp} in Theorem 3.6, we have $\tau_{\mathrm{sp}} < a_0$, so spectral clipping is active on every noise sample. Let $(A, \boldsymbol{U}, \boldsymbol{V})$ have the law of $(A_k, \boldsymbol{U}_k, \boldsymbol{V}_k)$, and let $\boldsymbol{\Xi} := A\boldsymbol{U}\boldsymbol{V}^\top$ be one noise sample.

Lemma D.1 (Properties of the noise)**.** Let Assumption 3.5 hold. Then the following statements hold.

(a) For any $\alpha < \kappa$, the amplitude A satisfies

$$\mathbb{E}\left[A^{-1}\right] = \frac{\kappa}{(\kappa+1)a_0}, \qquad \mathbb{E}\left[A^{-2}\right] = \frac{\kappa}{(\kappa+2)a_0^2}, \qquad \mathbb{E}\left[A^{\alpha}\right] = \frac{\kappa a_0^{\alpha}}{\kappa - \alpha},$$

while $\mathbb{E}\left[A^2\right] = \infty$.

(b) The noise $\boldsymbol{\Xi}_k$ satisfies $\mathbb{E}\left[\boldsymbol{\Xi}_k\right] = 0$ and $\|\boldsymbol{\Xi}_k\|_F = \sqrt{r}A_k$. Hence, Assumption 3.2 holds for every $\alpha \in (1, \kappa)$ with

$$\sigma^{\alpha} = \frac{r^{\alpha/2}\kappa a_0^{\alpha}}{\kappa - \alpha},$$

while $\mathbb{E}\left[\|\boldsymbol{\Xi}_k\|_F^2\right] = \infty$.

(c) For any orthogonal matrices $\boldsymbol{Q}_1, \boldsymbol{Q}_2 \in \mathbb{R}^{d\times d}$, the matrix $\boldsymbol{Q}_1\boldsymbol{\Xi}_k\boldsymbol{Q}_2^\top$ has the same law as $\boldsymbol{\Xi}_k$.

Proof. (a) By Assumption 3.5, A has the density $\kappa a_0^{\kappa} t^{-\kappa-1}$ on $[a_0, \infty)$. Therefore, for any $s < \kappa$, we have

$$\begin{aligned}\mathbb{E}\left[A^s\right] &= \int_{a_0}^{\infty} t^s \cdot \kappa a_0^{\kappa} t^{-\kappa-1}\, dt \\ &= \frac{\kappa a_0^s}{\kappa - s},\end{aligned}$$

while for $s \geq \kappa$ the integral diverges. Taking $s = -1$, $s = -2$ and $s = \alpha$ gives the three identities, and since $\kappa < 2$, taking $s = 2$ gives $\mathbb{E}\left[A^2\right] = \infty$.

(b) Since the columns of $\boldsymbol{U}_k$ and $\boldsymbol{V}_k$ are orthonormal, i.e., $\boldsymbol{U}_k^\top\boldsymbol{U}_k = \boldsymbol{V}_k^\top\boldsymbol{V}_k = \boldsymbol{I}$, we have

$$\begin{aligned}\|\boldsymbol{U}_k\boldsymbol{V}_k^\top\|_F^2 &= \mathrm{tr}\left(\boldsymbol{V}_k\boldsymbol{U}_k^\top\boldsymbol{U}_k\boldsymbol{V}_k^\top\right) \\ &= \mathrm{tr}\left(\boldsymbol{V}_k^\top\boldsymbol{V}_k\right) \\ &= r.\end{aligned}$$

Hence, $\|\boldsymbol{\Xi}_k\|_F = A_k\|\boldsymbol{U}_k\boldsymbol{V}_k^\top\|_F = \sqrt{r}A_k$. Moreover, the Haar distribution is invariant under orthogonal transformations, so $-\boldsymbol{U}_k$ has the same law as $\boldsymbol{U}_k$, and thus $\mathbb{E}[\boldsymbol{U}_k] = 0$. Since $\kappa > 1$, part (a) gives $\mathbb{E}[A_k] < \infty$, and by the independence of A_k, $\boldsymbol{U}_k$ and $\boldsymbol{V}_k$, we obtain

$$\mathbb{E}[\boldsymbol{\Xi}_k] \quad = \quad \mathbb{E}[A_k]\,\mathbb{E}[\boldsymbol{U}_k]\,\mathbb{E}[\boldsymbol{V}_k]^\top \quad = \quad 0.$$

Now, since $\boldsymbol{X}_k$ is a function of $\boldsymbol{X}_0$ and $\boldsymbol{\Xi}_0, \dots, \boldsymbol{\Xi}_{k-1}$, the noise $\boldsymbol{\Xi}_k$ is independent of $\boldsymbol{X}_k$. Therefore, $\mathbb{E}[\boldsymbol{G}_k \mid \boldsymbol{X}_k] = \nabla f(\boldsymbol{X}_k)$, and for every $\alpha \in (1, \kappa)$, part (a) gives

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F^\alpha \mid \boldsymbol{X}_k] &= r^{\alpha/2}\mathbb{E}[A_k^\alpha] \\ &= \frac{r^{\alpha/2}\kappa a_0^\alpha}{\kappa - \alpha},\end{aligned}$$

which is Assumption 3.2 with the stated σ^α. Similarly, $\mathbb{E}\left[\|\boldsymbol{\Xi}_k\|_F^2\right] = r\mathbb{E}\left[A_k^2\right] = \infty$.

(c) For any orthogonal $\boldsymbol{Q}_1, \boldsymbol{Q}_2 \in \mathbb{R}^{d\times d}$, we have

$$\boldsymbol{Q}_1\boldsymbol{\Xi}_k\boldsymbol{Q}_2^\top \quad = \quad A_k\,(\boldsymbol{Q}_1\boldsymbol{U}_k)\,(\boldsymbol{Q}_2\boldsymbol{V}_k)^\top.$$

By the invariance of the Haar distribution, $\boldsymbol{Q}_1\boldsymbol{U}_k$ and $\boldsymbol{Q}_2\boldsymbol{V}_k$ have the same laws as $\boldsymbol{U}_k$ and $\boldsymbol{V}_k$, respectively. Since A_k, $\boldsymbol{U}_k$ and $\boldsymbol{V}_k$ are independent, the triple $(A_k, \boldsymbol{Q}_1\boldsymbol{U}_k, \boldsymbol{Q}_2\boldsymbol{V}_k)$ has the same law as $(A_k, \boldsymbol{U}_k, \boldsymbol{V}_k)$, which proves the claim. □

Properties of the Clipping Maps.

Lemma D.2 (Clipping maps are projections)**.** Let $\mathcal{K}_{\mathrm{sp}} := \{\boldsymbol{Y} : \sigma_1(\boldsymbol{Y}) \le \tau_{\mathrm{sp}}\}$ and $\mathcal{K}_{\mathrm{F}} := \{\boldsymbol{Y} : \|\boldsymbol{Y}\|_F \le \tau_{\mathrm{F}}\}$. Then $\mathcal{C}_{\tau_{\mathrm{sp}}} = P_{\mathcal{K}_{\mathrm{sp}}}$ and $\mathrm{clip}(\cdot, \tau_{\mathrm{F}}) = P_{\mathcal{K}_{\mathrm{F}}}$, where $P_{\mathcal{K}}$ is the projection onto $\mathcal{K}$ in the Frobenius norm (Lemma B.9). Consequently, for any $\Pi \in \{\mathcal{C}_{\tau_{\mathrm{sp}}}, \mathrm{clip}(\cdot, \tau_{\mathrm{F}})\}$ and all $\boldsymbol{Y}, \boldsymbol{Z} \in \mathbb{R}^{d\times d}$, the following statements hold.

(a) The clipped matrix is bounded:

$$\|\Pi(\boldsymbol{Y})\|_F \le B_\Pi, \qquad \text{where } B_\Pi := \begin{cases}\sqrt{d}\,\tau_{\mathrm{sp}}, & \Pi = \mathcal{C}_{\tau_{\mathrm{sp}}}, \\ \tau_{\mathrm{F}}, & \Pi = \mathrm{clip}(\cdot, \tau_{\mathrm{F}}).\end{cases}$$

(b) The map Π is nonexpansive and monotone:

$$\|\Pi(\boldsymbol{Y}) - \Pi(\boldsymbol{Z})\|_F \le \|\boldsymbol{Y} - \boldsymbol{Z}\|_F, \qquad \langle \boldsymbol{Y} - \boldsymbol{Z}, \Pi(\boldsymbol{Y}) - \Pi(\boldsymbol{Z})\rangle \ge 0.$$

(c) The map Π is orthogonally equivariant: for all orthogonal $\boldsymbol{Q}_1, \boldsymbol{Q}_2 \in \mathbb{R}^{d\times d}$, we have

$$\Pi(\boldsymbol{Q}_1\boldsymbol{Y}\boldsymbol{Q}_2^\top) = \boldsymbol{Q}_1\Pi(\boldsymbol{Y})\boldsymbol{Q}_2^\top.$$

In particular, $\Pi(-\boldsymbol{Y}) = -\Pi(\boldsymbol{Y})$.

Proof. Both $\mathcal{K}_{\mathrm{sp}}$ and $\mathcal{K}_{\mathrm{F}}$ are closed norm balls, and hence nonempty, closed and convex sets. The map $\mathrm{clip}(\cdot, \tau_{\mathrm{F}})$ leaves $\boldsymbol{Y}$ unchanged if $\|\boldsymbol{Y}\|_F \le \tau_{\mathrm{F}}$ and rescales it to $\frac{\tau_{\mathrm{F}}}{\|\boldsymbol{Y}\|_F}\boldsymbol{Y}$ otherwise, which is exactly the projection onto the Frobenius ball $\mathcal{K}_{\mathrm{F}}$.

Now, let us consider spectral clipping. Let $\boldsymbol{G} = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\sigma})\,\boldsymbol{V}^\top$ be a singular value decomposition of $\boldsymbol{G} \in \mathbb{R}^{d\times d}$. By definition, all singular values of $\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{G}) = \boldsymbol{U}\,\mathrm{diag}\left(\mathrm{clamp}_{\tau_{\mathrm{sp}}}(\boldsymbol{\sigma})\right)\boldsymbol{V}^\top$ are at most τ_{sp}, i.e., $\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{G}) \in \mathcal{K}_{\mathrm{sp}}$. Moreover, for any $\boldsymbol{Y} \in \mathcal{K}_{\mathrm{sp}}$, we have $0 \le \sigma_i(\boldsymbol{Y}) \le \tau_{\mathrm{sp}}$ for all i. Then, by Mirsky's inequality (Lemma B.7), we have

$$\begin{aligned}\|\boldsymbol{G} - \boldsymbol{Y}\|_F^2 &\ge \sum_{i=1}^d (\sigma_i(\boldsymbol{G}) - \sigma_i(\boldsymbol{Y}))^2 \\ &\ge \sum_{i=1}^d (\sigma_i(\boldsymbol{G}) - \min\{\sigma_i(\boldsymbol{G}), \tau_{\mathrm{sp}}\})^2 \\ &= \|\boldsymbol{G} - \mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{G})\|_F^2,\end{aligned}$$

where the second inequality holds because $\min\{\sigma_i(\boldsymbol{G}), \tau_{\mathrm{sp}}\}$ is the point of $[0, \tau_{\mathrm{sp}}]$ closest to $\sigma_i(\boldsymbol{G}) \ge 0$, and the last equality follows from $\boldsymbol{G} - \mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{G}) = \boldsymbol{U}\,\mathrm{diag}\left(\boldsymbol{\sigma} - \mathrm{clamp}_{\tau_{\mathrm{sp}}}(\boldsymbol{\sigma})\right)\boldsymbol{V}^\top$. Hence, $\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{G})$ is a point of $\mathcal{K}_{\mathrm{sp}}$ closest to $\boldsymbol{G}$, and since the projection onto a closed convex set is unique, $\mathcal{C}_{\tau_{\mathrm{sp}}} = P_{\mathcal{K}_{\mathrm{sp}}}$.

(a) For spectral clipping, we have

$$\begin{aligned}\|\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{Y})\|_F^2 &= \sum_{i=1}^{d} \min\{\sigma_i(\boldsymbol{Y}), \tau_{\mathrm{sp}}\}^2 \\ &\leq d\tau_{\mathrm{sp}}^2,\end{aligned}$$

so $\|\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{Y})\|_F \leq \sqrt{d}\,\tau_{\mathrm{sp}} = B_\Pi$. For norm clipping, $\|\mathrm{clip}(\boldsymbol{Y}, \tau_{\mathrm{F}})\|_F = \min\{\|\boldsymbol{Y}\|_F, \tau_{\mathrm{F}}\} \leq \tau_{\mathrm{F}} = B_\Pi$.

(b) Since Π is the projection onto a nonempty, closed and convex set, both inequalities follow from Lemma B.9.

(c) Let $\boldsymbol{Y} = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\sigma})\,\boldsymbol{V}^\top$ be a singular value decomposition of $\boldsymbol{Y}$. Since $\boldsymbol{Q}_1\boldsymbol{U}$ and $\boldsymbol{Q}_2\boldsymbol{V}$ are orthogonal, we have

$$\boldsymbol{Q}_1\boldsymbol{Y}\boldsymbol{Q}_2^\top = (\boldsymbol{Q}_1\boldsymbol{U})\,\mathrm{diag}(\boldsymbol{\sigma})\,(\boldsymbol{Q}_2\boldsymbol{V})^\top,$$

which is a singular value decomposition of $\boldsymbol{Q}_1\boldsymbol{Y}\boldsymbol{Q}_2^\top$ with the same singular values $\boldsymbol{\sigma}$. Therefore, we obtain

$$\begin{aligned}\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{Q}_1\boldsymbol{Y}\boldsymbol{Q}_2^\top) &= (\boldsymbol{Q}_1\boldsymbol{U})\,\mathrm{diag}\left(\mathrm{clamp}_{\tau_{\mathrm{sp}}}(\boldsymbol{\sigma})\right)(\boldsymbol{Q}_2\boldsymbol{V})^\top \\ &= \boldsymbol{Q}_1\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{Y})\boldsymbol{Q}_2^\top.\end{aligned}$$

Similarly, since $\|\boldsymbol{Q}_1\boldsymbol{Y}\boldsymbol{Q}_2^\top\|_F = \|\boldsymbol{Y}\|_F$, we have $\mathrm{clip}(\boldsymbol{Q}_1\boldsymbol{Y}\boldsymbol{Q}_2^\top, \tau_{\mathrm{F}}) = \boldsymbol{Q}_1\mathrm{clip}(\boldsymbol{Y}, \tau_{\mathrm{F}})\boldsymbol{Q}_2^\top$. Finally, taking $\boldsymbol{Q}_1 = -\boldsymbol{I}$ and $\boldsymbol{Q}_2 = \boldsymbol{I}$ gives $\Pi(-\boldsymbol{Y}) = -\Pi(\boldsymbol{Y})$. □

Lemma D.3 (Equivariant linear maps are scalar)**.** Let $\mathcal{T}$ be a linear map on $\mathbb{R}^{d\times d}$ such that

$$\mathcal{T}[\boldsymbol{S}\boldsymbol{W}\boldsymbol{S}'] = \boldsymbol{S}\,\mathcal{T}[\boldsymbol{W}]\,\boldsymbol{S}'$$

for all $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ and all signed permutation matrices $\boldsymbol{S}, \boldsymbol{S}' \in \mathbb{R}^{d\times d}$. Then

$$\mathcal{T} = \frac{\mathrm{tr}\,\mathcal{T}}{d^2}\,\mathrm{Id}\,.$$

Proof. This lemma is an instance of Schur's lemma. Indeed, let $\mathcal{B}_d$ be the group of signed permutation matrices in $\mathbb{R}^{d\times d}$. Then $(\boldsymbol{S}, \boldsymbol{S}') \mapsto \left(\boldsymbol{W} \mapsto \boldsymbol{S}\boldsymbol{W}\boldsymbol{S}'^\top\right)$ is an absolutely irreducible representation of $\mathcal{B}_d \times \mathcal{B}_d$ on $\mathbb{R}^{d\times d} \cong \mathbb{R}^d \otimes \mathbb{R}^d$, and by the hypothesis, $\mathcal{T}$ commutes with it. For completeness, we give an elementary proof.

For any $i, j \in \{1, \ldots, d\}$, let us write

$$\mathcal{T}[\boldsymbol{e}_i\boldsymbol{e}_j^\top] = \sum_{k,l=1}^{d} t_{kl}^{ij}\,\boldsymbol{e}_k\boldsymbol{e}_l^\top.$$

We first show that $\mathcal{T}$ maps each $\boldsymbol{e}_i\boldsymbol{e}_j^\top$ to a multiple of itself. Let $\boldsymbol{S} = \mathrm{diag}(\boldsymbol{s})$ and $\boldsymbol{S}' = \mathrm{diag}(\boldsymbol{s}')$ be diagonal sign matrices, i.e., $\boldsymbol{s}, \boldsymbol{s}' \in \{-1, 1\}^d$. Then $\boldsymbol{S}\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{S}' = s_i s_j'\,\boldsymbol{e}_i\boldsymbol{e}_j^\top$, and by the linearity of $\mathcal{T}$ and the hypothesis, we have

$$\begin{aligned}\sum_{k,l=1}^{d} s_i s_j'\,t_{kl}^{ij}\,\boldsymbol{e}_k\boldsymbol{e}_l^\top &= \mathcal{T}[\boldsymbol{S}\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{S}'] \\ &= \boldsymbol{S}\,\mathcal{T}[\boldsymbol{e}_i\boldsymbol{e}_j^\top]\,\boldsymbol{S}' \\ &= \sum_{k,l=1}^{d} s_k s_l'\,t_{kl}^{ij}\,\boldsymbol{e}_k\boldsymbol{e}_l^\top.\end{aligned}$$

Since the matrices $\boldsymbol{e}_k\boldsymbol{e}_l^\top$ form a basis of $\mathbb{R}^{d\times d}$, comparing the coefficients gives $s_i s_j'\,t_{kl}^{ij} = s_k s_l'\,t_{kl}^{ij}$ for all k, l. Now, let $(k, l) \neq (i, j)$. If $k \neq i$, we choose $\boldsymbol{s} = \boldsymbol{1} - 2\boldsymbol{e}_k$ and $\boldsymbol{s}' = \boldsymbol{1}$, so that $s_i s_j' = 1$ and $s_k s_l' = -1$. If $k = i$ and $l \neq j$, we choose $\boldsymbol{s} = \boldsymbol{1}$ and $\boldsymbol{s}' = \boldsymbol{1} - 2\boldsymbol{e}_l$, so that again $s_i s_j' = 1$ and $s_k s_l' = -1$. In both cases, $t_{kl}^{ij} = -t_{kl}^{ij}$, and hence $t_{kl}^{ij} = 0$. Therefore, we obtain

$$\mathcal{T}[\boldsymbol{e}_i\boldsymbol{e}_j^\top] = c_{ij}\,\boldsymbol{e}_i\boldsymbol{e}_j^\top, \qquad c_{ij} := t_{ij}^{ij}.$$

Next, we show that all c_{ij} are equal. Let $\boldsymbol{P}$ and $\boldsymbol{P}'$ be permutation matrices with $\boldsymbol{P}\boldsymbol{e}_i = \boldsymbol{e}_{\pi(i)}$ and $\boldsymbol{P}'\boldsymbol{e}_j = \boldsymbol{e}_{\pi'(j)}$. Then $\boldsymbol{P}\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{P}'^\top = \boldsymbol{e}_{\pi(i)}\boldsymbol{e}_{\pi'(j)}^\top$, and the hypothesis with $(\boldsymbol{S}, \boldsymbol{S}') = (\boldsymbol{P}, \boldsymbol{P}'^\top)$ gives

$$\begin{aligned}c_{\pi(i)\pi'(j)}\,\boldsymbol{e}_{\pi(i)}\boldsymbol{e}_{\pi'(j)}^\top &= \mathcal{T}[\boldsymbol{P}\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{P}'^\top] \\ &= \boldsymbol{P}\,\mathcal{T}[\boldsymbol{e}_i\boldsymbol{e}_j^\top]\,\boldsymbol{P}'^\top \\ &= c_{ij}\,\boldsymbol{e}_{\pi(i)}\boldsymbol{e}_{\pi'(j)}^\top.\end{aligned}$$

Hence, $c_{\pi(i)\pi'(j)} = c_{ij}$ for all permutations π, π' of $\{1, \dots, d\}$. Since any pair (i, j) can be mapped to any other pair (k, l) in this way, all c_{ij} are equal to a common value c, i.e., $\mathcal{T} = c\,\mathrm{Id}$. Finally, taking the trace gives $\operatorname{tr}\mathcal{T} = c\,d^2$, and hence $c = \operatorname{tr}\mathcal{T}/d^2$. □

Part (b) of the next lemma gives the neighborhood on which Lemma D.7 differentiates the mean map.

Lemma D.4 (Derivatives of the clipping maps at a noise sample)**.** Let us define the sets

$$\mathcal{U}_{\mathrm{sp}} := \{\boldsymbol{Y} : \sigma_i(\boldsymbol{Y}) \neq \tau_{\mathrm{sp}} \text{ for all } i\}, \qquad \mathcal{U}_{\mathrm{F}} := \{\boldsymbol{Y} : \|\boldsymbol{Y}\|_F \neq \tau_{\mathrm{F}}\},$$

which are both open. Then the following statements hold.

(a) The map $\mathcal{C}_{\tau_{\mathrm{sp}}}$ is continuously differentiable on $\mathcal{U}_{\mathrm{sp}}$, and the map $\mathrm{clip}(\cdot, \tau_{\mathrm{F}})$ is smooth on $\mathcal{U}_{\mathrm{F}}$. Moreover, wherever the derivative $D\Pi(\boldsymbol{Y})$ exists, we have $\|D\Pi(\boldsymbol{Y})\|_{\mathrm{op}} \leq 1$.

(b) Let $\boldsymbol{\Xi} = A\boldsymbol{U}\boldsymbol{V}^\top$, where $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{d\times r}$ have orthonormal columns and $A > \tau_{\mathrm{sp}}$. Then $\boldsymbol{\Xi} \in \mathcal{U}_{\mathrm{sp}}$ and

$$\operatorname{tr} D\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{\Xi}) = (d-r)^2 + \frac{2r(d-r)\tau_{\mathrm{sp}}}{A} + \frac{r(r-1)\tau_{\mathrm{sp}}}{2A}. \tag{32}$$

If, in addition, $A \geq a_0$ and $\|\boldsymbol{Z}\|_F < \rho_0 := \min\{\tau_{\mathrm{sp}}, a_0 - \tau_{\mathrm{sp}}\}$, then $\boldsymbol{\Xi} + \boldsymbol{Z} \in \mathcal{U}_{\mathrm{sp}}$.

(c) Let $\boldsymbol{\Xi}$ be as in (b), and let $R := \sqrt{r}A$, which equals $\|\boldsymbol{\Xi}\|_F$ by Lemma D.1(b). If $R \neq \tau_{\mathrm{F}}$, then $\boldsymbol{\Xi} \in \mathcal{U}_{\mathrm{F}}$ and

$$\operatorname{tr} D\,\mathrm{clip}(\boldsymbol{\Xi}, \tau_{\mathrm{F}}) = \begin{cases} d^2, & R < \tau_{\mathrm{F}}, \\ \frac{\tau_{\mathrm{F}}}{R}(d^2-1), & R > \tau_{\mathrm{F}}. \end{cases} \tag{33}$$

Proof. (a) Let $\boldsymbol{Y} \in \mathcal{U}_{\mathrm{sp}}$ have a singular value decomposition $\boldsymbol{Y} = \sum_{i=1}^d \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, and let $\boldsymbol{P}(\boldsymbol{Y}) = \sum_{\sigma_i > \tau_{\mathrm{sp}}} \boldsymbol{u}_i \boldsymbol{u}_i^\top$ be the spectral projector of $\boldsymbol{Y}\boldsymbol{Y}^\top$ associated with its eigenvalues larger than τ_{sp}^2. Since $\sigma_i \neq \tau_{\mathrm{sp}}$ for all i, these eigenvalues are separated from the remaining ones, and hence $\boldsymbol{P}$ is smooth on $\mathcal{U}_{\mathrm{sp}}$ (Kato, 1995). Now, let us define

$$\begin{aligned} \boldsymbol{Y}_> &:= \boldsymbol{P}(\boldsymbol{Y})\boldsymbol{Y} = \sum_{\sigma_i > \tau_{\mathrm{sp}}} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top, \\ \boldsymbol{N}(\boldsymbol{Y}) &:= \boldsymbol{Y}_>\boldsymbol{Y}_>^\top + \boldsymbol{I} - \boldsymbol{P}(\boldsymbol{Y}) = \sum_{\sigma_i > \tau_{\mathrm{sp}}} \sigma_i^2 \boldsymbol{u}_i \boldsymbol{u}_i^\top + \boldsymbol{I} - \boldsymbol{P}(\boldsymbol{Y}). \end{aligned}$$

Then $\boldsymbol{N}(\boldsymbol{Y}) \succ 0$ and $\boldsymbol{N}(\boldsymbol{Y})^{-1/2}\boldsymbol{Y}_> = \sum_{\sigma_i > \tau_{\mathrm{sp}}} \boldsymbol{u}_i \boldsymbol{v}_i^\top$. Since spectral clipping replaces every singular value $\sigma_i > \tau_{\mathrm{sp}}$ by τ_{sp} and leaves the others unchanged, we obtain

$$\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{Y}) = \boldsymbol{Y} - \boldsymbol{Y}_> + \tau_{\mathrm{sp}}\,\boldsymbol{N}(\boldsymbol{Y})^{-1/2}\boldsymbol{Y}_>, \qquad \boldsymbol{Y} \in \mathcal{U}_{\mathrm{sp}}. \tag{34}$$

The right-hand side is a composition of smooth maps, and hence $\mathcal{C}_{\tau_{\mathrm{sp}}}$ is continuously differentiable on $\mathcal{U}_{\mathrm{sp}}$. On $\mathcal{U}_{\mathrm{F}}$, the map $\mathrm{clip}(\cdot, \tau_{\mathrm{F}})$ locally coincides either with $\boldsymbol{Y} \mapsto \boldsymbol{Y}$ or with $\boldsymbol{Y} \mapsto \tau_{\mathrm{F}}\boldsymbol{Y}/\|\boldsymbol{Y}\|_F$, and both maps are smooth. Finally, by Lemma D.2(b), Π is 1-Lipschitz, so $\|D\Pi(\boldsymbol{Y})\|_{\mathrm{op}} \leq 1$ wherever $D\Pi(\boldsymbol{Y})$ exists.

(b) The singular values of $\boldsymbol{\Xi}$ are A, with multiplicity r, and 0, with multiplicity $d - r$. Since $0 < \tau_{\mathrm{sp}} < A$, none of them equals τ_{sp}, and hence $\boldsymbol{\Xi} \in \mathcal{U}_{\mathrm{sp}}$. Now, let $A \geq a_0$ and $\|\boldsymbol{Z}\|_F < \rho_0$. Then, by Weyl's inequality (Lemma B.8), we have $|\sigma_i(\boldsymbol{\Xi} + \boldsymbol{Z}) - \sigma_i(\boldsymbol{\Xi})| \leq \|\boldsymbol{Z}\|_F < \rho_0$ for all i. Since $\rho_0 \leq a_0 - \tau_{\mathrm{sp}}$ and $\rho_0 \leq \tau_{\mathrm{sp}}$, this gives

$$\begin{aligned} \sigma_i(\boldsymbol{\Xi} + \boldsymbol{Z}) &> A - \rho_0 \geq a_0 - \rho_0 \geq \tau_{\mathrm{sp}} \quad \text{for } i \leq r, \\ \sigma_i(\boldsymbol{\Xi} + \boldsymbol{Z}) &< \rho_0 \leq \tau_{\mathrm{sp}} \quad \text{for } i > r, \end{aligned}$$

and hence $\boldsymbol{\Xi} + \boldsymbol{Z} \in \mathcal{U}_{\mathrm{sp}}$.

It remains to compute the trace. Let us complete $\boldsymbol{U}$ and $\boldsymbol{V}$ to orthogonal matrices $\bar{\boldsymbol{U}} = [\boldsymbol{U}, \boldsymbol{U}_\perp]$ and $\bar{\boldsymbol{V}} = [\boldsymbol{V}, \boldsymbol{V}_\perp]$, and let $\boldsymbol{\Sigma} := A\sum_{i=1}^r \boldsymbol{e}_i\boldsymbol{e}_i^\top$, so that $\boldsymbol{\Xi} = \bar{\boldsymbol{U}}\boldsymbol{\Sigma}\bar{\boldsymbol{V}}^\top$. By Lemma D.2(c), for any $\boldsymbol{W} \in \mathbb{R}^{d\times d}$, we have

$$\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{\Xi} + \boldsymbol{W}) = \bar{\boldsymbol{U}}\,\mathcal{C}_{\tau_{\mathrm{sp}}}\left(\boldsymbol{\Sigma} + \bar{\boldsymbol{U}}^\top\boldsymbol{W}\bar{\boldsymbol{V}}\right)\bar{\boldsymbol{V}}^\top.$$

Then, differentiating at $\boldsymbol{W} = 0$, we get $D\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{\Xi})[\boldsymbol{W}] = \bar{\boldsymbol{U}}\,D\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{\Sigma})[\bar{\boldsymbol{U}}^\top\boldsymbol{W}\bar{\boldsymbol{V}}]\,\bar{\boldsymbol{V}}^\top$. Since $\boldsymbol{W} \mapsto \bar{\boldsymbol{U}}^\top\boldsymbol{W}\bar{\boldsymbol{V}}$ is an isometry of $\mathbb{R}^{d\times d}$, the linear maps $D\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{\Xi})$ and $D\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{\Sigma})$ are similar, and hence

$$\begin{aligned} \operatorname{tr} D\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{\Xi}) &= \operatorname{tr} D\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{\Sigma}) \\ &= \sum_{i,j=1}^d \frac{\mathrm{d}}{\mathrm{d}s}\left[\mathcal{C}_{\tau_{\mathrm{sp}}}\left(\boldsymbol{\Sigma} + s\,\boldsymbol{e}_i\boldsymbol{e}_j^\top\right)\right]_{ij}\Big|_{s=0}. \end{aligned}$$

To compute each term, note that after a permutation of rows and columns, the matrix $\boldsymbol{\Sigma}+s\,\boldsymbol{e}_i\boldsymbol{e}_j^\top$ is block diagonal with rectangular blocks. Permutations are orthogonal, and a singular value decomposition of a block diagonal matrix is assembled from those of its blocks, so by Lemma D.2(c), $\mathcal{C}_{\tau_{\mathrm{sp}}}$ acts on each block separately. Therefore, the (i,j) entry of $\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{\Sigma}+s\,\boldsymbol{e}_i\boldsymbol{e}_j^\top)$ depends only on the block that contains it. Let $|s|<\min\{\tau_{\mathrm{sp}}, A-\tau_{\mathrm{sp}}\}$. There are four kinds of pairs (i,j).

If $i,j>r$, which happens for $(d-r)^2$ pairs, the block is the 1×1 matrix (s). Its singular value $|s|$ is smaller than τ_{sp}, so it is left unchanged, and the derivative of the (i,j) entry at $s=0$ equals 1.

If $i\le r<j$ or $j\le r<i$, which happens for $2r(d-r)$ pairs, the block is $(A\ \ s)$ or its transpose. It has rank one with the singular value $\sqrt{A^2+s^2}>\tau_{\mathrm{sp}}$, so spectral clipping replaces it by τ_{sp} times its normalization, $\frac{\tau_{\mathrm{sp}}}{\sqrt{A^2+s^2}}(A\ \ s)$. The (i,j) entry is $\frac{\tau_{\mathrm{sp}}s}{\sqrt{A^2+s^2}}$, whose derivative at $s=0$ equals $\frac{\tau_{\mathrm{sp}}}{A}$.

If $i=j\le r$, which happens for r pairs, the block is $(A+s)$ with the singular value $A+s>\tau_{\mathrm{sp}}$, so it is replaced by (τ_{sp}), and the derivative equals 0.

If $i\ne j$ and $i,j\le r$, which happens for $r(r-1)$ pairs, the block is

$$\boldsymbol{B}=\begin{pmatrix}A & s\\ 0 & A\end{pmatrix},$$

both singular values of which exceed τ_{sp} by Weyl's inequality (Lemma B.8), so $\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{B})=\tau_{\mathrm{sp}}\boldsymbol{Q}$, where $\boldsymbol{Q}$ is the orthogonal polar factor of $\boldsymbol{B}$. The matrix

$$\boldsymbol{Q}=\frac{1}{\sqrt{4A^2+s^2}}\begin{pmatrix}2A & s\\ -s & 2A\end{pmatrix}$$

is this polar factor, since it is orthogonal and

$$\boldsymbol{Q}^\top\boldsymbol{B} \;=\; \frac{1}{\sqrt{4A^2+s^2}}\begin{pmatrix}2A^2 & As\\ As & 2A^2+s^2\end{pmatrix}$$

is symmetric positive definite. The (i,j) entry of $\tau_{\mathrm{sp}}\boldsymbol{Q}$ is $\frac{\tau_{\mathrm{sp}}s}{\sqrt{4A^2+s^2}}$, whose derivative at $s=0$ equals $\frac{\tau_{\mathrm{sp}}}{2A}$.

Finally, summing the number of pairs times the derivative over the four kinds, we get

$$\operatorname{tr} D\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{\Xi}) \;=\; (d-r)^2\cdot 1+2r(d-r)\cdot\frac{\tau_{\mathrm{sp}}}{A}+r\cdot 0+r(r-1)\cdot\frac{\tau_{\mathrm{sp}}}{2A},$$

which is (32).

(c) Since $\|\boldsymbol{\Xi}\|_F=R\ne\tau_{\mathrm{F}}$, we have $\boldsymbol{\Xi}\in\mathcal{U}_{\mathrm{F}}$. If $R<\tau_{\mathrm{F}}$, then $\operatorname{clip}(\cdot,\tau_{\mathrm{F}})$ is the identity in a neighborhood of $\boldsymbol{\Xi}$, so its derivative is the identity map on $\mathbb{R}^{d\times d}$, whose trace is d^2. If $R>\tau_{\mathrm{F}}$, then $\operatorname{clip}(\boldsymbol{Y},\tau_{\mathrm{F}})=\tau_{\mathrm{F}}\boldsymbol{Y}/\|\boldsymbol{Y}\|_F$ in a neighborhood of $\boldsymbol{\Xi}$. Hence, with $\hat{\boldsymbol{\Xi}}:=\boldsymbol{\Xi}/R$, we have

$$D\operatorname{clip}(\boldsymbol{\Xi},\tau_{\mathrm{F}})[\boldsymbol{W}] \;=\; \frac{\tau_{\mathrm{F}}}{R}\left(\boldsymbol{W}-\langle\hat{\boldsymbol{\Xi}},\boldsymbol{W}\rangle\hat{\boldsymbol{\Xi}}\right),$$

which is $\frac{\tau_{\mathrm{F}}}{R}$ times the orthogonal projection onto the orthogonal complement of $\hat{\boldsymbol{\Xi}}$ in $\mathbb{R}^{d\times d}$. Therefore, its trace equals $\frac{\tau_{\mathrm{F}}}{R}(d^2-1)$. □

The Mean Clipped Update. Now, let us define the mean clipped update and the energy of the clipped noise at the solution,

$$\boldsymbol{M}(\boldsymbol{Z}):=\mathbb{E}_{\boldsymbol{\Xi}}\big[\Pi(\boldsymbol{Z}+\boldsymbol{\Xi})\big],\qquad q_0:=\mathbb{E}\left[\|\Pi(\boldsymbol{\Xi})\|_F^2\right]. \tag{35}$$

Lemma D.5 (Global properties of the mean map)**.** Let Assumption 3.5 hold.

(a) $\|\boldsymbol{M}(\boldsymbol{Z})\|_F\le B_\Pi$, $\|\boldsymbol{M}(\boldsymbol{Y})-\boldsymbol{M}(\boldsymbol{Z})\|_F\le\|\boldsymbol{Y}-\boldsymbol{Z}\|_F$ and $\langle\boldsymbol{Y}-\boldsymbol{Z},\boldsymbol{M}(\boldsymbol{Y})-\boldsymbol{M}(\boldsymbol{Z})\rangle\ge 0$ for all $\boldsymbol{Y},\boldsymbol{Z}$.

(b) $\boldsymbol{M}(\boldsymbol{Q}_1\boldsymbol{Z}\boldsymbol{Q}_2^\top)=\boldsymbol{Q}_1\boldsymbol{M}(\boldsymbol{Z})\boldsymbol{Q}_2^\top$ for all orthogonal $\boldsymbol{Q}_1,\boldsymbol{Q}_2$; in particular $\boldsymbol{M}(-\boldsymbol{Z})=-\boldsymbol{M}(\boldsymbol{Z})$ and $\boldsymbol{M}(0)=0$.

(c) For spectral clipping, $\Pi(\boldsymbol{\Xi})=\tau_{\mathrm{sp}}\boldsymbol{U}\boldsymbol{V}^\top$ and $q_0=r\tau_{\mathrm{sp}}^2$; for norm clipping, $\Pi(\boldsymbol{\Xi})=\frac{\min\{R,\tau_{\mathrm{F}}\}}{\sqrt{r}}\boldsymbol{U}\boldsymbol{V}^\top$ and $q_0=\mathbb{E}\left[\min\{R,\tau_{\mathrm{F}}\}^2\right]$. In both cases $q_0>0$ and

$$\mathbb{E}\left[\|\Pi(\boldsymbol{\Xi})_{j,:}\|^2\right]=\frac{q_0}{d},\qquad j=1,\dots,d. \tag{36}$$

Proof. (a) By Jensen's inequality and Lemma D.2(a), $\|\boldsymbol{M}(\boldsymbol{Z})\|_F \le \mathbb{E}_{\boldsymbol{\Xi}}\big[\|\Pi(\boldsymbol{Z}+\boldsymbol{\Xi})\|_F\big] \le B_\Pi$. Similarly, applying Lemma D.2(b) to $\boldsymbol{Y}+\boldsymbol{\Xi}$ and $\boldsymbol{Z}+\boldsymbol{\Xi}$ and taking expectations,

$$\begin{aligned} \|\boldsymbol{M}(\boldsymbol{Y})-\boldsymbol{M}(\boldsymbol{Z})\|_F &\le \mathbb{E}_{\boldsymbol{\Xi}}\big[\|\Pi(\boldsymbol{Y}+\boldsymbol{\Xi})-\Pi(\boldsymbol{Z}+\boldsymbol{\Xi})\|_F\big] \le \|\boldsymbol{Y}-\boldsymbol{Z}\|_F, \\ \langle \boldsymbol{Y}-\boldsymbol{Z}, \boldsymbol{M}(\boldsymbol{Y})-\boldsymbol{M}(\boldsymbol{Z})\rangle &= \mathbb{E}_{\boldsymbol{\Xi}}\big[\langle (\boldsymbol{Y}+\boldsymbol{\Xi})-(\boldsymbol{Z}+\boldsymbol{\Xi}), \Pi(\boldsymbol{Y}+\boldsymbol{\Xi})-\Pi(\boldsymbol{Z}+\boldsymbol{\Xi})\rangle\big] \ge 0. \end{aligned}$$

(b) By Lemma D.2(c) and Lemma D.1(c),

$$\boldsymbol{M}(\boldsymbol{Q}_1\boldsymbol{Z}\boldsymbol{Q}_2^\top) = \mathbb{E}_{\boldsymbol{\Xi}}\big[\Pi(\boldsymbol{Q}_1(\boldsymbol{Z}+\boldsymbol{Q}_1^\top\boldsymbol{\Xi}\boldsymbol{Q}_2)\boldsymbol{Q}_2^\top)\big] = \boldsymbol{Q}_1\,\mathbb{E}_{\boldsymbol{\Xi}}\big[\Pi(\boldsymbol{Z}+\boldsymbol{Q}_1^\top\boldsymbol{\Xi}\boldsymbol{Q}_2)\big]\,\boldsymbol{Q}_2^\top = \boldsymbol{Q}_1\boldsymbol{M}(\boldsymbol{Z})\boldsymbol{Q}_2^\top;$$

take $(\boldsymbol{Q}_1, \boldsymbol{Q}_2) = (-\boldsymbol{I}, \boldsymbol{I})$ for the last two claims.

(c) The nonzero singular values of $\boldsymbol{\Xi}$ all equal $A \ge a_0 > \tau_{\mathrm{sp}}$, and $\|\boldsymbol{\Xi}\|_F = R$, so

$$\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{\Xi}) = \tau_{\mathrm{sp}}\boldsymbol{U}\boldsymbol{V}^\top, \qquad \mathrm{clip}(\boldsymbol{\Xi}, \tau_{\mathrm{F}}) = \frac{\min\{R, \tau_{\mathrm{F}}\}}{R}\boldsymbol{\Xi};$$

both values of q_0 follow from $\|\boldsymbol{U}\boldsymbol{V}^\top\|_F^2 = r$. For a permutation matrix $\boldsymbol{P}$, $\boldsymbol{P}\boldsymbol{\Xi}$ has the law of $\boldsymbol{\Xi}$ (Lemma D.1(c)) and $\Pi(\boldsymbol{P}\boldsymbol{\Xi}) = \boldsymbol{P}\Pi(\boldsymbol{\Xi})$, so all rows of $\Pi(\boldsymbol{\Xi})$ have the same expected squared norm; these d numbers sum to q_0. □

Lemma D.6 (Expected derivative of the clipping map)**.** Let Assumption 3.5 hold. Then $D\Pi(\boldsymbol{\Xi})$ exists almost surely and

$$\mathbb{E}\left[D\Pi(\boldsymbol{\Xi})\right] = \lambda\,\mathrm{Id}, \qquad \lambda := \begin{cases} \lambda_{\mathrm{sp}}, & \Pi = \mathcal{C}_{\tau_{\mathrm{sp}}}, \\ \lambda_{\mathrm{F}}, & \Pi = \mathrm{clip}(\cdot, \tau_{\mathrm{F}}), \end{cases} \tag{37}$$

where λ_{sp} is the constant of Theorem 3.6 and

$$\lambda_{\mathrm{F}} := \Pr[R < \tau_{\mathrm{F}}] + \left(1 - d^{-2}\right)\tau_{\mathrm{F}}\,\mathbb{E}\left[R^{-1}\mathbf{1}\{R > \tau_{\mathrm{F}}\}\right]. \tag{38}$$

Moreover, $\lambda_{\mathrm{sp}} \ge (1 - r/d)^2 > 0$ and $\lambda_{\mathrm{F}} > 0$.

Proof. Almost sure existence follows from Lemma D.4(b),(c), since $A \ge a_0 > \tau_{\mathrm{sp}}$ and $\Pr[R = \tau_{\mathrm{F}}] = 0$ (A has a density). By Lemma D.4(a), $\mathcal{T} := \mathbb{E}\left[D\Pi(\boldsymbol{\Xi})\right]$ is well defined.

Differentiating Lemma D.2(c) at $\boldsymbol{Y}$ in the direction $\boldsymbol{W}$ gives

$$D\Pi(\boldsymbol{Q}_1\boldsymbol{Y}\boldsymbol{Q}_2^\top)[\boldsymbol{Q}_1\boldsymbol{W}\boldsymbol{Q}_2^\top] = \boldsymbol{Q}_1 D\Pi(\boldsymbol{Y})[\boldsymbol{W}]\boldsymbol{Q}_2^\top;$$

since $\boldsymbol{Q}_1\boldsymbol{\Xi}\boldsymbol{Q}_2^\top$ has the law of $\boldsymbol{\Xi}$ (Lemma D.1(c)),

$$\mathcal{T}[\boldsymbol{Q}_1\boldsymbol{W}\boldsymbol{Q}_2^\top] = \mathbb{E}\left[D\Pi(\boldsymbol{Q}_1\boldsymbol{\Xi}\boldsymbol{Q}_2^\top)[\boldsymbol{Q}_1\boldsymbol{W}\boldsymbol{Q}_2^\top]\right] = \boldsymbol{Q}_1\mathcal{T}[\boldsymbol{W}]\boldsymbol{Q}_2^\top$$

for all orthogonal $\boldsymbol{Q}_1, \boldsymbol{Q}_2$, in particular for signed permutation matrices. Lemma D.3 gives

$$\mathcal{T} = d^{-2}\mathbb{E}\left[\mathrm{tr}\, D\Pi(\boldsymbol{\Xi})\right]\mathrm{Id}\,.$$

For spectral clipping, (32), Lemma D.1(a), $2r(d-r) + \frac{r(r-1)}{2} = \frac{r(4d-3r-1)}{2}$ and the definition of λ_{sp} in Theorem 3.6 give

$$\frac{1}{d^2}\mathbb{E}\left[\mathrm{tr}\, D\mathcal{C}_{\tau_{\mathrm{sp}}}(\boldsymbol{\Xi})\right] = \left(1-\frac{r}{d}\right)^2 + \frac{r(4d-3r-1)}{2d^2}\,\tau_{\mathrm{sp}}\,\mathbb{E}\left[A^{-1}\right] = \left(1-\frac{r}{d}\right)^2 + \frac{r(4d-3r-1)\,\kappa\,\tau_{\mathrm{sp}}}{2(\kappa+1)\,a_0 d^2} = \lambda_{\mathrm{sp}}.$$

For norm clipping, (33) and $\Pr[R = \tau_{\mathrm{F}}] = 0$ give

$$d^{-2}\mathbb{E}\left[\mathrm{tr}\, D\,\mathrm{clip}(\boldsymbol{\Xi}, \tau_{\mathrm{F}})\right] = \Pr[R < \tau_{\mathrm{F}}] + (1 - d^{-2})\,\tau_{\mathrm{F}}\,\mathbb{E}\left[R^{-1}\mathbf{1}\{R > \tau_{\mathrm{F}}\}\right] = \lambda_{\mathrm{F}}.$$

Finally, $4d-3r-1 \ge d+2 > 0$ and $r < d$ give $\lambda_{\mathrm{sp}} \ge (1-r/d)^2 > 0$, and $\lambda_{\mathrm{F}} > 0$ because $\Pr[R < \tau_{\mathrm{F}}] > 0$ if $\tau_{\mathrm{F}} > \sqrt{r}\,a_0$, while $\Pr[R > \tau_{\mathrm{F}}] = 1$ if $\tau_{\mathrm{F}} \le \sqrt{r}\,a_0$. □

Lemma D.7 (Linearization of the mean map)**.** Let Assumption 3.5 hold, let λ be as in (37), and let ρ_0 be as in Lemma D.4(b) for spectral clipping and $\rho_0 := \infty$ for norm clipping.

(a) $\boldsymbol{M}$ is differentiable on $\{\boldsymbol{Z} : \|\boldsymbol{Z}\|_F < \rho_0\}$ with $D\boldsymbol{M}(\boldsymbol{Z}) = \mathbb{E}\,[D\Pi(\boldsymbol{Z} + \boldsymbol{\Xi})]$, and $\|D\boldsymbol{M}(\boldsymbol{Z}) - \lambda\,\mathrm{Id}\,\|_{\mathrm{op}} \to 0$ as $\boldsymbol{Z} \to 0$. In particular, $\boldsymbol{M}(\boldsymbol{Z}) = \lambda\boldsymbol{Z} + o(\|\boldsymbol{Z}\|_F)$ as $\boldsymbol{Z} \to 0$.

(b) For every $\epsilon > 0$ there is $\rho_\epsilon \in (0, \rho_0]$ such that, whenever $\|\boldsymbol{Z}\|_F < \rho_\epsilon$,

$$\|\boldsymbol{M}(\boldsymbol{Z})_{j,:} - \lambda\boldsymbol{Z}_{j,:}\| \le \epsilon\,\|\boldsymbol{Z}_{j,:}\|, \qquad j = 1, \dots, d. \tag{39}$$

(c) For every $\rho > 0$,

$$\inf_{\|\boldsymbol{Z}\|_F \ge \rho} \frac{\langle \boldsymbol{Z}, \boldsymbol{M}(\boldsymbol{Z})\rangle}{\|\boldsymbol{Z}\|_F} > 0. \tag{40}$$

Proof. (a) Fix $\boldsymbol{Z}$ with $\|\boldsymbol{Z}\|_F < \rho_0$. Almost surely, Π is continuously differentiable on a neighborhood of $\boldsymbol{Z} + \boldsymbol{\Xi}$: for spectral clipping, $\boldsymbol{Z} + \boldsymbol{\Xi} \in \mathcal{U}_{\mathrm{sp}}$ by Lemma D.4(b); for norm clipping, conditionally on $(\boldsymbol{U}, \boldsymbol{V})$ the equation $\|\boldsymbol{Z} + \boldsymbol{\Xi}\|_F^2 = rA^2 + 2A\langle \boldsymbol{U}\boldsymbol{V}^\top, \boldsymbol{Z}\rangle + \|\boldsymbol{Z}\|_F^2 = \tau_{\mathrm{F}}^2$ has at most two roots in A, so $\Pr[\|\boldsymbol{Z} + \boldsymbol{\Xi}\|_F = \tau_{\mathrm{F}}] = 0$. For $\boldsymbol{W} \neq 0$, the normalized remainder

$$\frac{\|\Pi(\boldsymbol{Z} + \boldsymbol{W} + \boldsymbol{\Xi}) - \Pi(\boldsymbol{Z} + \boldsymbol{\Xi}) - D\Pi(\boldsymbol{Z} + \boldsymbol{\Xi})[\boldsymbol{W}]\|_F}{\|\boldsymbol{W}\|_F}$$

tends to 0 almost surely as $\boldsymbol{W} \to 0$ and is bounded by 2 by Lemmas D.2(b) and D.4(a); dominated convergence gives differentiability of $\boldsymbol{M}$ at $\boldsymbol{Z}$ with $D\boldsymbol{M}(\boldsymbol{Z}) = \mathbb{E}\,[D\Pi(\boldsymbol{Z} + \boldsymbol{\Xi})]$. If $\boldsymbol{Z}_i \to 0$, then almost surely $D\Pi(\boldsymbol{Z}_i + \boldsymbol{\Xi}) \to D\Pi(\boldsymbol{\Xi})$ by continuity of $D\Pi$ near $\boldsymbol{\Xi}$, and dominated convergence together with Lemma D.6 yields $\|D\boldsymbol{M}(\boldsymbol{Z}_i) - \lambda\,\mathrm{Id}\,\|_{\mathrm{op}} \to 0$. The last claim follows from $\boldsymbol{M}(0) = 0$ (Lemma D.5(b)) and the mean value inequality.

(b) By (a), choose $\rho_\epsilon \in (0, \rho_0]$ with $\|D\boldsymbol{M}(\boldsymbol{W}) - \lambda\,\mathrm{Id}\,\|_{\mathrm{op}} \le \epsilon$ for $\|\boldsymbol{W}\|_F < \rho_\epsilon$. Fix $\boldsymbol{Z}$ with $\|\boldsymbol{Z}\|_F < \rho_\epsilon$ and an index j, and let $\boldsymbol{Z}^{(j)}$ be $\boldsymbol{Z}$ with its jth row replaced by zero. The segment from $\boldsymbol{Z}^{(j)}$ to $\boldsymbol{Z}$ lies in the ball of radius ρ_ϵ, so the mean value inequality gives

$$\big\|\boldsymbol{M}(\boldsymbol{Z}) - \boldsymbol{M}(\boldsymbol{Z}^{(j)}) - \lambda(\boldsymbol{Z} - \boldsymbol{Z}^{(j)})\big\|_F \le \epsilon\,\|\boldsymbol{Z} - \boldsymbol{Z}^{(j)}\|_F = \epsilon\,\|\boldsymbol{Z}_{j,:}\|.$$

For the reflection $\boldsymbol{S}_j := \boldsymbol{I} - 2\boldsymbol{e}_j\boldsymbol{e}_j^\top$ we have $\boldsymbol{S}_j\boldsymbol{Z}^{(j)} = \boldsymbol{Z}^{(j)}$, hence $\boldsymbol{M}(\boldsymbol{Z}^{(j)}) = \boldsymbol{S}_j\boldsymbol{M}(\boldsymbol{Z}^{(j)})$ by Lemma D.5(b), and the jth row of $\boldsymbol{M}(\boldsymbol{Z}^{(j)})$ vanishes. Taking the jth row in the last display proves (39).

(c) Fix $\rho > 0$ and, by (a), choose $\varepsilon_0 \in (0, \rho)$ with $\langle \boldsymbol{W}, \boldsymbol{M}(\varepsilon_0\boldsymbol{W})\rangle \ge \lambda\varepsilon_0/2$ for all $\boldsymbol{W}$ with $\|\boldsymbol{W}\|_F = 1$. For $\|\boldsymbol{Z}\|_F \ge \rho$ let $\boldsymbol{W} := \boldsymbol{Z}/\|\boldsymbol{Z}\|_F$. Monotonicity of $\boldsymbol{M}$ (Lemma D.5(a)) applied to $\boldsymbol{Z}$ and $\varepsilon_0\boldsymbol{W}$ gives $(\|\boldsymbol{Z}\|_F - \varepsilon_0)\langle \boldsymbol{W}, \boldsymbol{M}(\boldsymbol{Z}) - \boldsymbol{M}(\varepsilon_0\boldsymbol{W})\rangle \ge 0$, and since $\|\boldsymbol{Z}\|_F > \varepsilon_0$,

$$\frac{\langle \boldsymbol{Z}, \boldsymbol{M}(\boldsymbol{Z})\rangle}{\|\boldsymbol{Z}\|_F} = \langle \boldsymbol{W}, \boldsymbol{M}(\boldsymbol{Z})\rangle \ge \langle \boldsymbol{W}, \boldsymbol{M}(\varepsilon_0\boldsymbol{W})\rangle \ge \frac{\lambda\varepsilon_0}{2} > 0.$$

□

Localization of the Iterates under Harmonic Step Sizes.

Lemma D.8 (Localization of the iterates)**.** Let $\boldsymbol{Y}_0 \in \mathbb{R}^{d\times d}$ be deterministic, let $\boldsymbol{\Xi}_0, \boldsymbol{\Xi}_1, \dots$ be independent and identically distributed random matrices, and let

$$\boldsymbol{Y}_{k+1} = \boldsymbol{Y}_k - \eta_k\boldsymbol{\Phi}(\boldsymbol{Y}_k, \boldsymbol{\Xi}_k), \qquad \eta_k = \frac{c}{k + k_0}, \qquad c, k_0 > 0, \tag{41}$$

where $\boldsymbol{\Phi}$ is measurable with $\|\boldsymbol{\Phi}(\boldsymbol{Y}, \boldsymbol{\Xi})\|_F \le \bar{B}$ for all arguments, and the mean field $\boldsymbol{m}(\boldsymbol{Y}) := \mathbb{E}_{\boldsymbol{\Xi}}[\boldsymbol{\Phi}(\boldsymbol{Y}, \boldsymbol{\Xi})]$ satisfies

$$\inf_{\|\boldsymbol{Y}\|_F \ge \rho} \frac{\langle \boldsymbol{Y}, \boldsymbol{m}(\boldsymbol{Y})\rangle}{\|\boldsymbol{Y}\|_F} > 0 \qquad \text{for every } \rho > 0. \tag{42}$$

Then, for every $\rho > 0$, there is $\delta_\rho > 0$ such that $\Pr[\|\boldsymbol{Y}_k\|_F \ge \rho] \le \exp\{-k^{\delta_\rho}\}$ for all sufficiently large k. Consequently, for all $\rho, p > 0$,

$$\mathbb{E}\left[(1 + \|\boldsymbol{Y}_k\|_F^2)\,\mathbf{1}\{\|\boldsymbol{Y}_k\|_F \ge \rho\}\right] = o(k^{-p}). \tag{43}$$

Proof. Let $\mathcal{F}_i := \sigma(\boldsymbol{\Xi}_0, \dots, \boldsymbol{\Xi}_{i-1})$; then $\boldsymbol{Y}_i$ is $\mathcal{F}_i$-measurable and $\mathbb{E}\,[\boldsymbol{\Phi}(\boldsymbol{Y}_i, \boldsymbol{\Xi}_i) \mid \mathcal{F}_i] = \boldsymbol{m}(\boldsymbol{Y}_i)$. Since $\|\boldsymbol{Y}_{i+1} - \boldsymbol{Y}_i\|_F \le \bar{B}\eta_i$, there is a constant C_0 with

$$\|\boldsymbol{Y}_i\|_F \le C_0 + c\bar{B}\log(i + 2), \qquad i \ge 0. \tag{44}$$

Decompose $\|\boldsymbol{Y}_{i+1}\|_F - \|\boldsymbol{Y}_i\|_F = \mu_i + D_{i+1}$ with $\mu_i := \mathbb{E}\left[\|\boldsymbol{Y}_{i+1}\|_F - \|\boldsymbol{Y}_i\|_F \mid \mathcal{F}_i\right]$, so that $\mathbb{E}\left[D_{i+1} \mid \mathcal{F}_i\right] = 0$ and $|D_{i+1}| \le 2\bar{B}\eta_i$. All constants C_1, C_2, C_3 below depend on ρ but not on k.

Fix $\rho > 0$ and let $\nu > 0$ satisfy $\langle \boldsymbol{Y}, \boldsymbol{m}(\boldsymbol{Y})\rangle \ge 2\nu\|\boldsymbol{Y}\|_F$ whenever $\|\boldsymbol{Y}\|_F \ge \rho/4$, by (42). Let i_ρ be such that $\bar{B}\eta_i \le \rho/8$ and $4\bar{B}^2\eta_i/\rho \le \nu$ for all $i \ge i_\rho$. For $i \ge i_\rho$, on $\{\|\boldsymbol{Y}_i\|_F \ge \rho/4\}$ the segment from $\boldsymbol{Y}_i$ to $\boldsymbol{Y}_{i+1}$ stays at distance at least $\rho/8$ from the origin, where the Hessian of $\|\cdot\|_F$ has operator norm at most $8/\rho$; Taylor's theorem and conditional expectation give

$$\mu_i \le -\eta_i \frac{\langle \boldsymbol{Y}_i, \boldsymbol{m}(\boldsymbol{Y}_i)\rangle}{\|\boldsymbol{Y}_i\|_F} + \frac{4\bar{B}^2\eta_i^2}{\rho} \le -\nu\eta_i \qquad \text{on } \{\|\boldsymbol{Y}_i\|_F \ge \rho/4\}, \quad i \ge i_\rho. \tag{45}$$

Set $\delta := \nu/(2(\bar{B}+\nu)) \in (0, 1/2)$, $k_\delta := \lfloor k^\delta \rfloor$, and let k be so large that $k_\delta \ge \max\{i_\rho, 2\}$. The process $S_t := \sum_{i=k_\delta}^{t-1} D_{i+1}$, $t \ge k_\delta$, is a martingale with increments bounded by $2\bar{B}\eta_i$ and $\sum_{i\ge k_\delta}(2\bar{B}\eta_i)^2 \le C_1 k^{-\delta}$. The event $\{\|\boldsymbol{Y}_k\|_F \ge \rho\}$ is contained in $E_1 \cup E_2$, where:

E_1: $\|\boldsymbol{Y}_i\|_F \ge \rho/4$ for all $k_\delta \le i \le \lfloor k/2 \rfloor$. On E_1, (45) gives

$$0 \le \|\boldsymbol{Y}_{\lfloor k/2\rfloor}\|_F \le \|\boldsymbol{Y}_{k_\delta}\|_F - \nu \sum_{i=k_\delta}^{\lfloor k/2\rfloor - 1} \eta_i + S_{\lfloor k/2\rfloor}.$$

By (44), $\|\boldsymbol{Y}_{k_\delta}\|_F \le c\bar{B}\delta \log k + C_2$, while $\sum_{i=k_\delta}^{\lfloor k/2\rfloor-1}\eta_i \ge c(1-\delta)\log k - C_2$. Since $\nu(1-\delta) - \bar{B}\delta = \nu/2$, we obtain $S_{\lfloor k/2\rfloor} \ge \frac{c\nu}{2}\log k - C_3$ on E_1, and Lemma B.10 gives $\Pr[E_1] \le 2\exp\{-(\frac{c\nu}{2}\log k - C_3)^2 k^\delta/(2C_1)\}$.

E_2: $\|\boldsymbol{Y}_i\|_F < \rho/4$ for some $k_\delta \le i \le \lfloor k/2\rfloor$, and $\|\boldsymbol{Y}_k\|_F \ge \rho$. Let i_0 be the last index in $\{k_\delta, \dots, k-1\}$ with $\|\boldsymbol{Y}_{i_0}\|_F < \rho/4$. Then $\|\boldsymbol{Y}_{i_0+1}\|_F < \rho/4 + \bar{B}\eta_{k_\delta} \le 3\rho/8$, so $i_0 + 1 < k$, and $\|\boldsymbol{Y}_i\|_F \ge \rho/4$ for $i_0 < i < k$, where $\mu_i \le 0$ by (45). Hence

$$\frac{5\rho}{8} \le \|\boldsymbol{Y}_k\|_F - \|\boldsymbol{Y}_{i_0+1}\|_F \le \sum_{i=i_0+1}^{k-1} D_{i+1} = S_k - S_{i_0+1} \le 2\max_{k_\delta \le t \le k}|S_t|,$$

and Lemma B.10 gives $\Pr[E_2] \le 2\exp\{-(5\rho/16)^2 k^\delta/(2C_1)\}$.

Both bounds are at most $\frac{1}{2}\exp\{-k^{\delta/2}\}$ for all sufficiently large k, which proves the tail bound with $\delta_\rho := \delta/2$. Combining it with (44),

$$\mathbb{E}\left[(1 + \|\boldsymbol{Y}_k\|_F^2)\mathbf{1}\{\|\boldsymbol{Y}_k\|_F \ge \rho\}\right] \le \left(1 + (C_0 + c\bar{B}\log(k+2))^2\right)\exp\{-k^{\delta_\rho}\} = o(k^{-p}).$$

□

Asymptotic Excess Loss.

Proposition D.9. Let Assumption 3.5 hold, and let q_0 and λ be the constants of Lemma D.5(c) and Lemma D.6 for the chosen Π, and let $\Psi_{\boldsymbol{H}}$ and $t_{\boldsymbol{H}}$ be as in Theorem 3.6. Then

$$\lim_{k\to\infty} k\big(\mathbb{E}\left[f(\boldsymbol{X}_k)\right] - f_*\big) = \begin{cases} \dfrac{c^2 q_0}{2d}\displaystyle\sum_{j=1}^{d}\frac{h_j}{2c\lambda h_j - 1} = \frac{q_0}{\lambda^2}\Psi_{\boldsymbol{H}}(c\lambda), & c\lambda > \dfrac{1}{2h_{\min}}, \\ +\infty, & c\lambda \le \dfrac{1}{2h_{\min}}. \end{cases} \tag{46}$$

In particular, $\inf_{c>0}\lim_{k\to\infty} k(\mathbb{E}\left[f(\boldsymbol{X}_k)\right] - f_*) = \frac{q_0}{\lambda^2}\Psi_{\boldsymbol{H}}(t_{\boldsymbol{H}})$, attained at $c = t_{\boldsymbol{H}}/\lambda$.

Proof. Let $\boldsymbol{E}_k := \boldsymbol{X}_k - \boldsymbol{X}_*$. By (7) and (31),

$$\boldsymbol{E}_{k+1} = \boldsymbol{E}_k - \eta_k\Pi(\boldsymbol{H}\boldsymbol{E}_k + \boldsymbol{\Xi}_k), \qquad \mathbb{E}\left[f(\boldsymbol{X}_k)\right] - f_* = \tfrac{1}{2}\mathbb{E}\left[\langle \boldsymbol{E}_k, \boldsymbol{H}\boldsymbol{E}_k\rangle\right]. \tag{47}$$

Let $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top$ with $\boldsymbol{Q}$ orthogonal and $\boldsymbol{D} = \mathrm{diag}(h_1, \dots, h_d)$. By Lemma D.2(c), $\tilde{\boldsymbol{E}}_k := \boldsymbol{Q}^\top\boldsymbol{E}_k$ satisfies $\tilde{\boldsymbol{E}}_{k+1} = \tilde{\boldsymbol{E}}_k - \eta_k\Pi(\boldsymbol{D}\tilde{\boldsymbol{E}}_k + \boldsymbol{Q}^\top\boldsymbol{\Xi}_k)$, where the matrices $\boldsymbol{Q}^\top\boldsymbol{\Xi}_k$ are independent with the law of $\boldsymbol{\Xi}_k$ by Lemma D.1(c), $\tilde{\boldsymbol{E}}_0$ is deterministic, and $\langle \boldsymbol{E}_k, \boldsymbol{H}\boldsymbol{E}_k\rangle = \langle \tilde{\boldsymbol{E}}_k, \boldsymbol{D}\tilde{\boldsymbol{E}}_k\rangle$. Hence we may assume $\boldsymbol{H} = \boldsymbol{D}$, so that $(\boldsymbol{H}\boldsymbol{E})_{j,:} = h_j\boldsymbol{E}_{j,:}$ and

$$\mathbb{E}\left[f(\boldsymbol{X}_k)\right] - f_* = \tfrac{1}{2}\sum_{j=1}^{d} h_j\, s_{j,k}, \qquad s_{j,k} := \mathbb{E}\left[\|\boldsymbol{E}_{k,j,:}\|^2\right]. \tag{48}$$

Let $\boldsymbol{Y}_k := \boldsymbol{H}^{1/2}\boldsymbol{E}_k$. Then (47) reads $\boldsymbol{Y}_{k+1} = \boldsymbol{Y}_k - \eta_k \boldsymbol{\Phi}(\boldsymbol{Y}_k, \boldsymbol{\Xi}_k)$ with $\boldsymbol{\Phi}(\boldsymbol{Y}, \boldsymbol{\Xi}) := \boldsymbol{H}^{1/2}\Pi(\boldsymbol{H}^{1/2}\boldsymbol{Y} + \boldsymbol{\Xi})$, so $\|\boldsymbol{\Phi}\|_F \le \sqrt{h_{\max}}\, B_\Pi$ and $\boldsymbol{m}(\boldsymbol{Y}) = \boldsymbol{H}^{1/2}\boldsymbol{M}(\boldsymbol{H}^{1/2}\boldsymbol{Y})$. For $\|\boldsymbol{Y}\|_F \ge \rho$ and $\boldsymbol{Z} := \boldsymbol{H}^{1/2}\boldsymbol{Y}$ we have $\|\boldsymbol{Z}\|_F \ge \sqrt{h_{\min}}\,\|\boldsymbol{Y}\|_F \ge \sqrt{h_{\min}}\,\rho$ and $\langle \boldsymbol{Z}, \boldsymbol{M}(\boldsymbol{Z})\rangle > 0$ by (40), hence

$$\frac{\langle \boldsymbol{Y}, \boldsymbol{m}(\boldsymbol{Y})\rangle}{\|\boldsymbol{Y}\|_F} = \frac{\langle \boldsymbol{Z}, \boldsymbol{M}(\boldsymbol{Z})\rangle}{\|\boldsymbol{Y}\|_F} \ge \sqrt{h_{\min}} \inf_{\|\boldsymbol{W}\|_F \ge \sqrt{h_{\min}}\rho} \frac{\langle \boldsymbol{W}, \boldsymbol{M}(\boldsymbol{W})\rangle}{\|\boldsymbol{W}\|_F} > 0.$$

Lemma D.8 applies, and $\|\boldsymbol{E}_k\|_F \le h_{\min}^{-1/2}\|\boldsymbol{Y}_k\|_F$ yields, for all $\rho, p > 0$,

$$\mathbb{E}\left[(1 + \|\boldsymbol{E}_k\|_F^2)\,\mathbf{1}\{\|\boldsymbol{E}_k\|_F \ge \rho\}\right] = o(k^{-p}); \tag{49}$$

in particular, $\boldsymbol{E}_k \to 0$ in probability.

Since $\boldsymbol{\Xi}_k$ is independent of $\boldsymbol{E}_k$, expanding the jth row of (47) gives

$$s_{j,k+1} = s_{j,k} - 2\eta_k \mathbb{E}\left[\left\langle \boldsymbol{E}_{k,j,:}, \boldsymbol{M}(\boldsymbol{H}\boldsymbol{E}_k)_{j,:}\right\rangle\right] + \eta_k^2 \mathbb{E}\left[g_j(\boldsymbol{H}\boldsymbol{E}_k)\right], \qquad g_j(\boldsymbol{Z}) := \mathbb{E}_{\boldsymbol{\Xi}}\left[\|\Pi(\boldsymbol{Z} + \boldsymbol{\Xi})_{j,:}\|^2\right]. \tag{50}$$

By bounded convergence, g_j is continuous with $g_j \le B_\Pi^2$, and $g_j(0) = q_0/d$ by (36); since $\boldsymbol{H}\boldsymbol{E}_k \to 0$ in probability, $\mathbb{E}\left[g_j(\boldsymbol{H}\boldsymbol{E}_k)\right] \to q_0/d$. Fix $\epsilon \in (0, \min\{\lambda, q_0/(2d)\})$ and let ρ_ϵ be as in Lemma D.7(b). On $\{\|\boldsymbol{E}_k\|_F < \rho_\epsilon/h_{\max}\}$ we have $\|\boldsymbol{H}\boldsymbol{E}_k\|_F < \rho_\epsilon$, and (39) with $(\boldsymbol{H}\boldsymbol{E}_k)_{j,:} = h_j \boldsymbol{E}_{k,j,:}$ gives

$$(\lambda - \epsilon)h_j\|\boldsymbol{E}_{k,j,:}\|^2 \le \left\langle \boldsymbol{E}_{k,j,:}, \boldsymbol{M}(\boldsymbol{H}\boldsymbol{E}_k)_{j,:}\right\rangle \le (\lambda + \epsilon)h_j\|\boldsymbol{E}_{k,j,:}\|^2.$$

On the complementary event, the middle term is bounded in absolute value by $B_\Pi\|\boldsymbol{E}_k\|_F$ and $h_j\|\boldsymbol{E}_{k,j,:}\|^2 \le h_{\max}\|\boldsymbol{E}_k\|_F^2$, so (49) with $p = 3$ gives

$$(\lambda - \epsilon)h_j s_{j,k} - o(k^{-3}) \le \mathbb{E}\left[\left\langle \boldsymbol{E}_{k,j,:}, \boldsymbol{M}(\boldsymbol{H}\boldsymbol{E}_k)_{j,:}\right\rangle\right] \le (\lambda + \epsilon)h_j s_{j,k} + o(k^{-3}).$$

Substituting this and $|\mathbb{E}\left[g_j(\boldsymbol{H}\boldsymbol{E}_k)\right] - q_0/d| \le \epsilon$ into (50), and absorbing the $o(k^{-4})$ terms into $\epsilon\eta_k^2$, we obtain, for all sufficiently large k,

$$\Big(1 - \frac{2c(\lambda+\epsilon)h_j}{k+k_0}\Big)s_{j,k} + \frac{c^2(q_0/d - 2\epsilon)}{(k+k_0)^2} \;\le\; s_{j,k+1} \;\le\; \Big(1 - \frac{2c(\lambda-\epsilon)h_j}{k+k_0}\Big)s_{j,k} + \frac{c^2(q_0/d + 2\epsilon)}{(k+k_0)^2}. \tag{51}$$

If $2c\lambda h_j > 1$, take ϵ with $2c(\lambda - \epsilon)h_j > 1$; Lemma B.11 applied to both sides of (51) gives

$$\frac{c^2(q_0/d - 2\epsilon)}{2c(\lambda+\epsilon)h_j - 1} \le \liminf_{k\to\infty} k\, s_{j,k} \le \limsup_{k\to\infty} k\, s_{j,k} \le \frac{c^2(q_0/d + 2\epsilon)}{2c(\lambda-\epsilon)h_j - 1},$$

and $\epsilon \downarrow 0$ yields $\lim_k k s_{j,k} = \frac{c^2 q_0/d}{2c\lambda h_j - 1}$. If $2c\lambda h_j < 1$, take ϵ with $2c(\lambda+\epsilon)h_j < 1$; the lower bound in (51) and Lemma B.11(b) give $k s_{j,k} \to \infty$. If $2c\lambda h_j = 1$, the lower bound in (51) has coefficient $\gamma = 1 + 2c\epsilon h_j > 1$, so Lemma B.11(b) gives $\liminf_k k s_{j,k} \ge c^2(q_0/d - 2\epsilon)/(2c\epsilon h_j)$, which tends to $+\infty$ as $\epsilon \downarrow 0$. Summing over j in (48) proves (46); the second expression in the finite case is the substitution $t = c\lambda$ in the definition of $\Psi_{\boldsymbol{H}}$ in Theorem 3.6. Finally, the minimizer $t_{\boldsymbol{H}}$ of $\Psi_{\boldsymbol{H}}$ exists and is unique, since $\Psi_{\boldsymbol{H}}$ tends to $+\infty$ at both ends of its domain and $\Psi''_{\boldsymbol{H}}(t) = \frac{1}{d}\sum_{j=1}^d h_j(2th_j - 1)^{-3} > 0$; in particular, $0 < \Psi_{\boldsymbol{H}}(t_{\boldsymbol{H}}) < \infty$. Hence, for $c > 1/(2\lambda h_{\min})$ the limit is minimized at $c\lambda = t_{\boldsymbol{H}}$ with value $\frac{q_0}{\lambda^2}\Psi_{\boldsymbol{H}}(t_{\boldsymbol{H}})$, while the limit is $+\infty$ for all other c. □

Comparison of the Constants and Proof of Theorem 3.6.

Lemma D.10 (Norm clipping)**.** Under Assumption 3.5 with $\Pi = \mathrm{clip}(\cdot, \tau_{\mathrm{F}})$, for every $\tau_{\mathrm{F}} > 0$,

$$\frac{q_0}{\lambda_{\mathrm{F}}^2} \;\ge\; \frac{1}{\mathbb{E}\left[R^{-2}\right]} = \frac{r(\kappa+2)\,a_0^2}{\kappa}. \tag{52}$$

Proof. Since $1 - d^{-2} < 1$ and $\Pr[R = \tau_{\mathrm{F}}] = 0$, (38) gives

$$\lambda_{\mathrm{F}} \le \mathbb{E}\left[\mathbf{1}\{R \le \tau_{\mathrm{F}}\}\right] + \mathbb{E}\left[\tfrac{\tau_{\mathrm{F}}}{R}\mathbf{1}\{R > \tau_{\mathrm{F}}\}\right] = \mathbb{E}\left[\min\{R, \tau_{\mathrm{F}}\}/R\right].$$

By the Cauchy–Schwarz inequality and Lemma D.5(c),

$$\lambda_{\mathrm{F}}^2 \le \mathbb{E}\left[\min\{R, \tau_{\mathrm{F}}\}^2\right]\mathbb{E}\left[R^{-2}\right] = q_0\mathbb{E}\left[R^{-2}\right],$$

and $\mathbb{E}\left[R^{-2}\right] = \mathbb{E}\left[A^{-2}\right]/r = \kappa/(r(\kappa+2)a_0^2)$ by Lemma D.1(a). □

Lemma D.11 (Spectral clipping)**.** Let $\tau_{\mathrm{sp}} \le (1 - r/d)^2 a_0$, or $\tau_{\mathrm{sp}} \le a_0/2$ and $r \le d/2$. Then

$$\frac{\tau_{\mathrm{sp}}}{\lambda_{\mathrm{sp}}} < a_0 \sqrt{\frac{\kappa + 2}{\kappa}}. \tag{53}$$

Proof. By the definition of λ_{sp} in Theorem 3.6, $\lambda_{\mathrm{sp}} = \alpha_0 + \beta_0 \tau_{\mathrm{sp}}$ with

$$\alpha_0 := (1 - r/d)^2 > 0 \qquad \text{and} \qquad \beta_0 := \frac{r(4d - 3r - 1)\kappa}{2(\kappa + 1) a_0 d^2} > 0,$$

so $\lambda_{\mathrm{sp}} \ge \alpha_0$ and $\tau_{\mathrm{sp}} \mapsto \tau_{\mathrm{sp}}/\lambda_{\mathrm{sp}}$ is nondecreasing.

If $\tau_{\mathrm{sp}} \le \alpha_0 a_0$, then

$$\tau_{\mathrm{sp}}/\lambda_{\mathrm{sp}} \le \tau_{\mathrm{sp}}/\alpha_0 \le a_0 < a_0 \sqrt{(\kappa + 2)/\kappa}.$$

If $r \le d/2$ and $\tau_{\mathrm{sp}} \le a_0/2$, by monotonicity it suffices to take $\tau_{\mathrm{sp}} = a_0/2$. With $\varrho := r/d \in (0, 1/2]$ and $\frac{r(4d-3r-1)}{2} = 2r(d - r) + \frac{r(r-1)}{2} \ge 2r(d - r)$,

$$\lambda_{\mathrm{sp}} \;\ge\; (1 - \varrho)^2 + \varrho(1 - \varrho)\frac{\kappa}{\kappa + 1} \;:=\; \phi(\varrho).$$

Since $\phi'(\varrho) = -2(1 - \varrho) + (1 - 2\varrho)\frac{\kappa}{\kappa+1} < 0$ on $[0, 1/2]$, we get

$$\lambda_{\mathrm{sp}} \ge \phi(1/2) = \frac{2\kappa + 1}{4(\kappa + 1)},$$

and

$$(\kappa + 2)\lambda_{\mathrm{sp}}^2 - \frac{\kappa}{4} \;\ge\; \frac{(\kappa + 2)(2\kappa + 1)^2 - 4\kappa(\kappa + 1)^2}{16(\kappa + 1)^2} = \frac{4\kappa^2 + 5\kappa + 2}{16(\kappa + 1)^2} > 0,$$

that is, $\lambda_{\mathrm{sp}} > \frac{1}{2}\sqrt{\kappa/(\kappa + 2)}$, which is (53) for $\tau_{\mathrm{sp}} = a_0/2$. □

We now restate and prove Theorem 3.6.

Theorem 3.6. Consider problem (7) under Assumption 3.5, and let each method use the step sizes $\eta_k = \frac{c}{k + k_0}$ with $c, k_0 > 0$. Let $\tau_{\mathrm{sp}} > 0$ satisfy $\tau_{\mathrm{sp}} \le (1 - r/d)^2 a_0$, or $\tau_{\mathrm{sp}} \le a_0/2$ if $r \le d/2$. Set

$$\lambda_{\mathrm{sp}} := \left(1 - \tfrac{r}{d}\right)^2 + \tfrac{r(4d-3r-1)\,\kappa\tau_{\mathrm{sp}}}{2(\kappa+1)\,a_0 d^2}, \qquad \Psi_{\boldsymbol{H}}(t) := \tfrac{t^2}{2d} \textstyle\sum_{j=1}^{d} \frac{h_j}{2th_j - 1} \quad \text{for } t > \frac{1}{2h_{\min}},$$

and let $t_{\boldsymbol{H}}$ be the unique minimizer of $\Psi_{\boldsymbol{H}}$. Let $\boldsymbol{X}_k^{\mathrm{sp}}$ follow (2) with $\tau_k = \tau_{\mathrm{sp}}$ and $c = t_{\boldsymbol{H}}/\lambda_{\mathrm{sp}}$, and let $\boldsymbol{X}_k^{\mathrm{F}}$ follow (1) with the Frobenius norm, any threshold $\tau_{\mathrm{F}} > 0$ and any $c > 0$. Then

$$\lim_{k\to\infty} k\left(\mathbb{E}\left[f(\boldsymbol{X}_k^{\mathrm{sp}})\right] - f_*\right) = \frac{r\tau_{\mathrm{sp}}^2\,\Psi_{\boldsymbol{H}}(t_{\boldsymbol{H}})}{\lambda_{\mathrm{sp}}^2} < \frac{r(\kappa+2)a_0^2\,\Psi_{\boldsymbol{H}}(t_{\boldsymbol{H}})}{\kappa} \le \lim_{k\to\infty} k\left(\mathbb{E}\left[f(\boldsymbol{X}_k^{\mathrm{F}})\right] - f_*\right). \tag{9}$$

Proof. The setting of Section 3.1 is problem (7) under Assumption 3.5.

Spectral clipping. By Lemma D.5(c) and Lemma D.6, $(q_0, \lambda) = (r\tau_{\mathrm{sp}}^2, \lambda_{\mathrm{sp}})$. For $c = t_{\boldsymbol{H}}/\lambda_{\mathrm{sp}}$ we have $c\lambda = t_{\boldsymbol{H}} > 1/(2h_{\min})$, so Proposition D.9 gives

$$\lim_{k\to\infty} k\left(\mathbb{E}\left[f(\boldsymbol{X}_k^{\mathrm{sp}})\right] - f_*\right) = \frac{r\tau_{\mathrm{sp}}^2\,\Psi_{\boldsymbol{H}}(t_{\boldsymbol{H}})}{\lambda_{\mathrm{sp}}^2}.$$

Norm clipping. For any $\tau_{\mathrm{F}}, c > 0$, $(q_0, \lambda) = \left(\mathbb{E}\left[\min\{R, \tau_{\mathrm{F}}\}^2\right], \lambda_{\mathrm{F}}\right)$, and Proposition D.9 together with Lemma D.10 gives

$$\lim_{k\to\infty} k\left(\mathbb{E}\left[f(\boldsymbol{X}_k^{\mathrm{F}})\right] - f_*\right) \;\ge\; \frac{q_0}{\lambda_{\mathrm{F}}^2}\,\Psi_{\boldsymbol{H}}(t_{\boldsymbol{H}}) \;\ge\; \frac{r(\kappa+2)a_0^2\,\Psi_{\boldsymbol{H}}(t_{\boldsymbol{H}})}{\kappa}.$$

Comparison. Since $\Psi_{\boldsymbol{H}}(t_{\boldsymbol{H}}) > 0$, Lemma D.11 gives $\frac{r\tau_{\mathrm{sp}}^2\,\Psi_{\boldsymbol{H}}(t_{\boldsymbol{H}})}{\lambda_{\mathrm{sp}}^2} < \frac{r(\kappa+2)a_0^2\,\Psi_{\boldsymbol{H}}(t_{\boldsymbol{H}})}{\kappa}$. Combining the three displays yields (9). □

Remark D.12 (Extensions). (i) If the r nonzero singular values of $\mathbf{\Xi}_k$ are independent Pareto variables $A_{k,1}, \dots, A_{k,r}$ instead of a common A_k, the proof applies verbatim: in case (iv) of Lemma D.4(b), the block $\left(\begin{smallmatrix} A_i & s \\ 0 & A_j \end{smallmatrix}\right)$ has polar factor with $(1,2)$ entry $s/\sqrt{(A_i + A_j)^2 + s^2}$, so $\tau_{\mathrm{sp}}/(2A)$ becomes $\tau_{\mathrm{sp}}/(A_i + A_j)$ and

$$\lambda_{\mathrm{sp}} = \left(1 - \frac{r}{d}\right)^2 + \left(2r(d-r)\mathbb{E}\left[A^{-1}\right] + r(r-1)\mathbb{E}\left[(A_1 + A_2)^{-1}\right]\right)\frac{\tau_{\mathrm{sp}}}{d^2};$$

in Lemma D.10, $R = \|\mathbf{\Xi}\|_F = (\sum_i A_i^2)^{1/2}$ and $(\sum_i A_i^2)(\sum_i A_i^{-2}) \geq r^2$ give $\mathbb{E}\left[R^{-2}\right] \leq \mathbb{E}\left[A^{-2}\right]/r$. The conditions on τ_{sp} in Theorem 3.6 are unchanged.

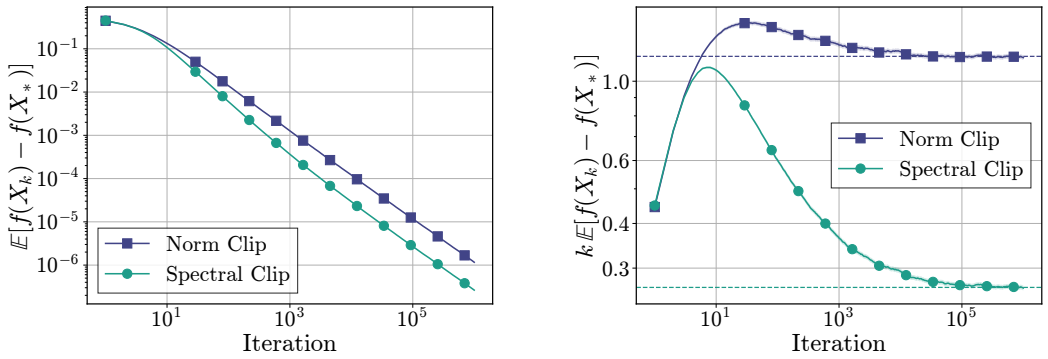
(ii) For $r = d$, the intercept $(1 - r/d)^2$ vanishes, $q_0/\lambda_{\mathrm{sp}}^2$ no longer depends on τ_{sp}, and the argument gives no separation.

Numerical Illustration. In this section, we numerically verify Theorem 3.6. Here, we consider problem (7) with $d = 3$ and rank-one noise ($r = 1$), where $\boldsymbol{H} = \boldsymbol{Q}\,\mathrm{diag}(1, 2, 4)\,\boldsymbol{Q}^\top$ for a fixed orthogonal $\boldsymbol{Q}$, $\boldsymbol{X}_*$ is a fixed dense matrix, and $\boldsymbol{X}_0 = 0$. The noise satisfies Assumption 3.5 with $a_0 = 1$ and $\kappa = 1.5$, and both methods follow (31) with $k_0 = 10$.

For spectral clipping, we use $\tau_{\mathrm{sp}} = a_0/2$ and $c = t_{\boldsymbol{H}}/\lambda_{\mathrm{sp}}$, as in Theorem 3.6. For norm clipping, we use $\tau_{\mathrm{F}} \approx 1.240$ and $c \approx 1.210$, which minimize its asymptotic constant $\frac{q_0}{\lambda_{\mathrm{F}}^2}\Psi_{\boldsymbol{H}}(c\lambda_{\mathrm{F}})$ from Proposition D.9; Theorem 3.6 holds for any τ_{F} and c. Since $t_{\boldsymbol{H}} \approx 0.801$ and $\Psi_{\boldsymbol{H}}(t_{\boldsymbol{H}}) \approx 0.354$, the limits in (46) equal 0.265 for spectral clipping and 1.174 for norm clipping.

We run 10^6 iterations for 4096 independent trajectories, and at each iteration both methods receive the same noise sample. In Figure 4, we plot the mean excess loss $\mathbb{E}\left[f(\boldsymbol{X}_k)\right] - f_*$ on log-log axes (a) and $k\big(\mathbb{E}\left[f(\boldsymbol{X}_k)\right] - f_*\big)$ together with the predicted limits (b). Both methods converge at the rate $1/k$.

At $k = 10^6$, the quantity $k\big(\mathbb{E}\left[f(\boldsymbol{X}_k)\right] - f_*\big)$ equals 0.264 ± 0.004 for spectral clipping and 1.164 ± 0.018 for norm clipping (95% confidence intervals), which is close to the predicted limits 0.265 and 1.174. Thus, the excess loss of norm clipping is 4.41 times larger than that of spectral clipping, while the predicted ratio is $1.174/0.265 \approx 4.43$. Moreover, spectral clipping has a lower mean excess loss than norm clipping at every recorded iteration $k \geq 3$, so the separation holds not only in the limit.



(a) Expected excess loss

(b) Rescaled excess loss

Figure 4: Spectral versus norm clipping on the problem of Theorem 3.6 with $d = 3$ and $r = 1$. (a) $\mathbb{E}\left[f(\boldsymbol{X}_k)\right] - f_*$. (b) $k\big(\mathbb{E}\left[f(\boldsymbol{X}_k)\right] - f_*\big)$; dashed lines show the limits predicted by (46). Curves show means over 4096 trajectories.

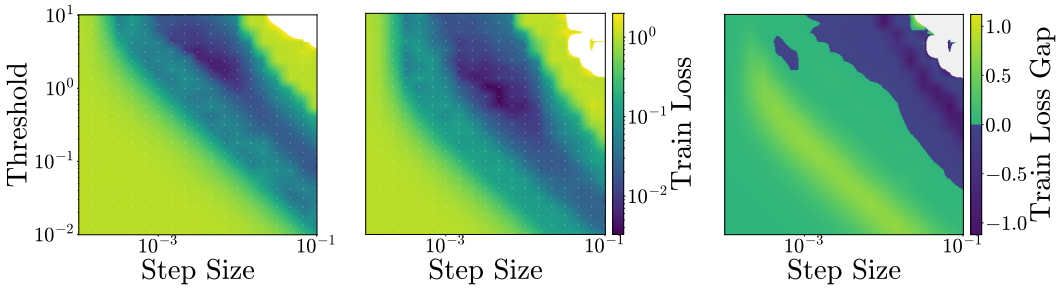
E Missing Pseudocode of Algorithm 3

Algorithm 3 SGDM with Quantile-based Spectral Clipping

Require: Initial point $\boldsymbol{X}_0 \in \mathbb{R}^{m\times n}$, momentum matrix $\boldsymbol{B}_{-1} = 0$, momentum parameter $\beta \in [0, 1)$, quantile $q \in (0, 1]$, window size $w \geq 1$, initial threshold $\tau_{-1} = +\infty$.

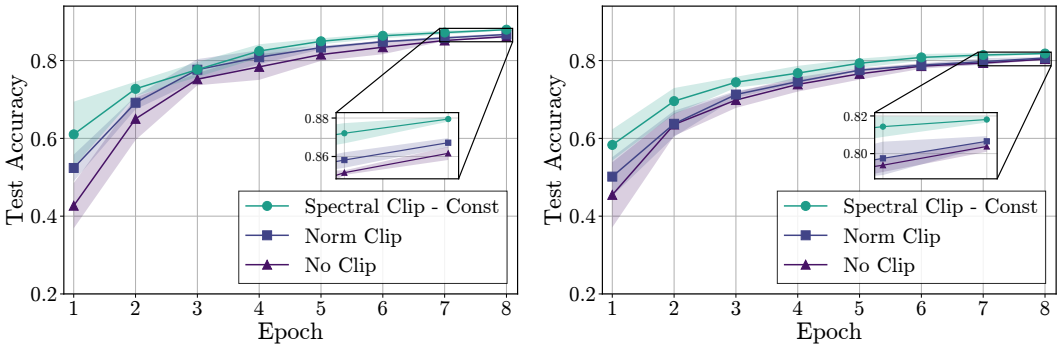
Initialize multiset $\mathcal{S}_{-1} \leftarrow \emptyset$.
for $k = 0, 1, \ldots, K-1$ **do**
  Compute gradient $\boldsymbol{G}_k$.
  $\boldsymbol{B}_k \leftarrow \beta \boldsymbol{B}_{k-1} + (1-\beta)\, \mathcal{C}_{\tau_{k-1}}(\boldsymbol{G}_k)$.
  $\boldsymbol{X}_{k+1} \leftarrow \boldsymbol{X}_k - \eta \boldsymbol{B}_k$.
  $\mathcal{S}_k \leftarrow \mathcal{S}_{k-1} \cup \sigma_{\max}(\boldsymbol{G}_k)$.
  if $k \geq w$ **then**
    $\mathcal{S}_k \leftarrow \mathcal{S}_k \setminus \sigma_{\max}(\boldsymbol{G}_{k-w})$.
  end if
  $\tau_k \leftarrow \mathcal{Q}_q(\mathcal{S}_k)$.
end for

F Missing Figures



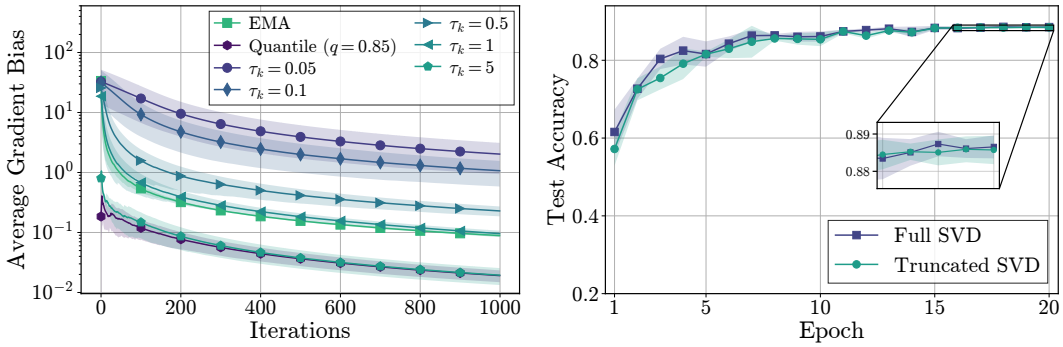
(a) Norm clipping (b) Spectral clipping (c) Comparison of Training Loss

Figure 5: Heatmap of train loss in logarithmic scale. Left: norm clipping. Middle: spectral clipping. Right: train loss gap between norm clipping and spectral clipping. Green color shows regions where spectral clipping is better, while purple one shows regions where norm clipping is better.



(a) Random-1: 17.18% label noise (b) Worst: 40.14% label noise

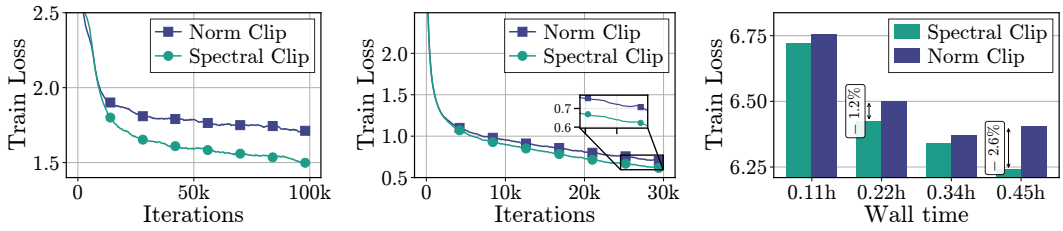
Figure 6: Test accuracy of CifarNet trained with SGDM using no clipping, norm clipping, and spectral clipping on CIFAR-10N.



(a) Comparison of τ_k

(b) Test accuracy on CIFAR-10N Aggregate: full vs. truncated SVD

Figure 7: In Figure 7a, we compare the performance of different spectral clipping strategies on trace regression problem (13). In Figure 7b, we compare full SVD and randomized truncated SVD ($r = 10$) for spectral clipping on CIFAR-10N Aggregate.



(a) SGDM

(b) Muon

(c) SGDM (wall-clock time)

Figure 8: In Figure 8a and Figure 8b, we show training loss of NanoGPT on the Shakespeare dataset for different optimizers. In Figure 8c we show the wall time comparison of GPT-2 pretraining experiment.

G Hyperparameter Tuning Details

All experiments were conducted on a system equipped with 4× NVIDIA A100 GPUs, and hyperparameters were tuned using Bayesian optimization via Weights & Biases (Biewald, 2020). Spectral clipping is implemented as in Algorithm 2: we call `torch.svd_lowrank`$(\boldsymbol{G}, q = r + p, \texttt{niter})$ and keep the top r components, with $r = 10$, $p = 5$, $\texttt{niter} = 1$ for CIFAR-10N; $r = 3$, $p = 5$, $\texttt{niter} = 2$ for NanoGPT; $r = 64$, $p = 10$, $\texttt{niter} = 2$ for GPT-2 with SGDM and Adam; and $r = 8$, $p = 5$, $\texttt{niter} = 1$ for GPT-2 with Muon. The MLP and trace regression experiments use the exact SVD.

Table 1: Best hyperparameters obtained for the MLP experiments.

Method	Learning rate η	Clipping parameter
SGDM Norm Clip	3.48×10^{-3}	$\texttt{max_norm} = 3.28$
SGDM Spectral Const	3.56×10^{-3}	$\tau = 0.891$
SGDM Spectral EMA	1.39×10^{-3}	—
SGDM Spectral Quantile	8.54×10^{-3}	$q = 0.754$
Adam No Clip	1.21×10^{-3}	—
Adam Norm Clip	9.00×10^{-4}	$\texttt{max_norm} = 1.64$

MLP. We sweep the learning rate $\eta \in [10^{-4}, 10^{-1}]$ for every method, the norm-clipping threshold and the constant spectral-clipping threshold over $[10^{-2}, 10]$, the EMA coefficient is fixed at $\theta = 0.9$, and the quantile level is swept over $q \in [0.5, 0.99]$ with a fixed sliding-window of length $w = 100$. Each Bayesian sweep uses 300 trials when two hyperparameters are searched and 100 trials otherwise. The resulting hyperparameters are reported in Table 1.

Table 2: Best hyperparameters obtained for the CIFAR-10N experiments. "filter / head / bias" denote the three parameter groups used by `airbench-94`.

Method	η_{filter}	η_{head}	η_{bias}	β_{filter}	β_{head}	Clipping parameter
SGDM No Clip	3.01×10^{-3}	8.32×10^{-2}	6.68×10^{-2}	0.868	0.752	—
SGDM Norm Clip	1.52×10^{-1}	0.102	3.48×10^{-2}	0.932	0.936	`max_norm`$= 5.29$
SGDM Spectral Const	7.67×10^{-2}	0.126	7.38×10^{-2}	0.649	0.574	$\tau = 2.239$

CV. For the CIFAR-10N experiments with SGDM, we searched over clipping thresholds in $[10^{-2}, 20]$ for norm clipping and in $[10^{-3}, 5]$ for spectral clipping. In addition, we searched over bias learning rates in $[5\times10^{-3}, 0.2]$, filter learning rates in $[5\times10^{-4}, 0.3]$, head learning rates in $[5\times10^{-2}, 1.5]$, and filter and head momentum parameters in $[0.5, 0.98]$. For each method, we evaluated 200 random configurations with 3 seeds each on Aggregate labels and selected the one with the highest mean validation accuracy (with test-time augmentation) on 5,000 held-out training images with their original clean labels; it was reused unchanged for Random-1 and Worst. The resulting hyperparameters are reported in Table 2. For convolutional layers, we apply spectral clipping after reshaping each gradient tensor of shape $C_{\text{out}} \times C_{\text{in}} \times h \times w$ into a matrix of size $C_{\text{out}} \times (C_{\text{in}} \cdot h \cdot w)$.

Table 3: Best hyperparameters obtained for the NanoGPT experiment.

Method	Learning rate η	Clipping parameter
SGDM Norm Clip	7.52×10^{-3}	`max_norm`$= 3.70$
SGDM Spectral Clip	5.39×10^{-3}	$\tau = 1.89\times10^{-2}$
Muon Norm Clip	8.74×10^{-3}	`max_norm`$= 5.53\times10^{-2}$
Muon Spectral Clip	7.87×10^{-3}	$\tau = 0.101$

NanoGPT. For the NanoGPT experiment with SGDM and Muon we searched the learning rate over $[10^{-4}, 10^{-1}]$ for both clipping rules, the norm–clipping threshold over $[10^{-2}, 10^{1}]$, and the spectral clipping threshold τ over $[10^{-2}, 10^{1}]$. Each search was run for 300 iterations. The resulting hyperparameters are reported in Table 3.

Table 4: Best hyperparameters obtained for the GPT-2 pretraining experiment.

Method	Learning rate η	Clipping parameter
SGDM Norm Clip	7.04×10^{-2}	`max_norm`$= 2.18$
SGDM Spectral Clip	9.42×10^{-2}	$\tau = 4.34\times10^{-3}$
Adam Norm Clip	1.13×10^{-3}	`max_norm`$= 4.56\times10^{-2}$
Adam Spectral Clip	1.89×10^{-3}	$\tau = 7.39\times10^{-3}$
Muon Norm Clip	1.24×10^{-2}	`max_norm`$= 2.06$
Muon Spectral Clip	1.32×10^{-2}	$\tau = 1.03\times10^{-2}$

GPT-2. For the GPT-2 pretraining experiment on FineWeb with SGDM, Adam and Muon, we searched the learning rate over $[10^{-4}, 10^{-1}]$ for all clipping rules, the norm-clipping threshold over $[10^{-2}, 5]$, and the spectral-clipping threshold τ over $[10^{-3}, 10^{-1}]$. Each search was run for 50 iterations. The resulting hyperparameters are reported in Table 4.

H Memory Overhead

We measure the GPU memory of one AdamW training step of GPT-2 models of four sizes on a single NVIDIA RTX PRO 6000 GPU, with bf16 autocast, sequence length 1024 and global batch size 128. We record the peak allocated memory of the whole step and, separately, the memory that the clipping routine allocates on top of the memory already in use. Spectral clipping uses Algorithm 2 with $r = 64$ and two power iterations.

Table 5 shows that spectral clipping changes the peak memory of a training step by at most 0.15%. The clipping routine itself allocates at most 88 MiB; it runs after the backward pass, when activations have been freed, so this memory stays well below the peak.

Table 5: Peak GPU memory of one training step and memory allocated by the clipping routine for GPT-2 models.

			Peak memory (GiB)		Clipping memory (MiB)	
Model	Parameters	Microbatch	Norm	Spectral	Norm	Spectral
GPT-2	124M	64	36.27	36.26	0.4	19.1
GPT-2 Medium	355M	16	23.55	23.55	1.0	34.0
GPT-2 Large	774M	8	28.16	28.15	2.0	51.9
GPT-2 XL	1.56B	4	38.74	38.80	3.0	87.8

I Gradient Spectral Profiles under Controlled Token Replacement

This appendix provides additional results supporting the motivation in the main text. Figure 9 shows the results of introducing bad samples in the batch of good samples for different layers and batch sizes.

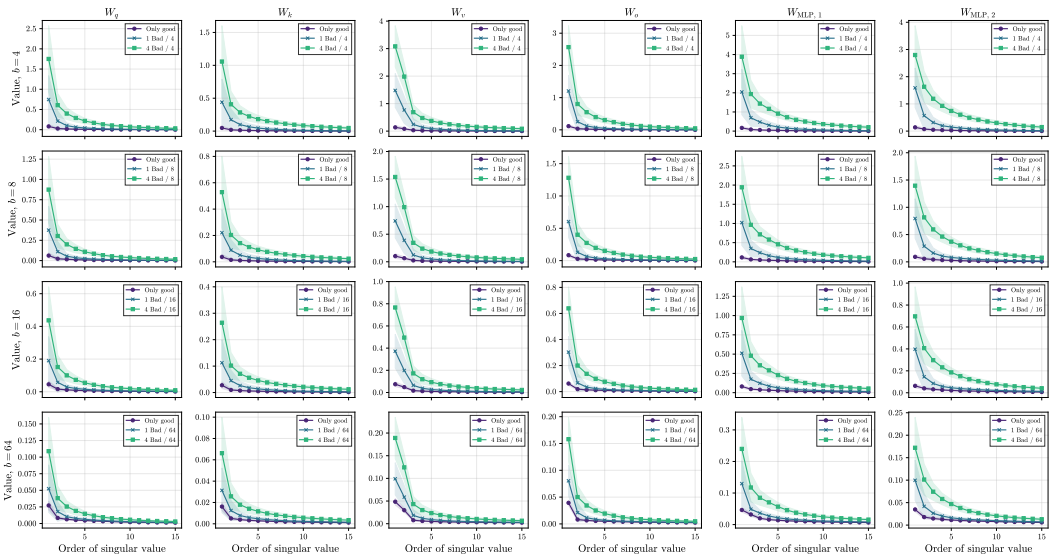


Figure 9: Spectral profile under controlled token replacement of the first feed-forward matrix $W_{\text{MLP},1}$ in layer **5** of GPT-2 Small for different batch sizes. First, the batch contains only good samples with $\tau_{\text{good}} = 0.95$, then we replacing one and four of them with bad samples with $\tau_{\text{bad}} = 10^{-7}$.

J Comparison of Spectral Clipping Before and After Momentum

We compare spectral clipping before and after momentum accumulation under heavy-tailed gradient noise, referred to as pre-clipping and post-clipping by Jiang et al. (2026).

For that we sample a $d \times d$ matrix with independent standard Gaussian entries and normalize it to obtain $\boldsymbol{G}$ with $\|\boldsymbol{G}\|_2 = 1$, where $d = 16$.

Let $\boldsymbol{G} = \nabla f(\boldsymbol{X})$ be the gradient at a fixed point $\boldsymbol{X}$. At each iteration, we observe

$$\boldsymbol{G}_k = \boldsymbol{G} + \sigma \boldsymbol{Z}_k + s_k \boldsymbol{u}_k \boldsymbol{v}_k^\top, \tag{54}$$

where $\sigma = 0.05$, $\boldsymbol{Z}_k$ has independent $\mathcal{N}(0,1)$ entries, $\boldsymbol{u}_k, \boldsymbol{v}_k \sim \text{Unif}(\mathbb{S}^{d-1})$, and s_k follows the two-sided Pareto distribution with shape $\alpha = 1.1$ and scale $x_m = 2.0$, as in Section 6.3.

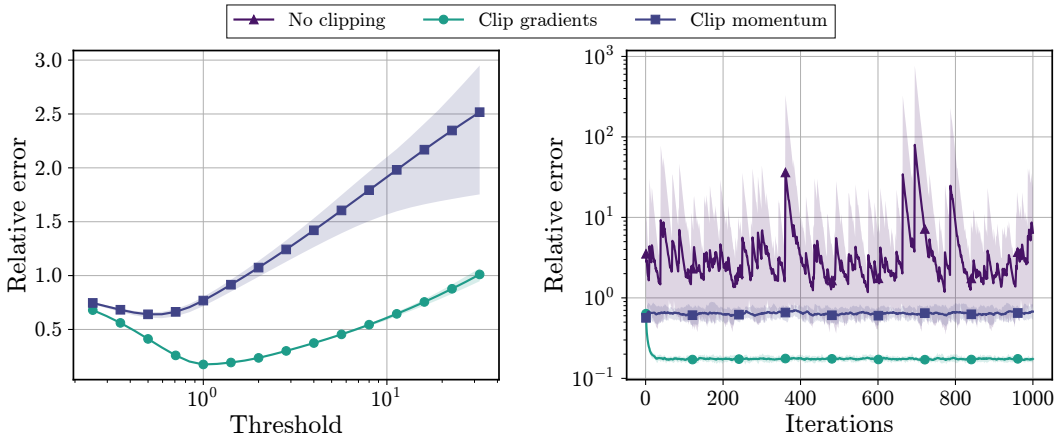


Figure 10: Comparison of spectral clipping of gradients and updates. Left: time-averaged relative Frobenius error versus threshold. Right: relative Frobenius error.

Let $\mathcal{C}_\tau$ denote the spectral clipping operator defined in Section 2.2. With $\beta = 0.9$ and $\boldsymbol{B}_{-1}^{\mathrm{raw}} = \boldsymbol{B}_{-1}^{\mathrm{grad}} = 0$, we compute,

$$\boldsymbol{B}_k^{\mathrm{raw}} = \beta \boldsymbol{B}_{k-1}^{\mathrm{raw}} + (1-\beta)\boldsymbol{G}_k, \tag{55}$$

$$\boldsymbol{B}_k^{\mathrm{grad}} = \beta \boldsymbol{B}_{k-1}^{\mathrm{grad}} + (1-\beta)\mathcal{C}_{\tau_{\mathrm{grad}}}(\boldsymbol{G}_k). \tag{56}$$

We compare the following bias-corrected estimates:

$$\widehat{\boldsymbol{G}}_k^{\mathrm{raw}} = \frac{\boldsymbol{B}_k^{\mathrm{raw}}}{1-\beta^{k+1}}, \tag{57}$$

$$\widehat{\boldsymbol{G}}_k^{\mathrm{grad}} = \frac{\boldsymbol{B}_k^{\mathrm{grad}}}{1-\beta^{k+1}}, \tag{58}$$

$$\widehat{\boldsymbol{G}}_k^{\mathrm{mom}} = \mathcal{C}_{\tau_{\mathrm{mom}}}\left(\frac{\boldsymbol{B}_k^{\mathrm{raw}}}{1-\beta^{k+1}}\right). \tag{59}$$

For each method, we measure the relative Frobenius error

$$\frac{\|\widehat{\boldsymbol{G}}_k - \boldsymbol{G}\|_F}{\|\boldsymbol{G}\|_F},$$

where $\widehat{\boldsymbol{G}}_k$ denotes the corresponding estimate. The thresholds τ_{grad} and τ_{mom} are tuned separately.

Figure 10 shows that gradient clipping gives lower estimation error than momentum clipping. Clipping each noisy gradient before momentum accumulation limits the contribution of large perturbations to the momentum buffer.