PROVABLY FINDING LOW NEURAL LOSS VIA SPECTRAL-GD

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimizers of the Muon family are rapidly becoming a competitive alternative to the standard adaptive methods for training LLMs. Such optimizers update the training parameters exploiting the intrinsically rich geometry of the space of the architectural weight matrices. In this work we do a first-of-its-kind analysis of Spectral-GD - the full-batch, momentum-free core of Muon - to show that it can provably train a shallow ReLU net in the large width regime. In terms of any set accuracy target and failure probability over random initialization, we give explicit estimates for the required width, step-length and number of iterations to reach the accuracy target. Our work rigorously establishes the surprise that gradient direction information, in the matrix sense, can be enough to train wide nets.

1 INTRODUCTION

The Muon family of optimisers, introduced by Jordan et al. (2024) in the context of the `modded-nanogpt` training-speed benchmark, has attracted growing attention as a competitive alternative to AdamW (Loshchilov & Hutter, 2017) for large-scale language-model training. A rapidly evolving ecosystem of Muon variants has since emerged including Polar Express (**?**), PolarGrad (Lau et al., 2025), NorMuon (Li et al., 2025), and Aurora (Dewulf et al., 2026), each modifying the basic polar-factor update to address aspects such as finite-precision arithmetic, gradient anisotropy, neuron-norm uniformity, and layer-wise heterogeneity in transformer architectures. At its core, every Muon step replaces the raw gradient with its *polar factor*, the matrix obtained by setting all nonzero singular values to one in the SVD, discarding all magnitude information. The full-batch, momentum-free special case of Muon, which we call *Spectral-GD*, was identified and given its geometric foundation by Bernstein & Newhouse (2024), it can be seen as the analytic predecessor of the momentum-augmented stochastic variant of Jordan et al. (2024). Given weights $\boldsymbol{W}_t \in \mathbb{R}^{d\times m}$ and gradient $\boldsymbol{G}_t = \nabla L(\boldsymbol{W}_t)$ with $\boldsymbol{G}_t = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, the Spectral-GD update with constant step size $\eta > 0$ is

$$\boldsymbol{W}_{t+1} \;=\; \boldsymbol{W}_t \;-\; \eta\,\mathrm{Polar}(\boldsymbol{G}_t), \qquad \mathrm{Polar}(\boldsymbol{G}_t) \;:=\; \boldsymbol{U}\boldsymbol{V}^\top, \tag{1}$$

Appendix A reviews that $-\mathrm{Polar}(\boldsymbol{G}_t)$ is the steepest direction inside a spectral-norm ball of radius η, so equation 1 is a linear minimization oracle for a trust region measured in the spectral norm; and that the same matrix is a subgradient of the nuclear norm evaluated at the gradient, which is the dual-norm counterpart of that statement.

Spectral-GD is the simplest member of the Muon family that retains the defining polar-factor step, and is the natural starting point for non-asymptotic convergence analysis.

We note that recently questions have been raised about the validity of the Muon algorithm (Shumaylov et al., 2026) and this work of ours can be seen as a counterpoint to that providing rigorous neural training guarantees for the core logic of Muon

Empirical Evidence for Spectral-GD and Muon. To the best of our knowledge empirical evidence for Spectral-GD as a standalone algorithm (full-batch, momentum-free) is limited in the literature; the bulk of practical work has focused on the stochastic, momentum-augmented variant Muon. Muon was introduced as part of the modded-nanogpt speedrun benchmark Jordan et al. (2024). In particular, from Figure 2 of Jordan et al. (2024), we note that Muon achieves a strictly lower validation loss than Adam, DistributedShampoo, and SOAP at every wallclock budget. Figure 3 of Jordan et al. (2024) further shows that this advantage persists at the 1.5B-parameter scale: Muon with "speedrun

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.

tweaks" reaches lower FineWeb validation loss than tuned AdamW across all token budgets, and, Muon trained on 25% fewer tokens still matches the AdamW baseline trained to convergence.

Muon has since been adopted in large-scale language-model training pipelines. Liu et al. (2025) establish its scalability by identifying two key modifications - weight decay and per-parameter update-scale adjustment - that allow Muon to match AdamW hyperparameters out of the box; their scaling-law experiments show that Muon achieves approximately $2\times$ the computational efficiency of AdamW under compute-optimal training, and their Moonlight model (3B/16B MoE, trained on 5.7T tokens) improves the frontier of performance versus training FLOPs over prior models. Team et al. (2025) extend this to the trillion-parameter regime with Kimi K2 (1T parameters, 15.5T tokens), where Muon substantially outperforms AdamW in token efficiency at fixed model size; they introduce MuonClip, a variant that resolves Muon's training instability from exploding attention logits via per-head query-key rescaling, enabling spike-free training at unprecedented scale.

Existing theory for Spectral-GD and Muon. For Spectral-GD itself, no non-asymptotic convergence guarantee has been established in any non-convex setting prior to this work. The most directly related theoretical analyses are present in this paper Bernstein & Newhouse (2024). For Muon, recent theoretical work has begun to clarify its optimization geometry and implicit regularization properties. Chen et al. (2025) place Muon within the Lion-$\mathcal{K}$ family of optimizers and prove that, for any L-smooth (possibly non-convex) objective, Muon with decoupled weight decay implicitly solves a spectral-norm-constrained problem and converges to its KKT points at rate $O(1/\sqrt{T})$.

Thus we conclude that existing guarantees for Spectral-GD or its variants do not explain why low neural net losses are attainable via these algorithms.

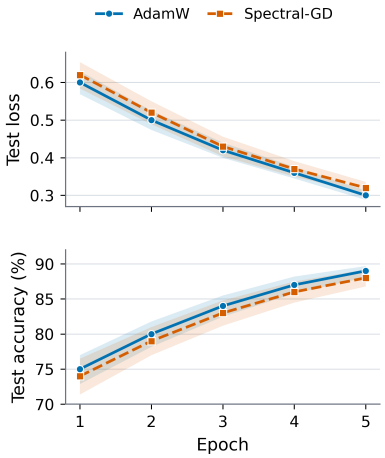
The Neural Tangent Kernel Regime Pioneering works like (Jacot et al., 2018; Du et al., 2018b; Chizat et al., 2018) provided the canonical "lazy-training" framework in which the network's prediction map is well-approximated by its first-order Taylor expansion throughout training. This powerful technique of proving trainabily of nets by (S)GD relies on the observation that, in the limit of asymptically large width, if the weights of the outer layer are scaled in an appropriate way then the weights can be initialized such that they always remain within a bounded set, throughout training, allowing the"neural tangent kernel" to be treated as approximately constant. These results are further reviewed in the last appendix and our result borrows methods and ideas from the NTK framework.

Motivating LLM Experiments for Spectral-GD. We compare Spectral-GD against both AdamW and Muon in two regimes: fine-tuning a pretrained model, and pretraining a model from a random initialization respectively.

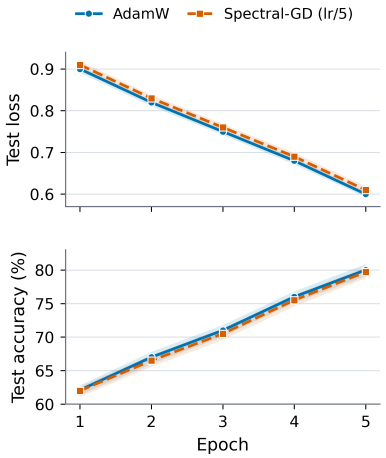
Fine-tuning. We compare Spectral-GD with AdamW (Loshchilov & Hutter, 2017) on two language-model fine-tuning benchmarks: the mathematical-reasoning dataset GSM8K (Cobbe et al., 2021) with the Qwen-3 1.7B model (Yang et al., 2025), and the preference-alignment dataset kto-mix-14k (Ethayarajh et al., 2024) with the Qwen-1.5 0.5B model (Yang et al., 2025), each over 5 epochs and 3 seeds. The test-loss and test-accuracy trajectories are shown in Figures 1a and 1b. On both benchmarks, Spectral-GD tracks AdamW closely throughout training, the curves overlap within their $\pm 1\sigma$ confidence bands for most of the training duration, with AdamW performing marginally better on later epochs. The competitive trajectory of Spectral-GD demonstrates that the polar-factor update can achieve AdamW-level fine-tuning performance.

Pretraining. We also compare full-batch Spectral-GD against full-batch Muon for training Pythia-70M language model (Biderman et al., 2023) from a random initialization on FineWeb (Penedo et al., 2024). Training and validation loss over 5000 iterations are shown in Figure 1c. Spectral-GD tracks Muon closely throughout training and separated only by a small, stable gap on the validation loss. We repeat this comparison for the modded-nanogpt model (Jordan et al., 2024), also trained from random initialization on FineWeb, shown in Figure 1d; the same pattern holds. Thus we give evidence that Spectral-GD can be a good model of Muon's behaviour in language model pretraining tasks.

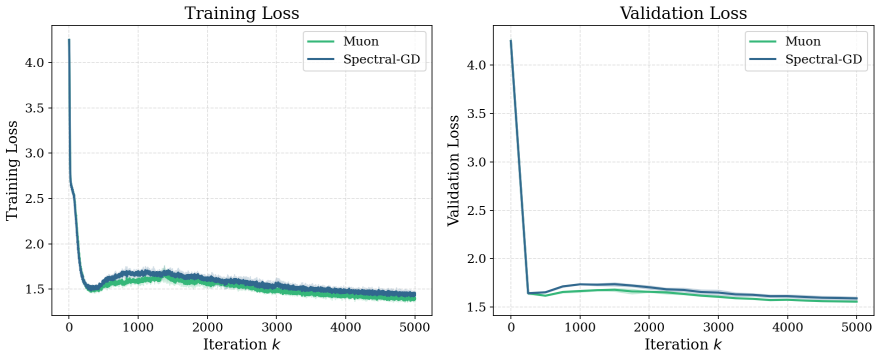
Theorem (**Informal Statement of the Main Result**). *Consider a width-m two-layer ReLU network whose first layer matrix is being updated via Spectral-GD on the empirical squared loss evaluated on this architecture. Suppose, n is the number of unit-norm training data with $\lambda_0 > 0$ being its smallest eigenvalue. Then for target accuracy $\varepsilon \in (0,1)$ and failure probability $\delta \in (0,1)$ and $m = \Omega\Big(\frac{n^5}{\lambda_0^4 \delta^3}\Big)$,*



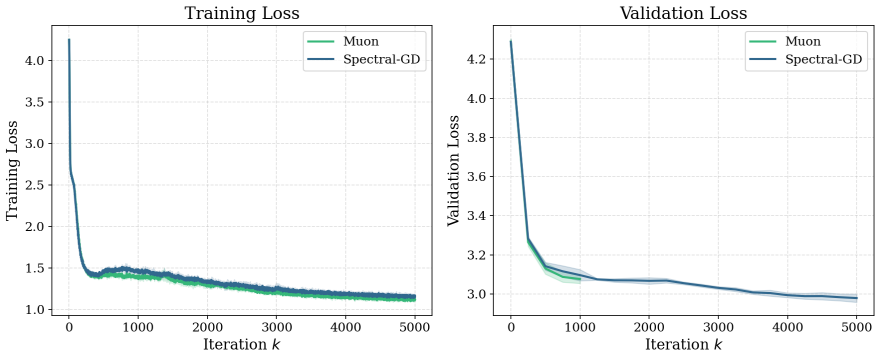
(a) GSM8K (Qwen-3 1.7B), fine-tuning.



(b) kto-mix-14k (Qwen-1.5 0.5B), fine-tuning.



(c) Pythia, pretraining from random initialization on the Fineweb.



(d) modded-nanogpt, pretraining from random initialization on the Fineweb.

Figure 1: (a)–(b) Test loss (top) and test accuracy (bottom) over 5 fine-tuning epochs. Blue: AdamW; orange: Spectral-GD (lr/5 on kto-mix-14k). Bands are $\pm\sigma$ over 3 seeds. (c)–(d) Training loss (left) and validation loss (right) over 5000 pretraining iterations from random initialization. Green: Muon; blue: Spectral-GD.

suppose full-batch, Spectral-GD uses a step length of $\eta = \Theta\left(\frac{\sqrt{\lambda_0 \varepsilon}}{n}\right)$. *Then after* $T = O\left(\frac{n^{3/2}}{\lambda_0 \sqrt{\varepsilon \delta}}\right)$ *updates the empirical loss remains* $O(\epsilon)$ *with probability at least* $1 - \delta$ *over initialization.*

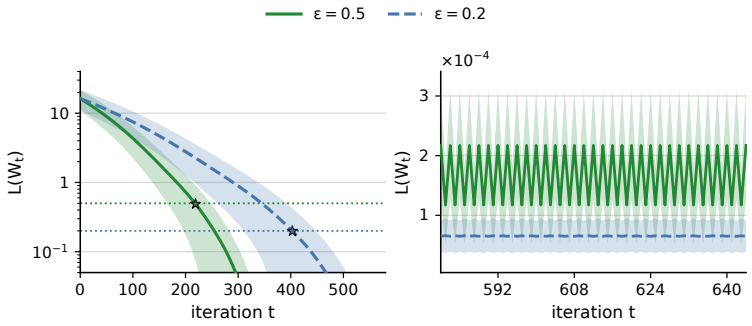


Figure 2: Mean training loss for $\varepsilon = 0.5$ (green) and $\varepsilon = 0.2$ (blue). The band is the pointwise ± 1 standard deviation over five initializations, and the stars mark the first time the mean reaches the target. Left: the descent of $L(\boldsymbol{W}_t)$ against iteration t, on a log axis. Right: the oscillation in loss during the later iterations.

Comparison with Standard Depth-two NTK Guarantees. In the same setup as above in Du et al. (2018b) the corresponding GD result required a sufficient width $m = \Omega(n^6/(\lambda_0^4\delta^3))$, and attains a geometric rate of decay of the loss. Arora et al. (2019b) analyzed generalization bounds along the training trajectory in the setup of Du et al. (2018b) and used and even higher value of m, scaling more steeply with n. In further development, Wu et al. (2019) improved the required learning rate while using the same width as in Du et al. (2018b).

In comparison to the above benchmarks, our result for Spectral-GD requires lower width scaling as $n^5/(\lambda_0^4\delta^3)$, while the convergence rate become slow as $O\big(1/\sqrt{\varepsilon}\big)$ instead of geometric.

An empirical illustration of the result. We train a two-layer ReLU student on n unit-norm Gaussian inputs in $\mathbb{R}^d$ whose labels come from a fixed random ReLU net of the same architecture. Only the first layer is trained, by full-batch Spectral-GD on the loss equation 3, from an initialization satisfying Assumption 1. We take $n = 20$, $d = 100$, $m = 1000$, $\alpha = 1$ and $\delta = 0.05$ and this dataset has $B \approx 1.68$ and $\lambda_0 \approx 0.32$, so equation 8 fixes η and T for each target ε before training starts. The pair of runs implemented are,

ε	η	T	first crossing	$\max_{t \geq t^\star} L(\boldsymbol{W}_t)$	$L(\boldsymbol{W}_T)$
0.500	1.41×10^{-2}	31645	210 ± 44	0.492	3.90×10^{-5}
0.200	8.94×10^{-3}	50035	377 ± 79	0.199	3.65×10^{-5}

Table 1: First crossing is the mean ± 1 s.d. over five initializations for the loss to cross the set ε threshold. The last two columns are from the full-budget run through T, on which the loss stays below ε after the hit.

The experimental results are shown in Figure 2. In each case we see that after the loss has fallen to cross the accuracy threshold set Spectral-GD enters an oscillation phase reflecting what we see in the proof of Theorem 3.1 that the error is not forever decreasing but is bounded in a small range after the corresponding threshold T number of iterations.

Organization. Section 2 fixes the model, the selected ReLU derivative, and the two kernel definitions, and states the two assumptions. Section 3 states the formal guarantee, Theorem 3.1, together with the three lemmas it rests on; Section 4.1 proves the theorem. Appendix A derives the two matrix characterizations of the polar update, as a linear minimization oracle over a spectral-norm ball and as a nuclear-norm subgradient;Appendix B proves the matrix identities used in Appendix A; Appendix C derives the closed form of the infinite-width kernel $\boldsymbol{H}^\infty$; and Appendix D proves the three lemmas of Section 3. Appendix E gives a brief survey of the NTK literature.

2 MATHEMATICAL SETUP

Notation. For vectors, $\|\cdot\|_2$ denotes the Euclidean norm. For a matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_2$ is its operator norm, $\|\boldsymbol{A}\|_F^2 = \sum_{i,j}[\boldsymbol{A}]_{ij}^2$ is its squared Frobenius norm, and $\|\boldsymbol{A}\|_* = \sum_j \sigma_j(\boldsymbol{A})$ is its nuclear norm. We use column-wise vectorization and the Frobenius inner product $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B}) = \mathrm{vec}(\boldsymbol{A})^\top \mathrm{vec}(\boldsymbol{B})$, so $\|\mathrm{vec}(\boldsymbol{A})\|_2 = \|\boldsymbol{A}\|_F$. We write $[n] = \{1, \dots, n\}$.

Model and loss. Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be the training data with $\|\boldsymbol{x}_i\|_2 = 1$ and define $B := \max_i |y_i| \geq 0$. The predictor is a two-layer ReLU network

$$f(\boldsymbol{x}; \boldsymbol{W}) = \frac{1}{\sqrt{m}} \sum_{r=1}^m a_r \, \sigma(\boldsymbol{w}_r^\top \boldsymbol{x}), \qquad a_r \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{\pm\alpha\}, \qquad \sigma(t) = \max\{0, t\}, \tag{2}$$

where $\boldsymbol{W} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_m) \in \mathbb{R}^{d \times m}$, and the second-layer signs a_r are fixed. The prediction vector $\boldsymbol{u}(\boldsymbol{W}) \in \mathbb{R}^n$ and the residual vector $\boldsymbol{e}(\boldsymbol{W}) \in \mathbb{R}^n$ are defined by $[\boldsymbol{u}(\boldsymbol{W})]_i := f(\boldsymbol{x}_i; \boldsymbol{W})$ and $[\boldsymbol{e}(\boldsymbol{W})]_i = y_i - [\boldsymbol{u}(\boldsymbol{W})]_i$, and hence the training loss is

$$L(\boldsymbol{W}) \triangleq \tfrac{1}{2}\|\boldsymbol{u}(\boldsymbol{W}) - \boldsymbol{y}\|_2^2. \tag{3}$$

Corresponding to the t-th iterate of the algorithm we define $\boldsymbol{u}_t := \boldsymbol{u}(\boldsymbol{W}_t)$, $\boldsymbol{e}_t := \boldsymbol{e}(\boldsymbol{W}_t)$, and $\Delta \boldsymbol{u}_t := \boldsymbol{u}(\boldsymbol{W}_{t+1}) - \boldsymbol{u}(\boldsymbol{W}_t)$.

Selected ReLU Derivative. We fix $\sigma'(t) := \mathbb{1}\{t \geq 0\}$, including at activation boundaries. At such points, $\boldsymbol{G}_t$ denotes the loss-gradient matrix selected by this convention. We do not assume that data-dependent iterates avoid boundaries; the remainder bound covers them explicitly. The r-th column of the selected loss gradient in equation 1 is

$$[\boldsymbol{G}_t]_{\cdot,r} = -\frac{1}{\sqrt{m}} \sum_{i=1}^n [\boldsymbol{e}_t]_i \, a_r \, \mathbb{1}_{\{\boldsymbol{w}_r(t)^\top \boldsymbol{x}_i \geq 0\}} \, \boldsymbol{x}_i, \qquad r \in [m]. \tag{4}$$

2.1 NTK DEFINITIONS

Definition 1 (Tangent Feature Matrix and Empirical NTK)**.** *For $\boldsymbol{W} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_m)$, define the tangent feature matrix $\boldsymbol{J}(\boldsymbol{W}) \in \mathbb{R}^{n \times dm}$ blockwise as,*

$$[\boldsymbol{J}(\boldsymbol{W})]_{i,(r-1)d+1:rd} = \frac{a_r}{\sqrt{m}} \mathbb{1}\{\boldsymbol{w}_r^\top \boldsymbol{x}_i \geq 0\} \boldsymbol{x}_i^\top, \qquad i \in [n], \; r \in [m].$$

At an iterate $\boldsymbol{W}_t$, the selected gradient in equation 4 therefore satisfies

$$\mathrm{vec}(\boldsymbol{G}_t) = -\boldsymbol{J}(\boldsymbol{W}_t)^\top \boldsymbol{e}_t.$$

The empirical neural tangent kernel is,

$$\boldsymbol{H}(\boldsymbol{W}) := \boldsymbol{J}(\boldsymbol{W}) \boldsymbol{J}(\boldsymbol{W})^\top \in \mathbb{R}^{n \times n}.$$

Writing $\boldsymbol{H}_t := \boldsymbol{H}(\boldsymbol{W}_t)$, its entries are

$$[\boldsymbol{H}_t]_{ij} = \frac{\alpha^2}{m} \sum_{r=1}^m \boldsymbol{x}_i^\top \boldsymbol{x}_j \, \mathbb{1}\{\boldsymbol{w}_r(t)^\top \boldsymbol{x}_i \geq 0\} \, \mathbb{1}\{\boldsymbol{w}_r(t)^\top \boldsymbol{x}_j \geq 0\}, \qquad i, j \in [n], \tag{5}$$

where we used $a_r^2 = \alpha^2$.

Definition 2 (Normalized Infinite-Width NTK)**.** *Under the Gaussian initialization in Assumption 1, define $\boldsymbol{H}^\infty \in \mathbb{R}^{n \times n}$ entrywise by*

$$H_{ij}^\infty := \mathbb{E}_{\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)}\left[\boldsymbol{x}_i^\top \boldsymbol{x}_j \, \mathbb{1}\{\boldsymbol{w}^\top \boldsymbol{x}_i \geq 0, \; \boldsymbol{w}^\top \boldsymbol{x}_j \geq 0\}\right] = \boldsymbol{x}_i^\top \boldsymbol{x}_j \, \frac{\pi - \arccos\left(\boldsymbol{x}_i^\top \boldsymbol{x}_j\right)}{2\pi}, \qquad i, j \in [n]. \tag{6}$$

Note that $\mathbb{E}[\boldsymbol{H}(\boldsymbol{W}_0)] = \alpha^2 \boldsymbol{H}^\infty$.

The kernel formula is derived in Appendix C.

Remark. *Since $\|\boldsymbol{x}_i\|_2 = 1$, we have $H_{ii}^\infty = 1/2$ for every $i \in [n]$. Consequently, $\mathrm{tr}(\boldsymbol{H}^\infty) = n/2$ and*

$$\lambda_0 = \lambda_{\min}(\boldsymbol{H}^\infty) \leq \frac{1}{n} \mathrm{tr}(\boldsymbol{H}^\infty) = \frac{1}{2}.$$

2.2 ASSUMPTIONS

We assume a random initialization independent of the training data and a positive definite infinite-width kernel.

Assumption 1 (Distribution of the Second-Layer Weights and Initialisation for the First Layer)**.** *The training data are fixed independently of the network initialization (or are sampled independently of it). The second-layer weights are sampled as* $a_r \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{\pm\alpha\}$ *and are independent of the initialization of the trainable weights, which are sampled as* $\boldsymbol{w}_r(0) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d), r = 1, \dots, m.$

Assumption 2 (Positive Definite Infinite-Width Kernel)**.** *Let* $\boldsymbol{H}^\infty$ *be as in Definition 2. Assume* $\lambda_{\min}(\boldsymbol{H}^\infty) = \lambda_0 > 0.$

A standard sufficient condition for the above is that no two inputs are parallel (Du et al., 2018b, Theorem 3.1); for unit-norm inputs this is $|\boldsymbol{x}_i^\top \boldsymbol{x}_j| < 1$ for every $i \neq j$.

3 MAIN RESULT: FINITE-BUDGET ACCURACY OF SPECTRAL-GD

Theorem 3.1 (Guarantee on Training a Wide Net by Spectral-GD)**.** *Let the training data be fixed independently of initialization, with* $\|\boldsymbol{x}_i\|_2 = 1$*, and let* $B \geq \max_i |y_i|$*. From Assumptions 1 and 2 recall* $\alpha > 0$ *and* $\lambda_0 = \lambda_{\min}(\boldsymbol{H}^\infty) > 0$*. For arbitrary* $\epsilon, \delta \in (0, 1)$ *let the width of the net satisfy,*

$$m \geq \max\left\{ \frac{8n^2}{\lambda_0^2} \log \frac{6n^2}{\delta}, \quad \Omega\left(\frac{n^5 (B^2 + \alpha^2/2)}{\alpha^2 \delta^3 \lambda_0^4} \right) \right\}. \tag{7}$$

Further choose,

$$\eta = \min\left\{ \frac{\sqrt{2\lambda_0 \varepsilon}}{2n\alpha}, \quad \frac{1}{\alpha\sqrt{\lambda_0}} \sqrt{\frac{3n(B^2 + \alpha^2/2)}{\delta}} \right\} \;\&\; T = \left\lceil \frac{4}{\alpha\sqrt{\lambda_0}\,\eta} \sqrt{\frac{3n(B^2 + \alpha^2/2)}{\delta}} \right\rceil. \tag{8}$$

Then, for updates $\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta\, \mathrm{Polar}(\boldsymbol{G}_t)$*, with probability at least* $1 - \delta$ *over initialization, for all* $t \geq T$*,*

$$L(\boldsymbol{W}_t) = \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_t)\|_2^2 \leq \varepsilon. \tag{9}$$

4 INTERMEDIATE LEMMAS

We use the following three lemmas to prove Theorem 3.1.

Lemma 4.1 (Initialization)**.** *In the setup of Theorem 3.1 we define the events,*

$$\mathcal{E}_1 := \left\{ \|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_0)\|_2 \leq \sqrt{\frac{3n(B^2 + \alpha^2/2)}{\delta}} \right\}, \tag{10}$$

$$\mathcal{E}_2 := \left\{ \|\boldsymbol{H}(\boldsymbol{W}_0) - \alpha^2 \boldsymbol{H}^\infty\|_2 \leq \frac{\alpha^2 \lambda_0}{4} \right\}, \tag{11}$$

$$\mathcal{E}_3 := \left\{ \frac{1}{m} \sum_{i=1}^{n} \sum_{r=1}^{m} \mathbb{1}\left\{ |\boldsymbol{w}_r(0)^\top \boldsymbol{x}_i| \leq \frac{\delta \lambda_0}{1536\sqrt{2/\pi}\, n} \right\} \leq \frac{\lambda_0}{512} \right\}. \tag{12}$$

Then

$$\mathbb{P}(\mathcal{E}_1^c) \leq \frac{\delta}{3}, \qquad \mathbb{P}(\mathcal{E}_2^c) \leq \frac{\delta}{3}, \qquad \mathbb{P}(\mathcal{E}_3^c) \leq \frac{\delta}{3}.$$

The proof is in Section D.1.

Lemma 4.2 (Uniform Kernel and Remainder Bounds)**.** *Fix* $\delta \in (0, 1)$ *and retain the hypotheses and events of Lemma 4.1. Let* $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ *and*

$$\mathcal{B} := \left\{ \boldsymbol{Z} \in \mathbb{R}^{d \times m} : \|\boldsymbol{Z} - \boldsymbol{W}_0\|_F \leq \frac{5n}{\alpha} \sqrt{\frac{3(B^2 + \alpha^2/2)}{\delta \lambda_0}} \right\}. \tag{13}$$

For $\boldsymbol{Z} = [\boldsymbol{z}_1, \dots, \boldsymbol{z}_m]$, write $g_{ir}(\boldsymbol{Z}) := \mathbb{1}\{\boldsymbol{z}_r^\top \boldsymbol{x}_i \geq 0\}$. For $\boldsymbol{W}, \boldsymbol{V} \in \mathcal{B}$, define

$$\mathcal{A}_i(\boldsymbol{W}, \boldsymbol{V}) := \{r \in [m] : g_{ir}(\boldsymbol{W}) \neq g_{ir}(\boldsymbol{W}_0) \text{ or } g_{ir}(\boldsymbol{V}) \neq g_{ir}(\boldsymbol{W}_0)\}.$$

On $\mathcal{E}$, the following bounds hold simultaneously for every $\boldsymbol{W}, \boldsymbol{V} \in \mathcal{B}$:

$$\frac{1}{m}\sum_{i=1}^{n} |\mathcal{A}_i(\boldsymbol{W}, \boldsymbol{V})| \leq \frac{\lambda_0}{256}, \tag{14}$$

$$\|\boldsymbol{J}(\boldsymbol{W}) - \boldsymbol{J}(\boldsymbol{W}_0)\|_F \leq \frac{\alpha\sqrt{\lambda_0}}{16}, \qquad \lambda_{\min}(\boldsymbol{H}(\boldsymbol{W})) \geq \frac{\alpha^2\lambda_0}{2}, \tag{15}$$

and

$$\|\boldsymbol{u}(\boldsymbol{V}) - \boldsymbol{u}(\boldsymbol{W}) - \boldsymbol{J}(\boldsymbol{W})\,\mathrm{vec}(\boldsymbol{V} - \boldsymbol{W})\|_2 \leq \frac{\alpha\sqrt{\lambda_0}}{8}\|\boldsymbol{V} - \boldsymbol{W}\|_2. \tag{16}$$

The proof is in Section D.2.

Lemma 4.3 (Spectral Steps and Residual Descent)**.** *Let $\boldsymbol{W} \in \mathbb{R}^{d\times m}$ and $\eta \geq 0$. Let $\boldsymbol{G}$ be the selected loss-gradient matrix at $\boldsymbol{W}$, so that $\mathrm{vec}(\boldsymbol{G}) = -\boldsymbol{J}(\boldsymbol{W})^\top(\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}))$, and set $\boldsymbol{V} := \boldsymbol{W} - \eta\,\mathrm{Polar}(\boldsymbol{G})$, where $\mathrm{Polar}(0) = 0$. Then*

$$\|\boldsymbol{V} - \boldsymbol{W}\|_2 \leq \eta, \qquad \|\boldsymbol{V} - \boldsymbol{W}\|_F \leq \sqrt{n}\,\eta, \qquad \|\boldsymbol{u}(\boldsymbol{V}) - \boldsymbol{u}(\boldsymbol{W})\|_2^2 \leq n\alpha^2\eta^2. \tag{17}$$

On the event $\mathcal{E}$ of Lemma 4.2, if $\boldsymbol{W}, \boldsymbol{V} \in \mathcal{B}$, then

$$\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{V})\|_2^2 \leq \|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W})\|_2^2 - \alpha\eta\sqrt{\lambda_0}\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W})\|_2 + n\alpha^2\eta^2. \tag{18}$$

If, in addition, $\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W})\|_2 > 0$ and $\eta \leq \sqrt{\lambda_0}\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W})\|_2/(2n\alpha)$, then

$$\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{V})\|_2 \leq \|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W})\|_2 - \frac{\alpha\eta\sqrt{\lambda_0}}{4}. \tag{19}$$

The proof is in Section D.3.

4.1 Proof of the finite-budget guarantee

We define, $R_0 := \sqrt{\frac{3n(B^2+\alpha^2/2)}{\delta}}$ to rewrite, the the following three key quantities as,

$$\eta = \min\left\{\frac{\sqrt{2\lambda_0\varepsilon}}{2n\alpha},\ \frac{R_0}{\alpha\sqrt{\lambda_0}}\right\}, \quad T = \left\lceil \max\left\{4,\ \frac{8nR_0}{\lambda_0\sqrt{2\varepsilon}}\right\}\right\rceil, \quad \mathcal{B} = \left\{\boldsymbol{Z} : \|\boldsymbol{Z} - \boldsymbol{W}_0\|_F \leq \frac{5\sqrt{n}\,R_0}{\alpha\sqrt{\lambda_0}}\right\}. \tag{20}$$

The argument is organised into five steps. Step 1 reduces the theorem to a statement about the residual norm. Step 2 shows that the good event $\mathcal{E}$ has probability at least $1 - \delta$. Step 3 rewrites the budget T in the form in which it is actually used, shows that *every* iterate up to time T stays inside the ball $\mathcal{B}$, and, on $\mathcal{E}$, applies Lemma 4.3 along the trajectory to get both the general descent recursion and the sharper per-step decrease used below. Step 4 shows that the target residual is reached at some time at most T, and Step 5 shows that once reached it is never lost, so that in particular it holds at time T itself.

Proof of Theorem 3.1. **Step 1:** We claim that it suffices to prove the implication

$$\mathcal{E} \implies \|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_T)\|_2 \leq \sqrt{2\varepsilon}, \tag{21}$$

together with $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$.

Recalling the definition of L from equation 3 we observe that Eq 21 implies, $L(\boldsymbol{W}_T) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_T)\|_2^2 \leq \varepsilon$.

Step 2: Recalling, $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, from Lemma 4.2 and invoking Lemma 4.1 and by union-bounding

$$\mathbb{P}(\mathcal{E}^c) = \mathbb{P}\left(\bigcup_{j=1}^{3} \mathcal{E}_j^c\right) \leq \sum_{j=1}^{3} \mathbb{P}(\mathcal{E}_j^c) \leq 3 \cdot \frac{\delta}{3} = \delta, \qquad \text{so} \qquad \mathbb{P}(\mathcal{E}) \geq 1 - \delta. \tag{22}$$

Step 3: We observe that η inequation 20 can be further re-written as,

$$\frac{1}{\eta} = \max\left\{\frac{2n\alpha}{\sqrt{2\lambda_0\varepsilon}}, \quad \frac{\alpha\sqrt{\lambda_0}}{R_0}\right\}. \tag{23}$$

Let,

$$x := \frac{4R_0}{\alpha\sqrt{\lambda_0}} \cdot \frac{1}{\eta} = \max\left\{\frac{8nR_0}{\lambda_0\sqrt{2\varepsilon}},\ 4\right\}, \qquad \text{hence} \qquad T = \lceil x \rceil, \qquad x \le T \le x+1. \tag{24}$$

Hence, we observe that,

$$T\eta \le (x+1)\eta = x\eta + \eta = \frac{4R_0}{\alpha\sqrt{\lambda_0}} + \eta \le \frac{4R_0}{\alpha\sqrt{\lambda_0}} + \frac{R_0}{\alpha\sqrt{\lambda_0}} = \frac{5R_0}{\alpha\sqrt{\lambda_0}} \qquad [\eta \le R_0/(\alpha\sqrt{\lambda_0})]. \tag{25}$$

We calculate for every $0 \le t \le T$,

$$\begin{aligned} \|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F &\le \sum_{j=0}^{t-1} \|\boldsymbol{W}_{j+1} - \boldsymbol{W}_j\|_F \le \sqrt{n}\, t\, \eta && \text{[equation 17]} \\ &\le \sqrt{n}\, T\eta \le \frac{5\sqrt{n}\, R_0}{\alpha\sqrt{\lambda_0}} && \text{[equation 25].} \end{aligned} \tag{26}$$

The above inequality holds trivially for $t = 0$. By equation 20 the right-hand side of equation 26 is precisely the radius of $\mathcal{B}$ in equation 13, and therefore

$$\boldsymbol{W}_t \in \mathcal{B} \qquad \text{for every } 0 \le t \le T. \tag{27}$$

Every iterate up to time T is in $\mathcal{B}$, so for each $t < T$ both $\boldsymbol{W}_t$ and $\boldsymbol{W}_{t+1}$ are in $\mathcal{B}$. On $\mathcal{E}$ we can therefore apply equation 18 of Lemma 4.3 to every update, which gives

$$\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_{t+1})\|_2^2 \le \|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_t)\|_2^2 - \alpha\eta\sqrt{\lambda_0}\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_t)\|_2 + n\alpha^2\eta^2, \qquad 0 \le t < T. \tag{28}$$

Take any $t < T$ whose residual is still above the target, i.e. $\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_t)\|_2 > \sqrt{2\varepsilon}$. The definition of η in equation 20 gives

$$\eta \le \frac{\sqrt{2\lambda_0\varepsilon}}{2n\alpha} = \frac{\sqrt{\lambda_0}\sqrt{2\varepsilon}}{2n\alpha} < \frac{\sqrt{\lambda_0}\,\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_t)\|_2}{2n\alpha},$$

which is exactly the step-size condition in equation 19 of Lemma 4.3.

Hence it follows from Lemma 4.3,

$$\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_{t+1})\|_2 \le \|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_t)\|_2 - \frac{\alpha\eta\sqrt{\lambda_0}}{4}. \tag{29}$$

Step 4: Fix an initialization for which $\mathcal{E}$ holds, and suppose towards a contradiction that

$$\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_t)\|_2 > \sqrt{2\varepsilon} \qquad \text{for every } 0 \le t \le T-1. \tag{30}$$

Then every $t < T$ has a residual above the target, so equation 29 applies at all T updates, i.e each update lowers the residual by at least $\alpha\eta\sqrt{\lambda_0}/4$. Telescoping on equation 29

$$\begin{aligned} \|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_T)\|_2 &\le \|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_0)\|_2 - \sum_{t=0}^{T-1} \frac{\alpha\eta\sqrt{\lambda_0}}{4} \\ &= \|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_0)\|_2 - \frac{T\alpha\eta\sqrt{\lambda_0}}{4}. \end{aligned} \tag{31}$$

Now we check that T is large enough to wipe out the initial residual. Since $\mathcal{E}1$ holds, We use $\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_0)\|_2 \le R_0$ and $T \ge x$ from equation 24:

$$\begin{aligned} \|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_T)\|_2 &\le R_0 - \frac{\alpha\eta\sqrt{\lambda_0}}{4}\, T \\ &\le R_0 - \frac{\alpha\eta\sqrt{\lambda_0}}{4} \cdot \frac{4R_0}{\alpha\sqrt{\lambda_0}\,\eta} && [T \ge x] \\ &= R_0 - R_0 = 0. \end{aligned}$$

So, if equation 30 holds, the residual at time T would be 0, hence at most $\sqrt{2\varepsilon}$. Otherwise, equation 30 fails, so some $t \le T-1$ already satisfies $\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_t)\|_2 \le \sqrt{2\varepsilon}$. Either way, on $\mathcal{E}$ some iterate reaches the target, and $t^\star$ below is well-defined:

$$t^\star := \min\{t \in \{0, 1, \dots, T\} : \|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_t)\|_2 \le \sqrt{2\varepsilon}\} \tag{32}$$

Step 5: Let $t^\star \le t \le T$ with $\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_t)\|_2 \le \sqrt{2\varepsilon}$. We show $\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_{t+1})\|_2 \le \sqrt{2\varepsilon}$ as well. Repeating this up to $t = T$ gives equation 21 as claimed.

Put $s := \|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_t)\|_2 \in [0, \sqrt{2\varepsilon}]$ and $c := \alpha\eta\sqrt{\lambda_0}/2$. By equation 28,

$$\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_{t+1})\|_2^2 \le s^2 - 2cs + n\alpha^2\eta^2 = (s-c)^2 - c^2 + n\alpha^2\eta^2. \tag{33}$$

From η in equation 20, i.e. $\eta \le \sqrt{2\lambda_0\varepsilon}/(2n\alpha)$, together with $\lambda_0 \le \mathrm{tr}(\boldsymbol{H}^\infty)/n = 1/2$ (Definition 2), gives

$$c \le \frac{\lambda_0\sqrt{2\varepsilon}}{4n} \le \frac{\sqrt{2\varepsilon}}{2} \text{ (weakening the upper bound)}, \qquad n\alpha^2\eta^2 \le n\alpha^2\eta \cdot \frac{\sqrt{2\lambda_0\varepsilon}}{2n\alpha} = c\sqrt{2\varepsilon}. \tag{34}$$

Now consider the following cases,

Case (a) : $s \ge c$. Here $|s-c| = s - c$. Since $s \le \sqrt{2\varepsilon}$, this is at most $\sqrt{2\varepsilon} - c$.

Case (b) : $s < c$. Here $|s-c| = c - s \le c$, since $s \ge 0$. And $c \le \sqrt{2\varepsilon} - c$ is exactly $c \le \sqrt{2\varepsilon}/2$ from equation 34, doubled and rearranged.

$$s \ge c: \quad |s-c| = s - c \le \sqrt{2\varepsilon} - c \qquad [s \le \sqrt{2\varepsilon}] \tag{35}$$

$$s < c: \quad |s-c| = c - s \le c \le \sqrt{2\varepsilon} - c \qquad [s \ge 0; \text{ then } 2c \le \sqrt{2\varepsilon}] \tag{36}$$

So in both cases $|s-c| \le \sqrt{2\varepsilon} - c$. Hence substituting in equation 33 we have,

$$\begin{aligned} \|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_{t+1})\|_2^2 &\le (\sqrt{2\varepsilon} - c)^2 - c^2 + n\alpha^2\eta^2 && \text{[by equation 33]} \\ &= 2\varepsilon - 2c\sqrt{2\varepsilon} + n\alpha^2\eta^2 \\ &\le 2\varepsilon - c\sqrt{2\varepsilon} \le 2\varepsilon && \text{[by equation 34]}, \end{aligned} \tag{37}$$

so $\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_{t+1})\|_2 \le \sqrt{2\varepsilon}$.

By Step 1 this gives $L(\boldsymbol{W}_T) \le \varepsilon$ on $\mathcal{E}$, and $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$ by equation 22, which is equation 9. □

5 Conclusion

We have given the first non-asymptotic training guarantee for Spectral-GD that defines the Muon family. The proof given here is in the large width of regime of neural network and assume positive definiteness of the infinite tangent kernel. Hence this work can be considered a part of the NTK literature. Relative to the gradient-descent analysis of Du et al. (2018b, Theorem 4.1), the Spectral-GD update buys a lower width requirement, n^5 in place of n^6, at the cost of a slower $O(1/\sqrt{\varepsilon})$ rate instead of a geometric one — a trade-off that can possibly be attributed to discarding the gradient's magnitude in deciding the update direction. Two key extensions possible are to incorporate Muon's momentum and stochastic mini-batch gradients into the theory, bringing its analysis closer to practical large-scale training that uses Muon. It would also be interesting to extend this result into the regime of high depth architecture and for LLMs in particular.

AI USE STATEMENT

We used generative AI tools to assist with formulating mathematical claims, providing critical ingredients for proofs, code generation for implementing parts of our method, identifying related literature, and assisting in the writing of proofs. All AI-generated code used in this work was reviewed by the authors. We did not use generative AI tools for generating synthetic datasets, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, cleaning or reformatting datasets, supporting qualitative or thematic data analysis, or interpreting results. The authors take full responsibility for the final content of this work, including all text and code produced with AI assistance.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. *Learning and generalization in overparameterized neural networks, going beyond two layers*. Curran Associates Inc., Red Hook, NY, USA, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252. PMLR, 09–15 Jun 2019b. URL https://proceedings.mlr.press/v97/allen-zhu19a.html.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. *On exact computation with an infinitely wide neural net*. Curran Associates Inc., Red Hook, NY, USA, 2019a.

Sanjeev Arora, Simon Shaolei Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *ArXiv*, abs/1901.08584, 2019b. URL https://api.semanticscholar.org/CorpusID:59222746.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. *ArXiv*, abs/2410.21265, 2024. URL https://api.semanticscholar.org/CorpusID:273655185.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Trans. Mach. Learn. Res.*, 2026, 2025. URL https://api.semanticscholar.org/CorpusID:279447438.

Lénaïc Chizat, Edouard Oyallon, and Francis R. Bach. On lazy training in differentiable programming. In *Neural Information Processing Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:189928159.

Karl Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John D. Schulman. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021. URL https://api.semanticscholar.org/CorpusID:239998651.

Alec Dewulf, Dhruv Pai, Li Yang, Ashley Zhang, and Ben Keigwin. Aurora: A leverage-aware spectral optimizer. *arXiv preprint arXiv:2606.27715*, 2026.

Simon Du and Jason Lee. On the power of over-parametrization in neural networks with quadratic activation. In *International conference on machine learning*, pp. 1329–1338. PMLR, 2018.

Simon Shaolei Du, J. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *ArXiv*, abs/1811.03804, 2018a. URL https://api.semanticscholar.org/CorpusID:53250419.

Simon Shaolei Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *ArXiv*, abs/1810.02054, 2018b. URL https://api.semanticscholar.org/CorpusID:52920808.

Simon Shaolei Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *ArXiv*, abs/1810.02054, 2018c. URL https://api.semanticscholar.org/CorpusID:52920808.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Daniel Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *ArXiv*, abs/2402.01306, 2024. URL https://api.semanticscholar.org/CorpusID:289701527.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *ArXiv*, abs/1806.07572, 2018. URL https://api.semanticscholar.org/CorpusID:49321232.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

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/.

Tim Tsz-Kit Lau, Qi Long, and Weijie Su. Polargrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective. *arXiv preprint arXiv:2505.21799*, 2025.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Narain Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Journal of Statistical Mechanics: Theory and Experiment*, 2020, 2019. URL https://api.semanticscholar.org/CorpusID:62841516.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *ArXiv*, abs/1808.01204, 2018. URL https://api.semanticscholar.org/CorpusID:51920936.

Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. Normuon: Making muon more efficient and scalable. *arXiv preprint arXiv:2510.05491*, 2025.

Jingyuan Liu, Jianling Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yi-Bo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Meng Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for llm training. *ArXiv*, abs/2502.16982, 2025. URL https://api.semanticscholar.org/CorpusID:276575212.

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

Guilherme Penedo, Hynek Kydlí$ek, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37 : 30811 – –30849, 2024.

Thomas Pethick, Wanyun Xie, Roman Machácek, and Volkan Cevher. Optimistic dual averaging unifies modern optimizers. *ArXiv*, abs/2605.11172, 2026. URL https://api.semanticscholar.org/CorpusID:288253779.

Zakhar Shumaylov, Nathaël Da Costa, Peter Zaika, Bálint Mucsányi, Alex Massucco, Yoav Gelberg, Carola B. Schönlieb, Yarin Gal, and Philipp Hennig. Muon is not that special: Random or inverted spectra work just as well. *ArXiv*, abs/2605.11181, 2026. URL https://api.semanticscholar.org/CorpusID:288258167.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Xiaoxia Wu, Simon S Du, and Rachel Ward. Global convergence of adaptive gradient methods for an over-parameterized neural network. *arXiv preprint arXiv:1902.07111*, 2019.

An Yang, An-Feng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bo-Wen Yu, Chang Gao, Chengen Huang, Chen-Xu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jian-Hong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Jingren Zhou, Junyan Lin, Kai Dang, Ke-Qin Bao, Ke-Pei Yang, Le Yu, Li-Chun Deng, Mei Li, Min Xue, Ming-Ze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tian-Yi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Ying-Er Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhen-Ru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. 2025. URL https://api.semanticscholar.org/CorpusID:278602855.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. *Advances in neural information processing systems*, 32, 2019.

A Motivating and Defining Spectral-GD

We characterize the polar update as a linear minimization oracle and as a nuclear-norm subgradient evaluated at the gradient matrix.

Theorem A.1 (Linear minimization over a spectral-norm ball)**.** *Let $f : \mathbb{R}^{d\times m} \to \mathbb{R}$ be differentiable and fix $\eta > 0$. Write the compact SVD of its gradient at $\boldsymbol{W}_t$ as $\nabla f(\boldsymbol{W}_t) = \boldsymbol{U}_t \boldsymbol{\Sigma}_t \boldsymbol{V}_t^\top$, $\boldsymbol{U}_t \in \mathbb{R}^{d\times\rho_t}, \boldsymbol{\Sigma}_t \in \mathbb{R}^{\rho_t\times\rho_t}, \boldsymbol{V}_t \in \mathbb{R}^{m\times\rho_t}$, where $\rho_t = \operatorname{rank}(\nabla f(\boldsymbol{W}_t))$. Retain only positive singular values and set $\boldsymbol{U}_t\boldsymbol{V}_t^\top = 0$ when the gradient is zero. Then the compact-polar Spectral-GD update is a choice of minimizer, not necessarily the unique minimizer:*

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta \boldsymbol{U}_t\boldsymbol{V}_t^\top \in \operatorname*{arg\,min}_{\|\boldsymbol{W}-\boldsymbol{W}_t\|_2\le\eta} \langle \nabla f(\boldsymbol{W}_t), \boldsymbol{W} - \boldsymbol{W}_t\rangle. \tag{38}$$

Next we formally note an observation that appeared recently in Pethick et al. (2026).

Theorem A.2 (Spectral-GD as a Subgradient Step on the Nuclear Norm of the Gradient)**.** *Let $f : \mathbb{R}^{d\times m} \to \mathbb{R}$ be differentiable, fix $\eta > 0$, and let $\boldsymbol{G}_t := \nabla f(\boldsymbol{W}_t)$ have a SVD as $\boldsymbol{G}_t = \boldsymbol{U}_t\boldsymbol{\Sigma}_t\boldsymbol{V}_t^\top$. Use the compact SVD, set $\operatorname{Polar}(0) = 0$, and equip $\mathbb{R}^{d\times m}$ with the spectral norm $\|\cdot\|_2$. Its dual is the nuclear norm, $\|\boldsymbol{G}\|_* = \sup_{\|\boldsymbol{X}\|_2\le1}\langle \boldsymbol{G}, \boldsymbol{X}\rangle$. The claim below is specific to this norm pair, not an arbitrary norm. The Spectral-GD update satisfies*

$$\boldsymbol{W}_{t+1} - \boldsymbol{W}_t = -\eta\, \boldsymbol{U}_t\boldsymbol{V}_t^\top \in -\eta\, \partial\|\boldsymbol{G}_t\|_*, \tag{39}$$

The subgradient in Theorem A.2 is with respect to the *gradient matrix argument*, not the network weights. It does not say that Spectral-GD minimizes $\|\nabla f(\boldsymbol{W})\|_*$ as a new objective. Instead, $\langle \nabla f(\boldsymbol{W}), -\operatorname{Polar}(\nabla f(\boldsymbol{W}))\rangle = -\|\nabla f(\boldsymbol{W})\|_*$, so this is a descent direction for f whenever its gradient is nonzero. A finite step still requires control of the remainder, as in the finite-budget proof.

The proofs of both theorems are provided in Appendix B.

B Detailed matrix proofs for the spectral update

B.1 Proof of the spectral-ball minimization formula

Proof of Theorem A.1. Objective and minimizer. Write $\boldsymbol{G} = \nabla f(\boldsymbol{W}_t)$ and $\Delta\boldsymbol{W} = \boldsymbol{W} - \boldsymbol{W}_t$. The problem in the theorem is to minimize the scalar $\langle \boldsymbol{G}, \Delta\boldsymbol{W}\rangle$ subject to $\|\Delta\boldsymbol{W}\|_2 \le \eta$. We will prove both statements

$$\min_{\|\Delta\boldsymbol{W}\|_2\le\eta} \langle \boldsymbol{G}, \Delta\boldsymbol{W}\rangle = -\eta\|\boldsymbol{G}\|_*, \qquad -\eta\operatorname{Polar}(\boldsymbol{G}) \in \operatorname*{argmin}_{\|\Delta\boldsymbol{W}\|_2\le\eta} \langle \boldsymbol{G}, \Delta\boldsymbol{W}\rangle.$$

The first right-hand side is a scalar value; the second identifies one matrix attaining it. If $\boldsymbol{G} = 0$, every feasible matrix has objective value zero, and our convention chooses $\Delta\boldsymbol{W} = 0$. Hence assume $\boldsymbol{G} \ne 0$ below.

Lower bound. Let $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \sum_{j=1}^{\rho}\sigma_j u_j v_j^\top$ be its compact SVD, with $\sigma_j > 0$ and orthonormal columns in $\boldsymbol{U}, \boldsymbol{V}$. For any feasible $\Delta\boldsymbol{W}$,

$$\langle \boldsymbol{G}, \Delta\boldsymbol{W}\rangle = \sum_{j=1}^{\rho}\sigma_j u_j^\top \Delta\boldsymbol{W} v_j.$$

Since u_j, v_j are unit vectors, $|u_j^\top\Delta\boldsymbol{W}v_j| \le \|u_j\|_2\|\Delta\boldsymbol{W}v_j\|_2 \le \|\Delta\boldsymbol{W}\|_2 \le \eta$. Each summand is therefore bounded below by $-\eta\sigma_j$, and

$$\langle \boldsymbol{G}, \Delta\boldsymbol{W}\rangle \ge -\eta\sum_{j=1}^{\rho}\sigma_j = -\eta\|\boldsymbol{G}\|_*.$$

This lower bound applies to every feasible matrix, not just a matrix whose singular vectors agree with those of $\boldsymbol{G}$.

Attainment. Choose $\Delta\boldsymbol{W}_* = -\eta\boldsymbol{U}\boldsymbol{V}^\top$. Its operator norm is η, so it is feasible. Moreover,

$$u_j^\top\Delta\boldsymbol{W}_* v_j = -\eta u_j^\top\boldsymbol{U}\boldsymbol{V}^\top v_j = -\eta \quad (1 \le j \le \rho),$$

because the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal. Substitution into the objective gives $\langle \boldsymbol{G}, \Delta \boldsymbol{W}_* \rangle = -\eta \sum_j \sigma_j = -\eta \|\boldsymbol{G}\|_*$. The universal lower bound has been attained, proving both statements at the start of the proof. Substituting $\boldsymbol{W}_{t+1} = \boldsymbol{W}_t + \Delta \boldsymbol{W}_*$ gives the claimed spectral update.

Nonuniqueness. Knowing $\boldsymbol{U}^\top \Delta \boldsymbol{W} \boldsymbol{V} = -\eta \boldsymbol{I}_\rho$ controls only the part of $\Delta \boldsymbol{W}$ seen by these singular subspaces. It does not determine the entire matrix. For example, let $\boldsymbol{G} = \mathrm{diag}(1, 0)$, so $\boldsymbol{U} = \boldsymbol{V} = (1, 0)^\top$. Both

$$\Delta \boldsymbol{W}_* = \mathrm{diag}(-\eta, 0), \qquad \widetilde{\Delta \boldsymbol{W}} = \mathrm{diag}(-\eta, \eta/2)$$

have operator norm η, both satisfy $\boldsymbol{U}^\top \Delta \boldsymbol{W} \boldsymbol{V} = -\eta$, and both have objective value $-\eta$, but they are different matrices. Thus the minimizer need not be unique. □

B.2 The subgradient characterization for a norm

Theorem B.1 (Dual feasibility and alignment for a norm subgradient)**.** *Let $\mathcal{V}$ be a finite-dimensional real inner-product space with norm $\|\cdot\|$. Define its dual norm by*

$$\|G\|_* := \sup_{\|X\| \le 1} \langle G, X \rangle. \tag{40}$$

Then $G \in \partial \|X\|$ if and only if $\|G\|_* \le 1$ and $\langle G, X \rangle = \|X\|$.

Proof. Subgradient inequality. The notation $G \in \partial \|X\|$ means

$$\|Y\| \ge \|X\| + \langle G, Y - X \rangle \quad \text{for every } Y \in \mathcal{V}. \tag{41}$$

We prove each direction of the claimed equivalence from this definition.

Alignment. Assume equation 41. With $Y = 0$, it gives $0 \ge \|X\| - \langle G, X \rangle$, hence $\langle G, X \rangle \ge \|X\|$. With $Y = 2X$, positive homogeneity of a norm gives $2\|X\| \ge \|X\| + \langle G, X \rangle$, hence the reverse inequality. Together these prove $\langle G, X \rangle = \|X\|$. This calculation also holds at $X = 0$, where both sides are zero.

Dual feasibility. Insert alignment back into equation 41. The constant terms cancel, leaving $\|Y\| \ge \langle G, Y \rangle$ for every Y. In particular, if $\|Y\| \le 1$, then $\langle G, Y \rangle \le 1$. Taking the supremum over that unit ball proves $\|G\|_* \le 1$.

Converse. Now suppose $\|G\|_* \le 1$ and $\langle G, X \rangle = \|X\|$. For any nonzero Y, the vector $Y/\|Y\|$ is in the unit ball, so the definition of the dual norm gives $\langle G, Y \rangle \le \|G\|_* \|Y\| \le \|Y\|$. For $Y = 0$ the same inequality is immediate. Therefore

$$\|Y\| \ge \langle G, Y \rangle = \langle G, X \rangle + \langle G, Y - X \rangle = \|X\| + \langle G, Y - X \rangle.$$

This is the subgradient inequality for every Y, completing the proof. □

B.3 Why spectral and nuclear norms are dual

Proposition 1 (Spectral–nuclear norm duality)**.** *On $\mathbb{R}^{d \times m}$, use $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B})$. For every matrix $\boldsymbol{G}$,*

$$\sup_{\|\boldsymbol{X}\|_2 \le 1} \langle \boldsymbol{G}, \boldsymbol{X} \rangle = \|\boldsymbol{G}\|_*, \qquad \sup_{\|\boldsymbol{X}\|_* \le 1} \langle \boldsymbol{G}, \boldsymbol{X} \rangle = \|\boldsymbol{G}\|_2.$$

Here the subscript $*$ denotes the nuclear norm, not an additional multiplication symbol.

Proof. Zero matrix and compact SVD. If $\boldsymbol{G} = 0$, both objective functions vanish for every feasible $\boldsymbol{X}$, and both norms of $\boldsymbol{G}$ are zero. Both equalities follow, with witness $\boldsymbol{X} = 0$. Hence suppose $\boldsymbol{G} \ne 0$ and use its compact SVD $\boldsymbol{G} = \sum_{j=1}^{\rho} \sigma_j u_j v_j^\top = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^\top$, where $\sigma_1 \ge \cdots \ge \sigma_\rho > 0$.

Spectral unit ball. If $\|\boldsymbol{X}\|_2 \le 1$, then $|u_j^\top \boldsymbol{X} v_j| \le 1$ for every j, by Cauchy–Schwarz and the definition of the operator norm. Thus

$$\langle \boldsymbol{G}, \boldsymbol{X} \rangle = \sum_j \sigma_j u_j^\top \boldsymbol{X} v_j \le \sum_j \sigma_j = \|\boldsymbol{G}\|_*.$$

Choose $\boldsymbol{X}_* = \boldsymbol{U} \boldsymbol{V}^\top$, which has operator norm one. For this choice, $u_j^\top \boldsymbol{X}_* v_j = 1$ for every retained singular direction, so $\langle \boldsymbol{G}, \boldsymbol{X}_* \rangle = \sum_j \sigma_j = \|\boldsymbol{G}\|_*$. This proves the first supremum identity. The

nonzero singular values of $\boldsymbol{X}_*$ are all one, but it also has zero singular values when its rank is less than $\min\{d, m\}$.

Nuclear unit ball. Take any matrix $\boldsymbol{X}$ with $\|\boldsymbol{X}\|_* \le 1$ and write a compact SVD $\boldsymbol{X} = \sum_{k=1}^{q} \gamma_k p_k z_k^\top$, with $\gamma_k > 0$. If $\boldsymbol{X} = 0$, the desired upper bound is immediate; otherwise,

$$\begin{aligned}\langle \boldsymbol{G}, \boldsymbol{X} \rangle &= \sum_k \gamma_k p_k^\top \boldsymbol{G} z_k \\ &\le \sum_k \gamma_k |p_k^\top \boldsymbol{G} z_k| \le \sum_k \gamma_k \|\boldsymbol{G}\|_2 = \|\boldsymbol{X}\|_* \|\boldsymbol{G}\|_2 \le \|\boldsymbol{G}\|_2.\end{aligned}$$

The unit lengths of p_k, z_k justify the second inequality. For equality, choose $\boldsymbol{X}_* = u_1 v_1^\top$. It has dimensions $d \times m$, a single nonzero singular value equal to one, and hence nuclear norm one. Since $\boldsymbol{G} v_1 = \sigma_1 u_1$,

$$\langle \boldsymbol{G}, u_1 v_1^\top \rangle = \operatorname{tr}(\boldsymbol{G}^\top u_1 v_1^\top) = v_1^\top \boldsymbol{G}^\top u_1 = u_1^\top \boldsymbol{G} v_1 = \sigma_1 = \|\boldsymbol{G}\|_2.$$

The order of the two singular vectors matters: $v_1 u_1^\top$ would be an $m \times d$ matrix and is not the required witness. This proves the second supremum identity. □

B.4 The compact polar factor is a nuclear-norm subgradient

Theorem B.2 (Subgradient of the nuclear norm)**.** *For a compact SVD $\boldsymbol{X} = \boldsymbol{U}_1 \boldsymbol{\Sigma}_1 \boldsymbol{V}_1^\top$ retaining only positive singular values, $\boldsymbol{U}_1 \boldsymbol{V}_1^\top \in \partial \|\boldsymbol{X}\|_*$. At $\boldsymbol{X} = 0$, zero is a valid subgradient.*

Proof. Dual feasibility. For $\boldsymbol{X} \neq 0$, set $\boldsymbol{Q} = \boldsymbol{U}_1 \boldsymbol{V}_1^\top$. Orthogonality of the retained columns gives $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{V}_1 \boldsymbol{V}_1^\top$, an orthogonal projection. Hence $\|\boldsymbol{Q}\|_2 = 1$. At $\boldsymbol{X} = 0$, set $\boldsymbol{Q} = 0$, with $\|\boldsymbol{Q}\|_2 = 0$. In either case the spectral norm is at most one. By Proposition 1, this is precisely dual feasibility for a subgradient of the nuclear norm.

Alignment. For $\boldsymbol{X} \neq 0$, the compact SVD and cyclicity of the trace yield

$$\begin{aligned}\langle \boldsymbol{Q}, \boldsymbol{X} \rangle &= \operatorname{tr}\big((\boldsymbol{U}_1 \boldsymbol{V}_1^\top)^\top \boldsymbol{U}_1 \boldsymbol{\Sigma}_1 \boldsymbol{V}_1^\top\big) \\ &= \operatorname{tr}(\boldsymbol{V}_1 \boldsymbol{U}_1^\top \boldsymbol{U}_1 \boldsymbol{\Sigma}_1 \boldsymbol{V}_1^\top) = \operatorname{tr}(\boldsymbol{V}_1 \boldsymbol{\Sigma}_1 \boldsymbol{V}_1^\top) \\ &= \operatorname{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{V}_1^\top \boldsymbol{V}_1) = \operatorname{tr}(\boldsymbol{\Sigma}_1) = \|\boldsymbol{X}\|_*.\end{aligned}$$

At $\boldsymbol{X} = 0$ and $\boldsymbol{Q} = 0$, alignment is the equality $0 = 0$.

Subgradient inequality. For any $\boldsymbol{Y} \in \mathbb{R}^{d \times m}$, duality and the preceding feasibility and alignment identities give

$$\|\boldsymbol{Y}\|_* \ge \langle \boldsymbol{Q}, \boldsymbol{Y} \rangle = \langle \boldsymbol{Q}, \boldsymbol{X} \rangle + \langle \boldsymbol{Q}, \boldsymbol{Y} - \boldsymbol{X} \rangle = \|\boldsymbol{X}\|_* + \langle \boldsymbol{Q}, \boldsymbol{Y} - \boldsymbol{X} \rangle.$$

This is the definition of $\boldsymbol{Q} \in \partial \|\boldsymbol{X}\|_*$, equivalently the characterization in Theorem B.1. Using compact factors makes every product dimensionally valid even for rectangular matrices, and avoids adding arbitrary directions associated with zero singular values. □

B.5 Applying the subgradient fact to the gradient matrix

Proof of Theorem A.2. Fix the current network weights and let $\boldsymbol{G}_t = \nabla f(\boldsymbol{W}_t)$. Apply Theorem B.2 to the matrix $\boldsymbol{X} = \boldsymbol{G}_t$, not to the weights $\boldsymbol{W}_t$. For the compact SVD $\boldsymbol{G}_t = \boldsymbol{U}_t \boldsymbol{\Sigma}_t \boldsymbol{V}_t^\top$, it gives $\boldsymbol{Q}_t = \boldsymbol{U}_t \boldsymbol{V}_t^\top \in \partial \|\boldsymbol{G}_t\|_*$; if $\boldsymbol{G}_t = 0$, use $\boldsymbol{Q}_t = 0$. Multiplication of this set-membership statement by $-\eta$ gives $-\eta \boldsymbol{Q}_t \in -\eta \partial \|\boldsymbol{G}_t\|_*$. The spectral update is $\boldsymbol{W}_{t+1} - \boldsymbol{W}_t = -\eta \boldsymbol{Q}_t$ by definition, so it has exactly the membership stated in the theorem.

This subgradient is taken with respect to the argument of the function $\boldsymbol{X} \mapsto \|\boldsymbol{X}\|_*$. It is not a derivative with respect to $\boldsymbol{W}$ of the composite function $\|\nabla f(\boldsymbol{W})\|_*$. To identify the actual descent direction, compute instead

$$\langle \nabla f(\boldsymbol{W}_t), -\boldsymbol{Q}_t \rangle = -\|\boldsymbol{G}_t\|_*.$$

When $\boldsymbol{G}_t \neq 0$, the right-hand side is strictly negative. Differentiability of f then gives, as $h \downarrow 0$,

$$f(\boldsymbol{W}_t - h \boldsymbol{Q}_t) = f(\boldsymbol{W}_t) - h \|\boldsymbol{G}_t\|_* + o(h).$$

For sufficiently small positive h, the negative linear term dominates the remainder, so the direction decreases f. A prescribed finite step still requires a bound on that remainder; the finite-budget lemmas provide such a bound for the ReLU training problem using its selected gates and exact prediction change. □

C Derivation of the infinite-width kernel

Proof. Since $\boldsymbol{x}_i^\top \boldsymbol{x}_j$ is deterministic (it does not depend on $\boldsymbol{w}$), we pull it outside the expectation:

$$H_{ij}^{\infty} = \boldsymbol{x}_i^\top \boldsymbol{x}_j \cdot \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)}\left[\mathbf{1}_{\{\boldsymbol{w}^\top \boldsymbol{x}_i \geq 0,\, \boldsymbol{w}^\top \boldsymbol{x}_j \geq 0\}}\right] = \boldsymbol{x}_i^\top \boldsymbol{x}_j \cdot \mathbb{P}_{\boldsymbol{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)}(\boldsymbol{w}^\top \boldsymbol{x}_i \geq 0,\, \boldsymbol{w}^\top \boldsymbol{x}_j \geq 0).$$

It remains to show that

$$\mathbb{P}_{\boldsymbol{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)}(\boldsymbol{w}^\top \boldsymbol{x}_i \geq 0,\, \boldsymbol{w}^\top \boldsymbol{x}_j \geq 0) = \frac{\pi - \arccos\left(\boldsymbol{x}_i^\top \boldsymbol{x}_j\right)}{2\pi}. \tag{42}$$

Since $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ and $\|\boldsymbol{x}_i\|_2 = \|\boldsymbol{x}_j\|_2 = 1$, the pair $(\boldsymbol{w}^\top \boldsymbol{x}_i,\, \boldsymbol{w}^\top \boldsymbol{x}_j)$ is a zero-mean bivariate Gaussian with covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathrm{Var}(\boldsymbol{w}^\top \boldsymbol{x}_i) & \mathrm{Cov}(\boldsymbol{w}^\top \boldsymbol{x}_i, \boldsymbol{w}^\top \boldsymbol{x}_j) \\ \mathrm{Cov}(\boldsymbol{w}^\top \boldsymbol{x}_i, \boldsymbol{w}^\top \boldsymbol{x}_j) & \mathrm{Var}(\boldsymbol{w}^\top \boldsymbol{x}_j) \end{pmatrix} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where $\rho := \boldsymbol{x}_i^\top \boldsymbol{x}_j \in [-1, 1]$, using $\mathrm{Var}(\boldsymbol{w}^\top \boldsymbol{x}_i) = \|\boldsymbol{x}_i\|_2^2 = 1$ and $\mathrm{Cov}(\boldsymbol{w}^\top \boldsymbol{x}_i, \boldsymbol{w}^\top \boldsymbol{x}_j) = \boldsymbol{x}_i^\top \boldsymbol{x}_j = \rho$. It therefore suffices to compute $\mathbb{P}(Z_1 \geq 0,\, Z_2 \geq 0)$ for $(Z_1, Z_2) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

Set $\theta := \arccos(\rho) \in [0, \pi]$. We can write $(Z_1, Z_2) = (Y_1,\, \cos\theta \cdot Y_1 + \sin\theta \cdot Y_2)$ where $Y_1, Y_2 \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. This is valid since $\mathrm{Cov}(Y_1, \cos\theta \cdot Y_1 + \sin\theta \cdot Y_2) = \cos\theta = \rho$.

For $\theta = 0$, we have $Z_2 = Z_1$, so the probability is $1/2$. For $\theta = \pi$, we have $Z_2 = -Z_1$, so the probability is $\mathbb{P}(Z_1 = 0) = 0$. Both agree with equation 42.

For $0 < \theta < \pi$, write the isotropic Gaussian as

$$(Y_1, Y_2) = R(\cos\phi, \sin\phi), \qquad R > 0 \text{ a.s.}, \qquad \phi \sim \mathrm{Unif}[-\pi/2, 3\pi/2).$$

The angular coordinate is chosen on this interval so that neither of the following arcs wraps around its endpoints. For $R > 0$,

$$Z_1 \geq 0 \iff \cos\phi \geq 0 \iff \phi \in [-\pi/2, \pi/2],$$
$$Z_2 \geq 0 \iff \cos(\phi - \theta) \geq 0 \iff \phi \in [\theta - \pi/2, \theta + \pi/2].$$

Their intersection is $[\theta - \pi/2, \pi/2]$, of length $\pi - \theta$. Since the uniform angular distribution assigns zero probability to endpoints,

$$\mathbb{P}(Z_1 \geq 0, Z_2 \geq 0) = \frac{\pi - \theta}{2\pi} = \frac{\pi - \arccos\left(\boldsymbol{x}_i^\top \boldsymbol{x}_j\right)}{2\pi}.$$

Multiplying by $\boldsymbol{x}_i^\top \boldsymbol{x}_j$ gives the claimed kernel entry. □

D Proofs of the Intermediate Lemmas

Notation. For $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m]$, write $g_{ir}(\boldsymbol{Z}) := \mathbb{1}\{\boldsymbol{z}_r^\top \boldsymbol{x}_i \geq 0\}$. We call $g_{ir}(\boldsymbol{Z})$ the *gate of unit r on input $\boldsymbol{x}_i$ at $\boldsymbol{Z}$*: it is the value of the selected ReLU derivative σ' at the pre-activation $\boldsymbol{z}_r^\top \boldsymbol{x}_i$, and it equals one at the activation boundary. Saying that this gate *changes* between two weight matrices $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ means $g_{ir}(\boldsymbol{Z}) \neq g_{ir}(\boldsymbol{Z}')$, i.e. the pre-activations $\boldsymbol{z}_r^\top \boldsymbol{x}_i$ and $\boldsymbol{z}_r'^\top \boldsymbol{x}_i$ have opposite signs under the convention above.

D.1 Initialization

Proof of Lemma 4.1. We prove the three parts in the order used in the statement. The initialization assumptions give $\boldsymbol{w}_r(0) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ and $a_r \sim \mathrm{Unif}\{\pm\alpha\}$, independently over r and between the two layers. Since $\|\boldsymbol{x}_i\|_2 = 1$,

$$\boldsymbol{w}_r(0)^\top \boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{x}_i^\top \boldsymbol{I}_d \boldsymbol{x}_i) = \mathcal{N}(0, 1).$$

Moreover, $\mathbb{E}a_r = 0$, $\mathbb{E}a_r^2 = \alpha^2$, and $\mathbb{E}(a_r a_s) = 0$ for $r \neq s$.

Part 1. Let $Z \sim \mathcal{N}(0, 1)$. Symmetry of the Gaussian density gives

$$\mathbb{E}\sigma(Z)^2 = \mathbb{E}[Z^2 \mathbb{1}\{Z \geq 0\}] = \frac{1}{2}\mathbb{E}Z^2 = \frac{1}{2}.$$

For a fixed input $\boldsymbol{x}_i$, the initial prediction is $u_i(\boldsymbol{W}_0) = m^{-1/2}\sum_{r=1}^m a_r\sigma(\boldsymbol{w}_r(0)^\top \boldsymbol{x}_i)$. Independence of the two layers and $\mathbb{E}a_r = 0$ imply

$$\mathbb{E}u_i(\boldsymbol{W}_0) = \frac{1}{\sqrt{m}}\sum_{r=1}^m \mathbb{E}a_r\,\mathbb{E}\sigma(\boldsymbol{w}_r(0)^\top \boldsymbol{x}_i) = 0.$$

Expanding the square of the sum produces a double sum over the hidden units, and the two layers are independent, so each expectation factorises:

$$\begin{aligned}
\mathbb{E}u_i(\boldsymbol{W}_0)^2 &= \frac{1}{m}\sum_{r=1}^m\sum_{s=1}^m \mathbb{E}(a_r a_s)\,\mathbb{E}\big[\sigma(\boldsymbol{w}_r(0)^\top \boldsymbol{x}_i)\sigma(\boldsymbol{w}_s(0)^\top \boldsymbol{x}_i)\big] \\
&= \frac{\alpha^2}{m}\sum_{r=1}^m \mathbb{E}\sigma(\boldsymbol{w}_r(0)^\top \boldsymbol{x}_i)^2 && [\mathbb{E}(a_r a_s) = 0 \text{ for } r \neq s,\ \mathbb{E}a_r^2 = \alpha^2] \\
&= \frac{\alpha^2}{m}\cdot m\cdot\frac{1}{2} = \frac{\alpha^2}{2} && [\mathbb{E}\sigma(Z)^2 = \tfrac{1}{2}].
\end{aligned}$$

Expanding the squared residual coordinatewise and inserting the two moments just computed now gives

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_0)\|_2^2 &= \sum_{i=1}^n \big(y_i^2 - 2y_i\mathbb{E}u_i(\boldsymbol{W}_0) + \mathbb{E}u_i(\boldsymbol{W}_0)^2\big) && \text{[expanding the square]} \\
&= \sum_{i=1}^n \left(y_i^2 + \frac{\alpha^2}{2}\right) && [\mathbb{E}u_i(\boldsymbol{W}_0) = 0] \\
&\le n\left(B^2 + \frac{\alpha^2}{2}\right) && [y_i^2 \le B^2 \text{ for every } i].
\end{aligned}$$

Set $X := \|\boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W}_0)\|_2^2$ and $q := 3n(B^2 + \alpha^2/2)/\delta$. Then $\mathcal{E}_1^c = \{X > q\}$, so Markov's inequality gives

$$\mathbb{P}(\mathcal{E}_1^c) = \mathbb{P}(X > q) \le \frac{\mathbb{E}X}{q} \le \frac{n(B^2 + \alpha^2/2)}{3n(B^2 + \alpha^2/2)/\delta} = \frac{\delta}{3}.$$

This proves Part 1.

Part 2. Dividing the defining inequality for $\mathcal{E}_2$ by $\alpha^2 > 0$, we may write

$$\mathcal{E}_2 = \left\{\left\|\frac{\boldsymbol{H}(\boldsymbol{W}_0)}{\alpha^2} - \boldsymbol{H}^\infty\right\|_2 \le \frac{\lambda_0}{4}\right\}.$$

For fixed $i, j \in [n]$ and $r \in [m]$, define

$$Y_r^{ij} := \boldsymbol{x}_i^\top \boldsymbol{x}_j\, g_{ir}(\boldsymbol{W}_0) g_{jr}(\boldsymbol{W}_0).$$

The variables $Y_1^{ij}, \dots, Y_m^{ij}$ are independent, $\mathbb{E}Y_r^{ij} = H_{ij}^\infty$, and $H_{ij}(\boldsymbol{W}_0)/\alpha^2 = m^{-1}\sum_{r=1}^m Y_r^{ij}$. Each Y_r^{ij} lies between 0 and $\boldsymbol{x}_i^\top \boldsymbol{x}_j$, in either order, so its range has length $|\boldsymbol{x}_i^\top \boldsymbol{x}_j| \le \|\boldsymbol{x}_i\|_2\|\boldsymbol{x}_j\|_2 = 1$. Hoeffding's inequality therefore gives, for every $h > 0$,

$$\mathbb{P}\left(\left|\frac{H_{ij}(\boldsymbol{W}_0)}{\alpha^2} - H_{ij}^\infty\right| > h\right) \le 2e^{-2mh^2}.$$

We take $h = \lambda_0/(4n)$, for which the exponent becomes $2mh^2 = 2m\lambda_0^2/(16n^2) = m\lambda_0^2/(8n^2)$, and we apply the union bound over the n^2 pairs (i, j):

$$\begin{aligned}
\mathbb{P}\left(\max_{i,j\in[n]}\left|\frac{H_{ij}(\boldsymbol{W}_0)}{\alpha^2} - H_{ij}^\infty\right| > \frac{\lambda_0}{4n}\right) &\le 2n^2 e^{-m\lambda_0^2/(8n^2)} && \text{[union bound over } n^2 \text{ pairs]} \\
&\le 2n^2 e^{-\log(6n^2/\delta)} && \text{[first width bound equation 7]} \\
&= 2n^2\cdot\frac{\delta}{6n^2} = \frac{\delta}{3}.
\end{aligned}$$

The middle step is the only place where the first entry of the width requirement equation 7 is used: $m \ge (8n^2/\lambda_0^2)\log(6n^2/\delta)$ is exactly the statement that $m\lambda_0^2/(8n^2) \ge \log(6n^2/\delta)$, and the exponential is decreasing.

It remains to convert this entrywise estimate into the operator-norm statement that defines $\mathcal{E}_2$. Let $\boldsymbol{D} \coloneqq \boldsymbol{H}(\boldsymbol{W}_0)/\alpha^2 - \boldsymbol{H}^\infty$ and suppose every entry of $\boldsymbol{D}$ has absolute value at most $\lambda_0/(4n)$. Then

$$\begin{aligned}
\|\boldsymbol{D}\|_2 \le \|\boldsymbol{D}\|_F &= \left(\sum_{i=1}^{n}\sum_{j=1}^{n} |D_{ij}|^2\right)^{1/2} && \text{[operator norm} \le \text{Frobenius norm]} \\
&\le n \max_{i,j\in[n]} |D_{ij}| && [n^2 \text{ summands}] \\
&\le \frac{\lambda_0}{4} && \text{[entrywise hypothesis]}.
\end{aligned}$$

Contrapositively, failure of $\mathcal{E}_2$ forces at least one entry error to exceed $\lambda_0/(4n)$, which is exactly the event bounded above by $\delta/3$. Hence $\mathbb{P}(\mathcal{E}_2^c) \le \delta/3$, which proves Part 2.

Part 3. Set $\tau \coloneqq \delta\lambda_0/(1536\sqrt{2/\pi}\, n)$. Since $\boldsymbol{w}_r(0)^\top \boldsymbol{x}_i$ is standard normal, the probability of landing in a window of half-width τ is at most the length of the window times the peak of the density:

$$\begin{aligned}
\mathbb{P}\big(|\boldsymbol{w}_r(0)^\top \boldsymbol{x}_i| \le \tau\big) &= \int_{-\tau}^{\tau} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz \\
&\le \frac{2\tau}{\sqrt{2\pi}} && [e^{-z^2/2} \le 1] \\
&= \frac{\delta\lambda_0}{1536n} && \text{[definition of } \tau].
\end{aligned}$$

Let $N \coloneqq \sum_{i=1}^{n}\sum_{r=1}^{m} \mathbb{1}\{|\boldsymbol{w}_r(0)^\top \boldsymbol{x}_i| \le \tau\}$. Linearity of expectation, which does not require independence among the summands, gives

$$\mathbb{E}N = \sum_{i=1}^{n}\sum_{r=1}^{m} \mathbb{P}\big(|\boldsymbol{w}_r(0)^\top \boldsymbol{x}_i| \le \tau\big) \le \frac{m\delta\lambda_0}{1536}.$$

Since $\mathcal{E}_3^c = \{N > m\lambda_0/512\}$, Markov's inequality gives

$$\mathbb{P}(\mathcal{E}_3^c) \le \frac{\mathbb{E}N}{m\lambda_0/512} \le \frac{512}{m\lambda_0}\frac{m\delta\lambda_0}{1536} = \frac{\delta}{3}.$$

This proves Part 3 and completes the proof. □

D.2 Uniform kernel and remainder bounds

Proof of Lemma 4.2. We prove the three parts in the order used in the statement. Fix an initialization in $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ and arbitrary $\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_m]$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_m]$ in $\mathcal{B}$. Set

$$\tau \coloneqq \frac{\delta\lambda_0}{1536\sqrt{2/\pi}\, n}, \qquad R_{\mathcal{B}} \coloneqq \frac{5n}{\alpha}\sqrt{\frac{3(B^2 + \alpha^2/2)}{\delta\lambda_0}}.$$

Then $\|\boldsymbol{W} - \boldsymbol{W}_0\|_F \le R_{\mathcal{B}}$ and $\|\boldsymbol{V} - \boldsymbol{W}_0\|_F \le R_{\mathcal{B}}$.

Part 1. For each $i \in [n]$, define the small-margin set $\mathcal{S}_i \coloneqq \{r \in [m] : |\boldsymbol{w}_r(0)^\top \boldsymbol{x}_i| \le \tau\}$. For $\boldsymbol{Z} = [\boldsymbol{z}_1, \dots, \boldsymbol{z}_m] \in \{\boldsymbol{W}, \boldsymbol{V}\}$, define $\mathcal{I}(\boldsymbol{Z}) \coloneqq \{r \in [m] : \|\boldsymbol{z}_r - \boldsymbol{w}_r(0)\|_2 > \tau\}$.

We first show that no gate g_{ir} can change between $\boldsymbol{W}_0$ and $\boldsymbol{Z}$ unless r lies in $\mathcal{S}_i \cup \mathcal{I}(\boldsymbol{Z})$. If $r \notin \mathcal{I}(\boldsymbol{Z})$, then Cauchy–Schwarz gives

$$|\boldsymbol{z}_r^\top \boldsymbol{x}_i - \boldsymbol{w}_r(0)^\top \boldsymbol{x}_i| = |(\boldsymbol{z}_r - \boldsymbol{w}_r(0))^\top \boldsymbol{x}_i| \le \|\boldsymbol{z}_r - \boldsymbol{w}_r(0)\|_2 \|\boldsymbol{x}_i\|_2 \le \tau.$$

If also $r \notin \mathcal{S}_i$, then $|\boldsymbol{w}_r(0)^\top \boldsymbol{x}_i| > \tau$. The two possible signs give

$$\begin{aligned}
\boldsymbol{w}_r(0)^\top \boldsymbol{x}_i > \tau &\implies \boldsymbol{z}_r^\top \boldsymbol{x}_i \ge \boldsymbol{w}_r(0)^\top \boldsymbol{x}_i - \tau > 0, \\
\boldsymbol{w}_r(0)^\top \boldsymbol{x}_i < -\tau &\implies \boldsymbol{z}_r^\top \boldsymbol{x}_i \le \boldsymbol{w}_r(0)^\top \boldsymbol{x}_i + \tau < 0.
\end{aligned}$$

Hence $g_{ir}(\boldsymbol{Z}) = g_{ir}(\boldsymbol{W}_0)$ whenever $r \notin \mathcal{S}_i \cup \mathcal{I}(\boldsymbol{Z})$. Therefore

$$\mathcal{A}_i(\boldsymbol{W}, \boldsymbol{V}) \subseteq \mathcal{S}_i \cup \mathcal{I}(\boldsymbol{W}) \cup \mathcal{I}(\boldsymbol{V}). \tag{43}$$

We now bound how many columns can lie in $\mathcal{I}(\boldsymbol{Z})$. For either $\boldsymbol{Z} \in \{\boldsymbol{W}, \boldsymbol{V}\}$, each index $r \in \mathcal{I}(\boldsymbol{Z})$ contributes more than τ^2 to the squared Frobenius displacement, by the very definition of $\mathcal{I}(\boldsymbol{Z})$, and the Frobenius norm sums these column contributions over *all* $r \in [m]$. Hence

$$\begin{aligned}
|\mathcal{I}(\boldsymbol{Z})|\,\tau^2 &\le \sum_{r\in\mathcal{I}(\boldsymbol{Z})} \|\boldsymbol{z}_r - \boldsymbol{w}_r(0)\|_2^2 && [\|\boldsymbol{z}_r - \boldsymbol{w}_r(0)\|_2 > \tau \text{ on } \mathcal{I}(\boldsymbol{Z})] \\
&\le \sum_{r=1}^{m} \|\boldsymbol{z}_r - \boldsymbol{w}_r(0)\|_2^2 = \|\boldsymbol{Z} - \boldsymbol{W}_0\|_F^2 && [\text{extra terms are} \ge 0] \\
&\le R_{\mathcal{B}}^2 && [\boldsymbol{Z} \in \mathcal{B}].
\end{aligned}$$

Dividing by τ^2 and substituting the definitions of τ and $R_{\mathcal{B}}$ gives

$$|\mathcal{I}(\boldsymbol{Z})| \le \frac{353894400\, n^4 (B^2 + \alpha^2/2)}{\pi \alpha^2 \delta^3 \lambda_0^3}, \qquad \boldsymbol{Z} \in \{\boldsymbol{W}, \boldsymbol{V}\}. \tag{44}$$

We can now assemble the two estimates. Taking cardinalities in the inclusion equation 43, the cardinality of a union is at most the sum of the cardinalities; summing the result over the n inputs and dividing by m gives

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{n} |\mathcal{A}_i(\boldsymbol{W}, \boldsymbol{V})| &\le \frac{1}{m}\sum_{i=1}^{n} \Big(|\mathcal{S}_i| + |\mathcal{I}(\boldsymbol{W})| + |\mathcal{I}(\boldsymbol{V})| \Big) && [\text{by equation 43}] \\
&= \frac{1}{m}\sum_{i=1}^{n} |\mathcal{S}_i| + \frac{n}{m}\,(|\mathcal{I}(\boldsymbol{W})| + |\mathcal{I}(\boldsymbol{V})|) && [\mathcal{I}(\boldsymbol{W}), \mathcal{I}(\boldsymbol{V}) \text{ do not depend on } i].
\end{aligned}$$

The two terms are controlled by different hypotheses. Event $\mathcal{E}_3$ is exactly the statement that the first term is at most $\lambda_0/512$. For the second term, $\Omega(\cdot)$ in the width requirement equation 7 stands for the explicit bound

$$m \ge \frac{362387865600\, n^5 (B^2 + \alpha^2/2)}{\pi \alpha^2 \delta^3 \lambda_0^4}, \tag{45}$$

which the column-count bound converts into

$$\frac{n}{m}\,(|\mathcal{I}(\boldsymbol{W})| + |\mathcal{I}(\boldsymbol{V})|) \le \frac{707788800\, n^5 (B^2 + \alpha^2/2)}{\pi m \alpha^2 \delta^3 \lambda_0^3} \le \frac{\lambda_0}{512}.$$

Adding the two estimates yields

$$\frac{1}{m}\sum_{i=1}^{n} |\mathcal{A}_i(\boldsymbol{W}, \boldsymbol{V})| \le \frac{\lambda_0}{512} + \frac{\lambda_0}{512} = \frac{\lambda_0}{256}.$$

This proves Part 1.

Part 2. The initialization $\boldsymbol{W}_0$ belongs to $\mathcal{B}$. For the pair $(\boldsymbol{W}, \boldsymbol{W}_0)$,

$$\mathcal{A}_i(\boldsymbol{W}, \boldsymbol{W}_0) = \{r \in [m] : g_{ir}(\boldsymbol{W}) \ne g_{ir}(\boldsymbol{W}_0)\}.$$

The (i, r) block of $\boldsymbol{J}(\boldsymbol{W}) - \boldsymbol{J}(\boldsymbol{W}_0)$ equals

$$\frac{a_r}{\sqrt{m}} \big(g_{ir}(\boldsymbol{W}) - g_{ir}(\boldsymbol{W}_0) \big) \boldsymbol{x}_i^\top.$$

Summing the squared Frobenius norms of these blocks, and using $a_r^2 = \alpha^2$ together with $\|\boldsymbol{x}_i\|_2 = 1$, gives

$$\begin{aligned}
\|\boldsymbol{J}(\boldsymbol{W}) - \boldsymbol{J}(\boldsymbol{W}_0)\|_F^2 &= \frac{\alpha^2}{m}\sum_{i=1}^{n}\sum_{r=1}^{m} \big(g_{ir}(\boldsymbol{W}) - g_{ir}(\boldsymbol{W}_0) \big)^2 && [\text{blockwise sum}] \\
&= \frac{\alpha^2}{m}\sum_{i=1}^{n} |\mathcal{A}_i(\boldsymbol{W}, \boldsymbol{W}_0)| && [\text{gates are binary}] \\
&\le \frac{\alpha^2 \lambda_0}{256} && [\text{Part 1 applied to } (\boldsymbol{W}, \boldsymbol{W}_0)].
\end{aligned}$$

The middle step uses that a difference of two zero-one gates is 0 when they agree and ± 1 when they differ, so its square indicates that $g_{ir}(\boldsymbol{W}) \neq g_{ir}(\boldsymbol{W}_0)$, and summing those indicators over r counts exactly $|\mathcal{A}_i(\boldsymbol{W}, \boldsymbol{W}_0)|$. Taking square roots proves

$$\|\boldsymbol{J}(\boldsymbol{W}) - \boldsymbol{J}(\boldsymbol{W}_0)\|_F \le \frac{\alpha\sqrt{\lambda_0}}{16}.$$

We now lower-bound the empirical kernel, first at the initialization and then at an arbitrary point of $\mathcal{B}$. For any $v \in \mathbb{R}^n$, add and subtract the infinite-width kernel:

$$\begin{aligned}
\|\boldsymbol{J}(\boldsymbol{W}_0)^\top v\|_2^2 &= v^\top \boldsymbol{H}(\boldsymbol{W}_0) v && [\boldsymbol{H} = \boldsymbol{J}\boldsymbol{J}^\top] \\
&= \alpha^2 v^\top \boldsymbol{H}^\infty v + v^\top \big(\boldsymbol{H}(\boldsymbol{W}_0) - \alpha^2 \boldsymbol{H}^\infty\big) v && [\text{add and subtract}] \\
&\ge \alpha^2 \lambda_0 \|v\|_2^2 - \|\boldsymbol{H}(\boldsymbol{W}_0) - \alpha^2 \boldsymbol{H}^\infty\|_2 \|v\|_2^2 && [v^\top \boldsymbol{A} v \ge -\|\boldsymbol{A}\|_2 \|v\|_2^2] \\
&\ge \alpha^2 \lambda_0 \|v\|_2^2 - \frac{\alpha^2 \lambda_0}{4} \|v\|_2^2 = \frac{3\alpha^2 \lambda_0}{4} \|v\|_2^2 && [\text{event } \mathcal{E}_2].
\end{aligned}$$

The third line bounds the first term below using $\lambda_{\min}(\boldsymbol{H}^\infty) = \lambda_0$ and the second term below using the symmetry of $\boldsymbol{H}(\boldsymbol{W}_0) - \alpha^2 \boldsymbol{H}^\infty$. Taking square roots,

$$\|\boldsymbol{J}(\boldsymbol{W}_0)^\top v\|_2 \ge \frac{\sqrt{3}}{2} \alpha \sqrt{\lambda_0} \|v\|_2.$$

To transfer this from $\boldsymbol{W}_0$ to $\boldsymbol{W}$, note that $\|(\boldsymbol{J}(\boldsymbol{W}) - \boldsymbol{J}(\boldsymbol{W}_0))^\top v\|_2 \le \|\boldsymbol{J}(\boldsymbol{W}) - \boldsymbol{J}(\boldsymbol{W}_0)\|_F \|v\|_2 \le (\alpha\sqrt{\lambda_0}/16)\|v\|_2$ by the Jacobian estimate just proved. The reverse triangle inequality then gives

$$\begin{aligned}
\|\boldsymbol{J}(\boldsymbol{W})^\top v\|_2 &\ge \|\boldsymbol{J}(\boldsymbol{W}_0)^\top v\|_2 - \|(\boldsymbol{J}(\boldsymbol{W}) - \boldsymbol{J}(\boldsymbol{W}_0))^\top v\|_2 && [\text{reverse triangle inequality}] \\
&\ge \left(\frac{\sqrt{3}}{2} - \frac{1}{16}\right) \alpha \sqrt{\lambda_0} \|v\|_2 && [\text{the two displays above}].
\end{aligned}$$

The bracket is positive and $(\sqrt{3}/2 - 1/16)^2 = 193/256 - \sqrt{3}/16 > 1/2$, so squaring gives

$$v^\top \boldsymbol{H}(\boldsymbol{W}) v = \|\boldsymbol{J}(\boldsymbol{W})^\top v\|_2^2 \ge \frac{\alpha^2 \lambda_0}{2} \|v\|_2^2.$$

This holds for every v, so $\lambda_{\min}(\boldsymbol{H}(\boldsymbol{W})) \ge \alpha^2 \lambda_0 / 2$. Together with the Jacobian estimate, this proves Part 2.

Part 3. For scalars z, h, define

$$R(z, h) := \max\{0, z + h\} - \max\{0, z\} - \mathbb{1}\{z \ge 0\} h.$$

The four sign configurations give

$$R(z, h) = \begin{cases} 0, & z \ge 0,\ z + h \ge 0, \\ 0, & z < 0,\ z + h < 0, \\ -z - h, & z \ge 0,\ z + h < 0, \\ z + h, & z < 0,\ z + h \ge 0. \end{cases}$$

The scalar z here stands for a pre-activation $\boldsymbol{w}_r^\top \boldsymbol{x}_i$ and $z + h$ for the corresponding pre-activation $\boldsymbol{v}_r^\top \boldsymbol{x}_i$ after the step, so the selected derivative $\sigma' = \mathbb{1}\{\cdot \ge 0\}$ takes the value $\mathbb{1}\{z \ge 0\}$ before the step and $\mathbb{1}\{z + h \ge 0\}$ after it. The first two cases of the display are exactly the cases in which these two indicators agree, and there $R(z, h) = 0$: the selected derivative is the true derivative and the first-order expansion is exact. The last two cases are the ones in which they differ; there, the third case gives $0 < -(z + h) \le -h = |h|$ and the fourth gives $0 \le z + h \le h = |h|$. Hence

$$R(z, h) = 0 \ \text{ whenever } \ \mathbb{1}\{z \ge 0\} = \mathbb{1}\{z + h \ge 0\}, \qquad |R(z, h)| \le |h| \text{ in every case.} \tag{46}$$

Let $\Delta\boldsymbol{W} := \boldsymbol{V} - \boldsymbol{W}$, with columns $\Delta\boldsymbol{w}_r := \boldsymbol{v}_r - \boldsymbol{w}_r$, and define

$$\boldsymbol{\xi} := \boldsymbol{u}(\boldsymbol{V}) - \boldsymbol{u}(\boldsymbol{W}) - \boldsymbol{J}(\boldsymbol{W}) \operatorname{vec}(\Delta\boldsymbol{W}).$$

By definition, $\xi_i = [\boldsymbol{u}(\boldsymbol{V})]_i - [\boldsymbol{u}(\boldsymbol{W})]_i - [\boldsymbol{J}(\boldsymbol{W})\,\mathrm{vec}(\Delta\boldsymbol{W})]_i$. The first two terms are given by the network output equation 2, with $[\boldsymbol{u}(\boldsymbol{Z})]_i = f(\boldsymbol{x}_i; \boldsymbol{Z})$. For the third term, vec stacks the columns, so $\mathrm{vec}(\Delta\boldsymbol{W})$ consists of the m blocks $\Delta\boldsymbol{w}_1, \ldots, \Delta\boldsymbol{w}_m$, each of length d. By Definition 1, the i-th row of $\boldsymbol{J}(\boldsymbol{W})$ is cut into the same m blocks, and its r-th block is $\frac{a_r}{\sqrt{m}} g_{ir}(\boldsymbol{W})\boldsymbol{x}_i^\top$. Multiplying block by block,

$$[\boldsymbol{J}(\boldsymbol{W})\,\mathrm{vec}(\Delta\boldsymbol{W})]_i = \sum_{r=1}^{m} \underbrace{[\boldsymbol{J}(\boldsymbol{W})]_{i,(r-1)d+1:rd}}_{=\frac{a_r}{\sqrt{m}} g_{ir}(\boldsymbol{W})\,\boldsymbol{x}_i^\top} \underbrace{\Delta\boldsymbol{w}_r}_{r\text{-th block of } \mathrm{vec}(\Delta\boldsymbol{W})} = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r\, g_{ir}(\boldsymbol{W})\, \Delta\boldsymbol{w}_r^\top \boldsymbol{x}_i. \tag{47}$$

Now put $z := \boldsymbol{w}_r^\top \boldsymbol{x}_i$ and $h := \Delta\boldsymbol{w}_r^\top \boldsymbol{x}_i$ in the r-th summand. Then $z+h = \boldsymbol{v}_r^\top \boldsymbol{x}_i$ because $\boldsymbol{v}_r = \boldsymbol{w}_r + \Delta\boldsymbol{w}_r$, and $g_{ir}(\boldsymbol{W}) = \mathbb{1}\{z \ge 0\}$ by definition. The underbraces record the value of each term in terms of z and h, and the three of them together are exactly $R(z,h)$:

$$\begin{aligned}
\xi_i &= [\boldsymbol{u}(\boldsymbol{V})]_i - [\boldsymbol{u}(\boldsymbol{W})]_i - [\boldsymbol{J}(\boldsymbol{W})\,\mathrm{vec}(\Delta\boldsymbol{W})]_i \\
&= \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \Big[\underbrace{\sigma(\boldsymbol{v}_r^\top \boldsymbol{x}_i)}_{=\max\{0,\, z+h\}} - \underbrace{\sigma(\boldsymbol{w}_r^\top \boldsymbol{x}_i)}_{=\max\{0,\, z\}} - \underbrace{g_{ir}(\boldsymbol{W})\, \Delta\boldsymbol{w}_r^\top \boldsymbol{x}_i}_{=\mathbb{1}\{z\ge 0\}\, h} \Big] && \text{[equation 2, equation 47]} \\
&= \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r\, R(z,h) && \text{[definition of } R\text{]}. \qquad (48)
\end{aligned}$$

Define $\mathcal{C}_i := \{r \in [m] : g_{ir}(\boldsymbol{V}) \ne g_{ir}(\boldsymbol{W})\}$. The scalar calculation shows that the summand is zero outside $\mathcal{C}_i$. Also, $\mathcal{C}_i \subseteq \mathcal{A}_i(\boldsymbol{W}, \boldsymbol{V})$, because two different values cannot both equal $g_{ir}(\boldsymbol{W}_0)$. We may therefore restrict the sum to $\mathcal{A}_i(\boldsymbol{W}, \boldsymbol{V})$ and bound each surviving term by $|R(z,h)| \le |h|$ from equation 46. Using $|a_r| = \alpha$ and then Cauchy–Schwarz on a sum of $|\mathcal{A}_i(\boldsymbol{W}, \boldsymbol{V})|$ terms,

$$\begin{aligned}
|\xi_i| &\le \frac{\alpha}{\sqrt{m}} \sum_{r \in \mathcal{A}_i(\boldsymbol{W}, \boldsymbol{V})} |\Delta\boldsymbol{w}_r^\top \boldsymbol{x}_i| && \text{[by equation 46]} \\
&\le \frac{\alpha}{\sqrt{m}} \sqrt{|\mathcal{A}_i(\boldsymbol{W}, \boldsymbol{V})|} \left(\sum_{r \in \mathcal{A}_i(\boldsymbol{W}, \boldsymbol{V})} |\Delta\boldsymbol{w}_r^\top \boldsymbol{x}_i|^2 \right)^{1/2} && \text{[Cauchy–Schwarz]} \\
&\le \alpha \sqrt{\frac{|\mathcal{A}_i(\boldsymbol{W}, \boldsymbol{V})|}{m}}\, \|\Delta\boldsymbol{W}^\top \boldsymbol{x}_i\|_2 && \text{[adding the remaining } r\text{]} \\
&\le \alpha \sqrt{\frac{|\mathcal{A}_i(\boldsymbol{W}, \boldsymbol{V})|}{m}}\, \|\boldsymbol{V} - \boldsymbol{W}\|_2 && [\|\Delta\boldsymbol{W}^\top \boldsymbol{x}_i\|_2 \le \|\Delta\boldsymbol{W}\|_2,\ \|\boldsymbol{x}_i\|_2 = 1].
\end{aligned}$$

The third line only enlarges a sum of nonnegative squares, from $r \in \mathcal{A}_i(\boldsymbol{W}, \boldsymbol{V})$ to all $r \in [m]$, which is exactly $\|\Delta\boldsymbol{W}^\top \boldsymbol{x}_i\|_2^2$. Squaring, summing over i, and then applying Part 1 gives

$$\|\boldsymbol{\xi}\|_2^2 \le \alpha^2 \|\boldsymbol{V} - \boldsymbol{W}\|_2^2 \frac{1}{m} \sum_{i=1}^{n} |\mathcal{A}_i(\boldsymbol{W}, \boldsymbol{V})| \le \frac{\alpha^2 \lambda_0}{256} \|\boldsymbol{V} - \boldsymbol{W}\|_2^2.$$

Therefore

$$\|\boldsymbol{\xi}\|_2 \le \frac{\alpha\sqrt{\lambda_0}}{16} \|\boldsymbol{V} - \boldsymbol{W}\|_2 \le \frac{\alpha\sqrt{\lambda_0}}{8} \|\boldsymbol{V} - \boldsymbol{W}\|_2.$$

This proves Part 3 and completes the proof. □

D.3 Spectral steps and residual descent

Proof of Lemma 4.3. Let $\boldsymbol{e} := \boldsymbol{y} - \boldsymbol{u}(\boldsymbol{W})$ and $\Delta\boldsymbol{W} := \boldsymbol{V} - \boldsymbol{W} = -\eta\,\mathrm{Polar}(\boldsymbol{G})$.

Part 1. The r-th column of the selected gradient is

$$[\boldsymbol{G}]_{\cdot,r} = -\frac{1}{\sqrt{m}} \sum_{i=1}^{n} e_i a_r g_{ir}(\boldsymbol{W}) \boldsymbol{x}_i.$$

Every column of $\boldsymbol{G}$ therefore lies in $\mathrm{span}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, and hence

$$\mathrm{rank}(\boldsymbol{G}) \le \dim \mathrm{span}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \le n.$$

For $\boldsymbol{G} \neq 0$, let $\boldsymbol{G} = \boldsymbol{U}_{\boldsymbol{G}} \boldsymbol{\Sigma}_{\boldsymbol{G}} \boldsymbol{V}_{\boldsymbol{G}}^{\top}$ be its compact SVD.

Then $\mathrm{Polar}(\boldsymbol{G}) = \boldsymbol{U}_{\boldsymbol{G}} \boldsymbol{V}_{\boldsymbol{G}}^{\top}$, where $\boldsymbol{U}_{\boldsymbol{G}}$ and $\boldsymbol{V}_{\boldsymbol{G}}$ have orthonormal columns, so

$$\mathrm{Polar}(\boldsymbol{G})^{\top}\,\mathrm{Polar}(\boldsymbol{G}) = \boldsymbol{V}_{\boldsymbol{G}} \boldsymbol{U}_{\boldsymbol{G}}^{\top} \boldsymbol{U}_{\boldsymbol{G}} \boldsymbol{V}_{\boldsymbol{G}}^{\top} = \boldsymbol{V}_{\boldsymbol{G}} \boldsymbol{V}_{\boldsymbol{G}}^{\top},$$

the orthogonal projection onto the row space of $\boldsymbol{G}$. The singular values of $\mathrm{Polar}(\boldsymbol{G})$ are therefore the square roots of the eigenvalues of a projection of rank $\mathrm{rank}(\boldsymbol{G})$: that is, $\mathrm{rank}(\boldsymbol{G})$ of them equal one and the rest are zero. Reading off the largest of them and their sum of squares,

$$\|\,\mathrm{Polar}(\boldsymbol{G})\|_2 = 1, \qquad \|\,\mathrm{Polar}(\boldsymbol{G})\|_F^2 = \mathrm{rank}(\boldsymbol{G}) \leq n,$$

the last inequality being the rank bound established above. For $\boldsymbol{G} = 0$ the same two upper bounds hold trivially, because $\mathrm{Polar}(0) = 0$. Since $\Delta \boldsymbol{W} = -\eta\,\mathrm{Polar}(\boldsymbol{G})$,

$$\|\boldsymbol{V} - \boldsymbol{W}\|_2 \leq \eta, \qquad \|\boldsymbol{V} - \boldsymbol{W}\|_F \leq \sqrt{n}\,\eta.$$

Write $\Delta \boldsymbol{w}_r$ for the r-th column of $\Delta \boldsymbol{W}$. The ReLU is 1-Lipschitz, so each of the m summands defining the prediction moves by at most $|\Delta \boldsymbol{w}_r^{\top} \boldsymbol{x}_i|$, and $|a_r| = \alpha$. Hence, for every i,

$$\begin{aligned}
|u_i(\boldsymbol{V}) - u_i(\boldsymbol{W})| &\leq \frac{\alpha}{\sqrt{m}} \sum_{r=1}^{m} |\Delta \boldsymbol{w}_r^{\top} \boldsymbol{x}_i| && \text{[ReLU is 1-Lipschitz]} \\
&\leq \frac{\alpha}{\sqrt{m}} \sqrt{m} \left(\sum_{r=1}^{m} |\Delta \boldsymbol{w}_r^{\top} \boldsymbol{x}_i|^2 \right)^{1/2} && \text{[Cauchy–Schwarz]} \\
&= \alpha \|\Delta \boldsymbol{W}^{\top} \boldsymbol{x}_i\|_2 \leq \alpha \|\Delta \boldsymbol{W}\|_2 \|\boldsymbol{x}_i\|_2 \leq \alpha\eta && [\|\Delta \boldsymbol{W}\|_2 \leq \eta,\ \|\boldsymbol{x}_i\|_2 = 1].
\end{aligned}$$

The factor $\sqrt{m}$ produced by Cauchy–Schwarz over m terms cancels the $1/\sqrt{m}$ in the network's normalization, which is why the bound is independent of the width. Squaring the coordinate bounds and summing over i gives

$$\|\boldsymbol{u}(\boldsymbol{V}) - \boldsymbol{u}(\boldsymbol{W})\|_2^2 = \sum_{i=1}^{n} |u_i(\boldsymbol{V}) - u_i(\boldsymbol{W})|^2 \leq n\alpha^2\eta^2.$$

This proves Part 1.

Part 2. Assume $\mathcal{E}$ holds and $\boldsymbol{W}, \boldsymbol{V} \in \mathcal{B}$. The selected-gradient identity is $\mathrm{vec}(\boldsymbol{G}) = -\boldsymbol{J}(\boldsymbol{W})^{\top} \boldsymbol{e}$. For $\boldsymbol{G} \neq 0$, the compact SVD gives

$$\begin{aligned}
\langle \boldsymbol{G}, \mathrm{Polar}(\boldsymbol{G}) \rangle &= \mathrm{tr}\big(\boldsymbol{V}_{\boldsymbol{G}} \boldsymbol{\Sigma}_{\boldsymbol{G}} \boldsymbol{U}_{\boldsymbol{G}}^{\top} \boldsymbol{U}_{\boldsymbol{G}} \boldsymbol{V}_{\boldsymbol{G}}^{\top}\big) && [\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{tr}(\boldsymbol{A}^{\top} \boldsymbol{B})] \\
&= \mathrm{tr}(\boldsymbol{V}_{\boldsymbol{G}} \boldsymbol{\Sigma}_{\boldsymbol{G}} \boldsymbol{V}_{\boldsymbol{G}}^{\top}) = \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{G}}) && [\boldsymbol{U}_{\boldsymbol{G}}^{\top} \boldsymbol{U}_{\boldsymbol{G}} = \boldsymbol{I}, \text{ cyclicity of trace}] \\
&= \|\boldsymbol{G}\|_* && \text{[sum of singular values]}.
\end{aligned}$$

The same identity holds when $\boldsymbol{G} = 0$. Therefore

$$\boldsymbol{e}^{\top} \boldsymbol{J}(\boldsymbol{W})\,\mathrm{vec}(\Delta \boldsymbol{W}) = (-\mathrm{vec}(\boldsymbol{G}))^{\top} (-\eta\,\mathrm{vec}(\mathrm{Polar}(\boldsymbol{G}))) = \eta \|\boldsymbol{G}\|_*. \tag{49}$$

The singular values are nonnegative, so the square of their sum dominates the sum of their squares, which is to say $\|\boldsymbol{G}\|_*^2 \geq \|\boldsymbol{G}\|_F^2$. Combining this with the selected-gradient identity and the kernel lower bound of Lemma 4.2,

$$\begin{aligned}
\|\boldsymbol{G}\|_*^2 \geq \|\boldsymbol{G}\|_F^2 &= \|\boldsymbol{J}(\boldsymbol{W})^{\top} \boldsymbol{e}\|_2^2 && [\mathrm{vec}(\boldsymbol{G}) = -\boldsymbol{J}(\boldsymbol{W})^{\top} \boldsymbol{e}] \\
&= \boldsymbol{e}^{\top} \boldsymbol{J}(\boldsymbol{W}) \boldsymbol{J}(\boldsymbol{W})^{\top} \boldsymbol{e} = \boldsymbol{e}^{\top} \boldsymbol{H}(\boldsymbol{W}) \boldsymbol{e} && [\text{definition of } \boldsymbol{H}] \\
&\geq \frac{\alpha^2 \lambda_0}{2} \|\boldsymbol{e}\|_2^2 && \text{[by equation 15]}.
\end{aligned}$$

Taking square roots yields

$$\|\boldsymbol{G}\|_* \geq \frac{\alpha \sqrt{\lambda_0}}{\sqrt{2}} \|\boldsymbol{e}\|_2.$$

Let $\boldsymbol{\xi} := \boldsymbol{u}(\boldsymbol{V}) - \boldsymbol{u}(\boldsymbol{W}) - \boldsymbol{J}(\boldsymbol{W})\,\mathrm{vec}(\Delta \boldsymbol{W})$. The remainder bound of Lemma 4.2 and the step bound proved above give

$$\|\boldsymbol{\xi}\|_2 \leq \frac{\alpha \sqrt{\lambda_0}}{8} \|\Delta \boldsymbol{W}\|_2 \leq \frac{\alpha\eta \sqrt{\lambda_0}}{8}.$$

Since $\boldsymbol{y}-\boldsymbol{u}(\boldsymbol{V}) = \boldsymbol{e}-(\boldsymbol{u}(\boldsymbol{V})-\boldsymbol{u}(\boldsymbol{W}))$, expanding the squared norm and then splitting the prediction change into its linear part and the remainder $\boldsymbol{\xi}$ gives

$$\begin{aligned}
\|\boldsymbol{y}-\boldsymbol{u}(\boldsymbol{V})\|_2^2 &= \|\boldsymbol{e}\|_2^2 - 2\boldsymbol{e}^\top(\boldsymbol{u}(\boldsymbol{V})-\boldsymbol{u}(\boldsymbol{W})) + \|\boldsymbol{u}(\boldsymbol{V})-\boldsymbol{u}(\boldsymbol{W})\|_2^2 && \text{[expanding the square]}\\
&= \|\boldsymbol{e}\|_2^2 - 2\boldsymbol{e}^\top \boldsymbol{J}(\boldsymbol{W})\,\mathrm{vec}(\Delta\boldsymbol{W}) - 2\boldsymbol{e}^\top\boldsymbol{\xi} + \|\boldsymbol{u}(\boldsymbol{V})-\boldsymbol{u}(\boldsymbol{W})\|_2^2 && \text{[definition of } \boldsymbol{\xi}]\\
&= \|\boldsymbol{e}\|_2^2 - 2\eta\|\boldsymbol{G}\|_* - 2\boldsymbol{e}^\top\boldsymbol{\xi} + \|\boldsymbol{u}(\boldsymbol{V})-\boldsymbol{u}(\boldsymbol{W})\|_2^2 && \text{[by equation 49].}
\end{aligned}$$

The three terms now play distinct roles: the second is the guaranteed first-order gain, the third is the linearization error, and the fourth is the second-order cost of a finite step. Cauchy–Schwarz and the remainder bound imply

$$-2\boldsymbol{e}^\top\boldsymbol{\xi} \le 2\|\boldsymbol{e}\|_2\|\boldsymbol{\xi}\|_2 \le \frac{\alpha\eta\sqrt{\lambda_0}}{4}\|\boldsymbol{e}\|_2.$$

Substituting the three bounds in turn — the nuclear-norm lower bound into the second term, the remainder estimate into the third, and the prediction-change bound of Part 1 into the fourth — gives

$$\begin{aligned}
\|\boldsymbol{y}-\boldsymbol{u}(\boldsymbol{V})\|_2^2 &\le \|\boldsymbol{e}\|_2^2 - \sqrt{2}\,\alpha\eta\sqrt{\lambda_0}\|\boldsymbol{e}\|_2 + \frac{\alpha\eta\sqrt{\lambda_0}}{4}\|\boldsymbol{e}\|_2 + n\alpha^2\eta^2 && \text{[the three bounds above]}\\
&= \|\boldsymbol{e}\|_2^2 - \left(\sqrt{2}-\frac{1}{4}\right)\alpha\eta\sqrt{\lambda_0}\|\boldsymbol{e}\|_2 + n\alpha^2\eta^2 && \text{[collecting like terms]}\\
&\le \|\boldsymbol{e}\|_2^2 - \alpha\eta\sqrt{\lambda_0}\|\boldsymbol{e}\|_2 + n\alpha^2\eta^2 && [\sqrt{2}-\tfrac{1}{4} > 1].
\end{aligned}$$

So the first-order gain survives the linearization error with a constant to spare: the error can consume at most a quarter of the step, while the gain is of size $\sqrt{2}$. This proves Part 2.

Part 3. Set $s := \|\boldsymbol{e}\|_2 > 0$ and assume $\eta \le \sqrt{\lambda_0}s/(2n\alpha)$. Multiplying this step-size bound by the nonnegative quantity $n\alpha^2\eta$ converts the quadratic term of Part 2 into a linear one,

$$n\alpha^2\eta^2 \le n\alpha^2\eta\cdot\frac{\sqrt{\lambda_0}s}{2n\alpha} = \frac{\alpha\eta\sqrt{\lambda_0}s}{2},$$

so that it can only cancel half of the first-order gain. Substituting into Part 2 and factoring out s^2,

$$\|\boldsymbol{y}-\boldsymbol{u}(\boldsymbol{V})\|_2^2 \le s^2 - \alpha\eta\sqrt{\lambda_0}s + \frac{\alpha\eta\sqrt{\lambda_0}s}{2} = s^2\left(1-\frac{\alpha\eta\sqrt{\lambda_0}}{2s}\right).$$

For unit-norm inputs, $H^\infty_{ii} = 1/2$, and therefore $\lambda_0 \le \mathrm{tr}(\boldsymbol{H}^\infty)/n = 1/2$. Set $x := \alpha\eta\sqrt{\lambda_0}/(2s)$, the quantity subtracted above. The step-size bound gives $x \le \lambda_0/(4n) \le 1/8$, so $1-x \ge 0$ and the square root below is real. Moreover $(1-x/2)^2-(1-x) = x^2/4 \ge 0$ with both sides nonnegative, which is the inequality $\sqrt{1-x} \le 1-x/2$. Taking square roots in the preceding bound therefore gives

$$\begin{aligned}
\|\boldsymbol{y}-\boldsymbol{u}(\boldsymbol{V})\|_2 &\le s\sqrt{1-x}\\
&\le s\left(1-\frac{x}{2}\right) = s - \frac{\alpha\eta\sqrt{\lambda_0}}{4} && [\sqrt{1-x} \le 1-x/2]\\
&= \|\boldsymbol{y}-\boldsymbol{u}(\boldsymbol{W})\|_2 - \frac{\alpha\eta\sqrt{\lambda_0}}{4} && \text{[definition of } s].
\end{aligned}$$

This proves Part 3 and completes the proof. □

E NTK REVIEW

The Neural Tangent Kernel regime (Jacot et al., 2018) provides the canonical lazy-training framework in which the network's prediction map is well-approximated by its first-order Taylor expansion throughout training (Chizat et al., 2018; Lee et al., 2019). The proof technique relies on the observation that, in the limit of asymptotically large width, if the weights of the outer layer are scaled in an appropriate way then the weights can be initialised such that they always remain within a bounded set throughout training, allowing the "neural tangent kernel" to be treated as approximately constant. Since this is the framework our results most directly improve upon, we survey the now-substantial NTK literature in some detail, organising it along two axes: depth-2 versus deeper architectures, and gradient descent (GD) versus stochastic gradient descent (SGD). A roughly chronological account in each of the four quadrants follows; we focus on training-loss minimisation first and then on population-risk guarantees.

GD for depth-2. The earliest results in this line studied two-layer networks trained by exact gradient descent. The landscape analysis of Du & Lee (2018) had already exhibited benign optimisation behaviour for depth-2 networks with quadratic activation under modest over-parameterisation. Building on this, Du et al. (2018b) gave the first proof that GD on sufficiently over-parameterised depth-2 ReLU networks converges to a global minimum of the empirical squared loss at a linear rate. Arora et al. (2019b) subsequently sharpened the analysis with a tighter characterisation of the training-speed dependence on data and provided a network-size-independent generalisation bound. Ji & Telgarsky (2019) then dramatically reduced the width requirement for two-layer ReLU networks trained with the logistic loss for binary classification, showing that width scaling only polylogarithmic in inverse target accuracy suffices albeit at the cost of a convergence rate that is itself polynomial in the inverse accuracy.

GD, depth > 2. The first proof that GD minimises the empirical loss of *deep* over-parameterised neural networks in the NTK regime is due to Du et al. (2018a), who established linear-rate convergence for deep fully-connected, convolutional, and residual architectures by analysing the stability of the Gram matrix induced by the network. Arora et al. (2019a) complemented this with the first efficient exact algorithm for computing the NTK of convolutional networks of arbitrary depth, and showed empirically that the GD-trained convolutional NTK predictor is competitive with practically deployed architectures on standard benchmarks. Lee et al. (2019) subsequently distilled the asymptotic picture: in the infinite-width limit, deep networks trained by GD with squared loss evolve as exact linear models in the first-order Taylor expansion around initialisation, yielding closed-form predictions equivalent to a Gaussian process with a compositional kernel. Du et al. (2018c) improved over Du et al. (2018a) along two axes: they relaxed the data assumption, replacing the strict non-degeneracy condition with a milder data-non-degeneracy condition for binary cross-entropy classification, and sharpened the trajectory-length analysis to obtain faster convergence rates, while leaving the over-parameterisation requirement (the required network width as a function of n) essentially unchanged. Zou & Gu (2019) then addressed precisely this remaining bottleneck: via a tighter gradient lower bound and a sharper trajectory-length argument, they reduced the required width from a high-degree polynomial in n - e.g. $\Omega(n^{24})$ in Allen-Zhu et al. (2019b) and polynomial-in-n in Du et al. (2018c) - to a significantly milder polynomial dependence, without changing the data assumption.

SGD for depth-2. The earliest SGD guarantee in the NTK regime can be found in Li & Liang (2018), who proved that SGD on overparameterised two-layer ReLU networks generalises well provided the data is drawn from *well-separated* mixture distributions. Concretely, they assume the data for each class is a mixture of several distributional components, and that components belonging to *different* classes are separated by a minimum pairwise distance $\gamma > 0$ - components within the same class may be arbitrarily close to one another. This geometric margin condition is the analogue of linear separability for mixture-structured data: it ensures that the class boundaries are not obscured by overlap between the mixture components of opposing classes.[1] Much of the technical apparatus they introduced - in particular the observation that overparameterised network weights barely move from initialisation during training - was implicitly reused by all subsequent NTK analyses. Ji & Telgarsky

[1]Formally, Li & Liang (2018) define a (k, γ, σ^2)-mixture in which the k components are sub-Gaussian with component means separated by at least γ across classes; the separation γ plays the same role as the NTK eigenvalue λ_0 in subsequent analyses, controlling both the convergence rate and the width requirement.

(2019) additionally proved an SGD counterpart to their GD result, showing that polylogarithmic width also suffices for SGD to minimise test error on depth-2 ReLU networks trained with the logistic loss, as described in quadrant (i) above.

SGD for depth > 2. For deep networks, the first SGD training-loss-minimisation result in the NTK regime is due to Allen-Zhu et al. (2019b), who proved that SGD finds global minima of the empirical loss in polynomial time on deep fully-connected, convolutional, and residual ReLU architectures, via the almost-convex and semi-smooth structure of the loss landscape near initialisation. For population risk under SGD on deep nets, a sequence of works progressively weakened the distributional assumptions required for a generalisation guarantee. Allen-Zhu et al. (2019a) extended the analysis to the population risk by going *beyond* the first-order NTK linearisation used in Allen-Zhu et al. (2019b): their key device is a *quadratic* approximation of the network function, which allows them to prove that SGD on two- and three-layer networks with smooth activations can learn target functions realisable by a small two- or three-layer net with smooth activations. Cao & Gu (2019) then gave what is arguably the first *general* SGD risk-minimisation result in the NTK regime: they introduced the neural tangent random feature (NTRF) model - the random feature model induced by the network's gradient at initialisation - and bounded the expected $0/1$ loss of an SGD-trained deep ReLU network by the training loss of this NTRF model, yielding an $\widetilde{\mathcal{O}}(n^{-1/2})$ generalisation bound independent of network width. This is strictly more general than both the structural mixture assumptions of Li & Liang (2018) and the realisability assumptions of Allen-Zhu et al. (2019a): it requires only that the data be classifiable by the NTRF model with small error, a condition that is implied by either of those earlier assumptions but not vice versa.