MUON–ADAM TRAINING FOR OVERPARAMETERIZED NEURAL NETWORKS: CONVERGENCE AND GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon, an optimizer designed for matrix parameters, has recently demonstrated strong empirical performance in large language model pretraining. However, Muon's underlying mechanisms remains elusive, particularly in practical settings that combine it with other optimizers. In this paper, we study the convergence and generalization of joint Muon–Adam training for overparameterized ReLU networks with two layers. Specifically, we apply Muon to the hidden weight matrix and Adam to the output weights. We show that, with paired random initialization, the empirical loss converges exponentially to zero. Our analysis reveals and separately characterizes the contributions of Muon and Adam to convergence. Furthermore, we establish a generalization bound throughout training, which is determined by the kernel complexity of the training data and reduces to $\widetilde{O}(1/n)$ after sufficiently many iterations when this complexity is uniformly bounded. Our results shed light on the roles of Muon and Adam in joint neural network training from a theoretical perspective.

1 INTRODUCTION

Large language models (LLMs) have achieved strong performance across a broad range of tasks (OpenAI et al., 2023; Grattafiori et al., 2024; Gemini Team et al., 2024; Yang et al., 2025). The choice of optimizer affects the efficiency and stability of LLM pretraining (Wen et al., 2026; Khona et al., 2026) and can also influence generalization and transferability (Ruan et al., 2026). In practice, Adam (Kingma & Ba, 2015) and its variants, particularly AdamW (Loshchilov & Hutter, 2019), are widely used for LLM pretraining (Liu et al., 2025). More recently, Muon has attracted attention as an optimizer for matrix-structured parameters (Jordan et al., 2024), with empirical comparisons reporting efficiency gains over AdamW (Liu et al., 2025; Shah et al., 2025; Khona et al., 2026).

Theoretical studies of Muon have begun to characterize its convergence behavior and the solutions its updates favor. For smooth nonconvex objectives, Shen et al. (2026) and Li & Hong (2025) establish convergence rates for idealized Muon based on exact polar decomposition. Kim & Oh (2026) extend this analysis to the finite Newton-Schulz iterations used in practice. Beyond convergence, implicit bias results show that Muon favors maximum-margin classifiers w.r.t. the spectral norm in separable multiclass linear classification (Fan et al., 2025) and KKT points of margin maximization problems in smooth homogeneous networks (Gronich & Vardi, 2026). These results provide insight into why Muon can generalize well.

However, most existing theoretical analyses of Muon study its matrix updates in isolation, whereas practical training often applies Muon to weight matrices and Adam or its variants to other parameter types, including vectors (Liu et al., 2025; Shah et al., 2025; DeepSeek-AI, 2026). In joint training, Adam and Muon interact through the evolving predictor, so the optimization dynamics differ from those of either optimizer alone and cannot be inferred directly from separate analyses. Under an assumption of directional convergence, Gronich & Vardi (2026) study the implicit bias of joint Muon–Adam training in smooth homogeneous networks.

In this paper, we study convergence and generalization for an overparameterized two-layer ReLU neural network trained jointly with Adam and Muon. Adam updates the vector of output weights,

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.

while Muon updates the hidden weight matrix as a whole along a direction computed by Newton-Schulz iterations. We adopt a paired initialization (Daniely, 2020; Munteanu et al., 2022) in which the contributions of the two neurons in each pair cancel, so the network output is zero at initialization. Using the positive definiteness of the two ReLU Gram matrices to lower bound both gradient norms, we analyze Adam and Muon jointly to separate their contributions to convergence and establish parameter path bounds used to control population loss.

Contributions. Our contributions are as follows.

- **Convergence of joint Muon–Adam training.** We prove that the empirical loss of the over-parameterized two-layer ReLU network trained by joint Muon–Adam converges linearly to zero, with high probability over the paired initialization. We separate the contributions of Adam and Muon to convergence, which depend on their respective ReLU Gram matrices and, for Muon, on the finite Newton-Schulz iterations. To the best of our knowledge, this is the first convergence guarantee from initialization for joint Muon–Adam training of a nonlinear neural network. Furthermore, under a stable rank condition, we quantify how polynomial order and iteration depth of the Newton-Schulz iterations affect convergence.
- **Data-dependent generalization bound.** We establish a generalization bound for the neural network trained by joint Muon–Adam that holds with high probability for all iterates Θ_t. The bound consists of an empirical optimization term and a statistical term of order $\widetilde{O}((1+\mathfrak{C}_a)/n)$, where $\mathfrak{C}_a$ measures complexity determined by the training data. As the number of iterations increases, the empirical optimization term decays to zero. Therefore, when $\mathfrak{C}_a = O(1)$, the generalization bound is reduced to $\widetilde{O}(1/n)$.
- **Empirical evaluation.** We experimentally validate our theoretical insights on synthetic data with uniform spherical inputs and real-world data from MNIST and CIFAR-10. The observed positive ReLU Gram margins and approximately geometric empirical loss decay are consistent with our convergence analysis. On MNIST and CIFAR-10, we compare finite Newton-Schulz updates with exact SVD. These experiments distinguish progress per iteration from update time. We also observe lower test square loss as the training set grows on MNIST subsets.

2 Related Work

Muon and spectral descent. Spectral descent updates a matrix in the polar direction of its gradient. This update direction can be viewed as a steepest descent direction w.r.t. the spectral norm (Bernstein & Newhouse, 2024). It has been applied to restricted Boltzmann machines and deep networks (Carlson et al., 2015a;b). Muon optimizer uses a momentum matrix in place of the current gradient and computes the corresponding direction with Newton-Schulz iterations (Jordan et al., 2024). Analyses of Muon and related spectral methods cover smooth nonconvex optimization (Shen et al., 2026; Li & Hong, 2025), matrix factorization and in-context learning (Ma et al., 2026), LoRA-style factorization (Kang et al., 2026), and phase retrieval (Braun et al., 2026). We study the Muon update jointly with Adam on the output weights and separate their contributions to the convergence rate of the empirical loss.

Convergence of overparameterized neural networks. The neural tangent kernel (NTK) characterizes gradient descent dynamics in the infinite width limit (Jacot et al., 2018). The limiting kernel remains constant during training (Jacot et al., 2018; Yang & Littwin, 2021), and the network follows the dynamics of its linearization at initialization (Lee et al., 2019; Chizat et al., 2019). At finite width, several analyses establish convergence of gradient descent while keeping the output weights fixed (Du et al., 2019; Oymak & Soltanolkotabi, 2020; Song & Yang, 2019; Munteanu et al., 2022). In this setting, Arora et al. (2019) further relate the trajectory of network predictions to limiting kernel dynamics. When the output weights are also trained, Du et al. (2019, Theorem 3.3) prove convergence for gradient flow, while subsequent work establishes convergence for discrete gradient descent in shallow and deep networks (Nguyen, 2021; Song et al., 2021; Razborov, 2022; Tinaz et al., 2026). Beyond gradient descent, Krichene et al. (2020) analyze heavy-ball dynamics with both layers trainable in the mean-field limit. We draw on stability estimates at finite width to control the network near initialization, but derive convergence directly from the joint Muon–Adam updates rather than from gradient flow driven by the limiting NTK.

Generalization. Kernel analyses of ReLU networks can relate generalization to data-dependent complexity. For gradient descent with fixed output weights, Arora et al. (2019) express this dependence through a quadratic form involving the inverse of the limiting tangent Gram matrix. In our setting, the corresponding data-dependent complexity is $y^\top \mathbf{K}_a^{-1} y$, where $\mathbf{K}_a$ is the limiting Gram matrix of random ReLU features associated with the output weights. This quantity alone does not account for changes in the predictor along the joint Muon–Adam trajectory. For square loss, localized complexity methods relate the complexity of a function class to population risk (Liang et al., 2015; Kanade et al., 2024). We combine the data-dependent complexity with bounds on the parameter path and localization to control population square loss at every iteration.

3 PRELIMINARIES

Notations. Let q be a positive integer and write $[q] := \{1, \dots, q\}$. Denote the $q \times q$ identity matrix by I_q. For vectors $u, v \in \mathbb{R}^q$, $\langle u, v \rangle := u^\top v$ denotes their Euclidean inner product, and the Euclidean norm of u is denoted by $\|u\|_2$. Given matrices $A, B \in \mathbb{R}^{m \times d}$, $\langle A, B \rangle := \mathrm{tr}(A^\top B)$ denotes their Frobenius inner product. The Frobenius, operator, and nuclear norms of A are denoted by $\|A\|_{\mathrm{F}}$, $\|A\|_2$, and $\|A\|_*$, respectively. For nonzero $A \in \mathbb{R}^{m \times d}$, its stable rank is $\mathrm{sr}(A) := \|A\|_{\mathrm{F}}^2 / \|A\|_2^2$. If $A = U_r \Sigma_r V_r^\top$ is a compact singular value decomposition, denote the direction from its polar decomposition by $\mathrm{polar}(A) := U_r V_r^\top$. The notation $u^{\odot 2}$ denotes the coordinatewise square of u. Given nonzero $x, x' \in \mathbb{R}^d$, we write $x \parallel x'$ if $x = \alpha x'$ for some $\alpha \neq 0$, and write $x \nparallel x'$ otherwise. For nonnegative u and v, we write $u \lesssim v$ if $u \leq Cv$ for some $C > 0$, and $u \gtrsim v$ if $u \geq cv$ for some $c > 0$. For convenience, we sometimes use asymptotic notation in place of the comparison symbols. Specifically, $u = O(v)$ and $u = \Omega(v)$ mean $u \lesssim v$ and $u \gtrsim v$, respectively. We write $u = \Theta(v)$ when $u = O(v)$ and $u = \Omega(v)$ both hold. The notations $\widetilde{O}(\cdot)$, $\widetilde{\Omega}(\cdot)$ and $\widetilde{\Theta}(\cdot)$ omit polylogarithmic factors. Unless stated otherwise, hidden constants are independent of the sample size n, input dimension d, layer width m, minimum eigenvalues λ_a and λ_W, failure probability δ, and training sample $\{(x_i, y_i)\}_{i=1}^n$.

3.1 PROBLEM SETUP

In this work, we consider the following two-layer ReLU neural network:

$$f_\Theta(x) := \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \sigma(w_j^\top x),$$

where $x \in \mathbb{R}^d$ is the input, $w_1, \dots, w_m \in \mathbb{R}^d$ are the weight vectors of the first layer, $a_1, \dots, a_m \in \mathbb{R}$ are the output weights, and $\sigma(z) := \max\{0, z\}$ denotes the ReLU activation function. We define $W := (w_1, \dots, w_m)^\top \in \mathbb{R}^{m \times d}$, $a := (a_1, \dots, a_m)^\top \in \mathbb{R}^m$, and $\Theta := (W, a) \in \mathbb{R}^{m \times d} \times \mathbb{R}^m$. With a slight abuse of notation, we apply σ coordinatewise to vectors, so that the neural network can equivalently be written as $f_\Theta(x) = \frac{1}{\sqrt{m}} a^\top \sigma(Wx)$.

We are given a training set $S := \{(x_i, y_i)\}_{i=1}^n$ consisting of n i.i.d. samples from a distribution $\mathcal{D}$ over $\mathbb{R}^d \times \mathbb{R}$. Following prior work (Du et al., 2019; Arora et al., 2019; Munteanu et al., 2022), we assume for simplicity that the inputs are normalized and the labels are bounded, i.e., $\|x\|_2 = 1$ and $|y| \leq C$. In particular, these conditions hold for every training pair.

We train the neural network by minimizing the quadratic loss on S:

$$\mathcal{L}(\Theta) := \frac{1}{2n} \sum_{i=1}^{n} \left(f_\Theta(x_i) - y_i\right)^2. \tag{1}$$

The corresponding population loss is defined by

$$\mathcal{L}_\mathcal{D}(\Theta) := \frac{1}{2} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\left(f_\Theta(x) - y\right)^2\right].$$

3.2 JOINT MUON–ADAM TRAINING

To minimize the empirical loss (1), the hidden weight matrix W is updated by Muon, while the output weights a are updated by Adam. The complete joint update is presented in Algorithm 1, with $X_{t,0} = Q_K(0) = 0$ when $M_t = 0$.

Algorithm 1 Joint Muon–Adam training

Require: Initialization $\Theta_0 = (W_0, a_0)$ and total iterations $T \in \mathbb{N}$
Adam: Step sizes $\{\gamma_t\}_{t=0}^{T-1}$, momentum $\beta_1 \in [0,1)$, $\beta_2 \in [0,1)$, $\varepsilon_0 > 0$
Muon: Step sizes $\{\eta_t\}_{t=0}^{T-1}$, momentum $\beta \in [0,1)$, an integer $K \in \mathbb{N}$, order $\kappa \geq 1$

1: **Initialize** $m_{-1} \leftarrow 0,\ v_{-1} \leftarrow 0,\ M_{-1} \leftarrow 0$
2: **for** $t = 0, \ldots, T-1$ **do**
3: $\quad g_t \leftarrow \nabla_a \mathcal{L}(\Theta_t), \quad G_t \leftarrow \nabla_W \mathcal{L}(\Theta_t)$ ∘ Compute gradients
4: $\quad m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$
5: $\quad v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^{\odot 2}$
6: $\quad \widehat{m}_t \leftarrow m_t/(1-\beta_1^{t+1}), \quad \widehat{v}_t \leftarrow v_t/(1-\beta_2^{t+1})$
7: $\quad a_{t+1} \leftarrow a_t - \gamma_t \widehat{m}_t / \left(\sqrt{\widehat{v}_t} + \varepsilon_0\right)$ ▶ Adam update of a (Lines 4–7)
8: $\quad M_t \leftarrow \beta M_{t-1} + (1-\beta) G_t, \quad X_{t,0} \leftarrow M_t / \|M_t\|_{\mathrm{F}}$
9: $\quad$ **for** $k = 0, \ldots, K-1$ **do** ∘ Newton-Schulz iterations
10: $\quad\quad X_{t,k+1} \leftarrow P_\kappa(X_{t,k} X_{t,k}^\top) X_{t,k}$
11: $\quad$ **end for**
12: $\quad W_{t+1} \leftarrow W_t - \eta_t X_{t,K}$ ▶ Muon update of W (Lines 8–12)
13: **end for**
14: **return** $\Theta_T = (W_T, a_T)$

Adam. At iteration t, Adam computes exponential moving averages m_t and v_t of the output gradients $\{g_s\}_{s=0}^t$ and their coordinatewise squares, respectively (Kingma & Ba, 2015). After correcting the bias introduced by zero initialization, it updates each output weight by

$$a_{t+1,j} = a_{t,j} - \gamma_t \widehat{m}_{t,j} / \left(\sqrt{\widehat{v}_{t,j}} + \varepsilon_0\right), \qquad j \in [m].$$

Here, γ_t is the step size, $\widehat{m}_{t,j}$ denotes the corrected gradient average in coordinate j, and $\widehat{v}_{t,j}$ is the corresponding average of squared gradients. The constant $\varepsilon_0 > 0$ is added to the denominator for numerical stability.

Muon. Muon updates the hidden weight matrix W as follows (Jordan et al., 2024):

$$M_t = \beta M_{t-1} + (1-\beta) G_t,$$
$$W_{t+1} = W_t - \eta_t Q_K(M_t).$$

Muon first computes the momentum matrix M_t using the hidden gradient G_t. It then applies K Newton-Schulz iterations to obtain $Q_K(M_t)$, which approximates the polar factor $\mathrm{polar}(M_t) := U_r V_r^\top$, where $M_t = U_r \Sigma_r V_r^\top$ is a compact SVD (Bernstein & Newhouse, 2024; Kim & Oh, 2026). When the matrix is a gradient, the negative of the direction supplied by its polar decomposition is steepest with respect to the spectral norm (Bernstein & Newhouse, 2024).

We next introduce the Newton-Schulz polynomial P_κ and the corresponding Newton-Schulz iteration for computing $Q_K(M_t)$.

Definition 3.1 (Kim & Oh, 2026)**.** Let order $\kappa \geq 1$ be an integer. The Newton-Schulz polynomial $P_\kappa : \mathbb{R} \to \mathbb{R}$ takes the form

$$P_\kappa(s) := \sum_{j=0}^{\kappa} \frac{\binom{2j}{j}}{4^j} (1-s)^j. \tag{2}$$

Given a nonzero matrix $A \in \mathbb{R}^{m \times d}$ and an integer $K \geq 0$, we compute $Q_K(A)$ using the following Newton-Schulz iterations:

$$Q_{k+1}(A) := P_\kappa\big(Q_k(A) Q_k(A)^\top\big)\, Q_k(A), \qquad k = 0, \ldots, K-1, \tag{3}$$

where $Q_0(A) = A / \|A\|_{\mathrm{F}}$.

Let $A = U_r \Sigma_r V_r^\top$ be a compact SVD. Each step of the Newton-Schulz iteration (3) preserves U_r and V_r and maps each singular value s to $sP_\kappa(s^2)$. Frobenius normalization places the nonzero singular values of $Q_0(A)$ in $(0, 1]$. With the Taylor polynomial (2), these singular values converge monotonically to one (Kim & Oh, 2026). Consequently, we obtain $Q_K(A) \to \mathrm{polar}(A)$ as $K \to \infty$.

3.3 Gram Matrices and Initialization

Let $X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n \times d}$ denote the input matrix of the training set S. We define the following Gram matrices induced by the ReLU activation and random initialization:

$$\mathbf{K}_a := \mathbb{E}_{u \sim \mathcal{N}(0, I_d)}[\sigma(Xu)\sigma(Xu)^\top], \qquad \mathbf{K}_W := \mathbb{E}_{u \sim \mathcal{N}(0, I_d)}[D(u)XX^\top D(u)], \tag{4}$$

where $D(u) := \mathrm{diag}\big(\mathbf{1}\{u^\top x_i > 0\} : i \in [n]\big)$. We make the following assumption on these matrices.

Assumption 3.1. Define $\lambda_a := \lambda_{\min}(\mathbf{K}_a)$ and $\lambda_W := \lambda_{\min}(\mathbf{K}_W)$. We assume $\lambda_0 := \min\{\lambda_a, \lambda_W\} > 0$.

Assumption 3.1 requires the Gram matrices $\mathbf{K}_a$ and $\mathbf{K}_W$ to be positive definite, with $0 < \lambda_0 \le 1/2$ under our input normalization. It is standard in convergence and generalization analyses of overparameterized two-layer neural networks (Du et al., 2019; Arora et al., 2019; Oymak & Soltanolkotabi, 2020). A sufficient condition for this assumption is that the training inputs are pairwise nonparallel.

Proposition 3.1 (Du et al. (2019))**.** *If $x_i \nparallel x_j$ for every $i \neq j$, then $\lambda_0 > 0$.*

We initialize the hidden neurons in pairs, with identical weight vectors and opposite output weights within each pair (Daniely, 2020; Munteanu et al., 2022).

Assumption 3.2. Let the hidden layer width m be even and set $p = m/2$ [1]. We initialize the parameters by setting

$$w_{0,2j-1} = w_{0,2j} = u_j, \qquad a_{0,2j-1} = s_j, \qquad a_{0,2j} = -s_j, \qquad j \in [p],$$

where $u_j \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$ and $s_j \overset{\text{iid}}{\sim} \mathrm{Unif}\{-1, +1\}$. Here, $w_{t,j} \in \mathbb{R}^d$ denotes the weight vector of hidden neuron j at iteration t, and $a_{t,j} \in \mathbb{R}$ is the corresponding output weight.

The pair initialization ensures $f_{\Theta_0}(x) = 0$ for every x, but does not force the hidden gradient to vanish. We then jointly train a with Adam and W with Muon, as in Algorithm 1.

4 Main Results

In this section, we analyze the convergence and generalization of joint Muon–Adam training.

4.1 Convergence Analysis of Muon–Adam Training

We first analyze Algorithm 1 with a loss-dependent learning rate of Muon. Theorem 4.1 establishes linear convergence of the empirical loss and separates the contributions of Adam and Muon to the convergence rate. Under a uniform stable-rank condition, Proposition 4.1 makes the dependence of the Muon contribution on the polynomial order and Newton-Schulz depth explicit. The detailed proof is deferred to Appendix B.

We set $\mathcal{B} := [P_\kappa(0)]^K$ for the subsequent analysis. For a nonzero matrix B, define

$$\Phi_{\kappa+1,K}(B) := \sum_{i=1}^{\mathrm{rank}(B)} s_i(B)\sqrt{1 - \{1 - s_i(B)^2\}^{(\kappa+1)^K}}, \tag{5}$$

where $s_i(B) := \sigma_i(B)/\|B\|_{\mathrm{F}}$, $i = 1, \ldots, \mathrm{rank}(B)$, are the singular values of B normalized by its Frobenius norm. The quantity $\Phi_{\kappa+1,K}(B)$ quantifies the Muon contribution to the convergence rate.

[1] For odd m, apply the paired initialization to the first $m - 1$ neurons and initialize the remaining neuron at zero. The remaining neuron stays at zero throughout training, so the conclusions remain valid for odd widths up to numerical constants.

Theorem 4.1. *Fix $\delta \in (0,1)$ and suppose Assumptions 3.1 and 3.2 hold. We assume $\beta_1^2 < \beta_2$, and let $(a_t, W_t)_{t\geq 0}$ be generated by Algorithm 1 with positive base step sizes γ, η satisfying $\gamma_t \equiv \gamma = O\left(\frac{\varepsilon_0\sqrt{\lambda_0}}{(1+\sqrt{\mathcal{B}})\sqrt{n}}\right)$ and $\eta_t := \eta\sqrt{2\mathcal{L}(\Theta_t)}$, where $\eta = O\left(\frac{\sqrt{\lambda_0}}{\mathcal{B}(1+\sqrt{\mathcal{B}})\sqrt{n}}\right)$, with sufficiently small implicit constants in the two upper bounds. If $m \gtrsim \frac{n^4}{\lambda_0^4}\log^2\left(\frac{n}{\lambda_0\delta}\right)$, then with probability at least $1-\delta$ over the random initialization, we have for every $T \geq 0$,*

$$\mathcal{L}(\Theta_T) \leq \mathcal{L}(\Theta_0)\exp\left[-\Omega\left(\frac{\gamma\lambda_a}{n}T + \eta\sqrt{\frac{\lambda_W}{n}}\sum_{t=0}^{T-1}\Phi_{\kappa+1,K}(M_t)\right)\right]. \tag{6}$$

Theorem 4.1 shows that, when the network width m is sufficiently large, the empirical loss of the neural network trained by joint Muon–Adam converges linearly to zero with high probability. By Lemma D.1, we have $\Phi_{\kappa+1,K}(M_t) \geq 1$. Substituting this estimate into (6), we obtain the simpler bound

$$\mathcal{L}(\Theta_T) \leq \mathcal{L}(\Theta_0)\exp\left[-\Omega\left(\frac{\gamma\lambda_a}{n} + \eta\sqrt{\frac{\lambda_W}{n}}\right)T\right].$$

Furthermore, note that the Muon learning rate $\eta_t = \eta\sqrt{2\mathcal{L}(\Theta_t)}$ decreases as the empirical loss falls. This scaling limits changes in the ReLU features caused by updating W while retaining an explicit Muon contribution to the loss contraction. This dependence on the loss is not necessary for convergence to zero. Theorem B.1 proves exponential convergence with a geometrically decaying Muon learning rate $\eta_t = \eta_0\rho^t$ chosen before training.

Contributions of Adam and Muon. The bound in (6) separates the contributions of Adam and Muon to empirical loss decay. The contribution $\gamma\lambda_a T/n$ arises from Adam updates to the output weights, whereas $\eta\sqrt{\lambda_W/n}\sum_{t=0}^{T-1}\Phi_{\kappa+1,K}(M_t)$ arises from Muon updates to the weight matrix.

Network width m. The sufficient width in Theorem 4.1 is $m \gtrsim n^4\lambda_0^{-4}\log^2\left(\frac{n}{\lambda_0\delta}\right)$. The quartic dependence on n in our proof arises from controlling ReLU activation changes and momentum tracking during joint training. This sufficient condition may not be optimal. We note that gradient descent analyses obtain quadratic dependence on n, up to kernel and logarithmic factors, with fixed output weights and coupled initialization (Munteanu et al., 2022), or with both layers trained under LeCun initialization and additional data assumptions (Nguyen, 2021). Motivated by these results, we leave improving the dependence on n in our width requirement to future work.

Effect of Newton-Schulz iterations. The factor $\Phi_{\kappa+1,K}(M_t)$ captures the dependence of the Muon term in (6) on the polynomial order and Newton-Schulz depth. If $M_t \neq 0$, then $\Phi_{\kappa+1,0}(M_t) = 1$, and $\Phi_{\kappa+1,K}(M_t)$ is nondecreasing in K. When $\mathrm{rank}(M_t) > 1$, it increases strictly with K and, for $K \geq 1$, with κ, see Lemma D.2. However, increasing K or κ also enlarges $\mathcal{B} = [P_\kappa(0)]^K$, which restricts learning rates γ and η.

Under the additional assumption that $\mathrm{sr}(M_t) \geq 1+\nu$ for all $t \geq 0$, the next proposition gives an upper bound in which the Muon term depends explicitly on κ and K.

Proposition 4.1. *Under the conditions of Theorem 4.1, assume that, for some $\nu > 0$, the training dynamics satisfy $\mathrm{sr}(M_t) \geq 1+\nu$ for all $t \geq 0$. Then, with probability at least $1-\delta$ over the random initialization, we have for every $T \geq 0$,*

$$\mathcal{L}(\Theta_T) \leq \mathcal{L}(\Theta_0)\exp\left[-\Omega\left(\frac{\gamma\lambda_a}{n} + \eta\sqrt{\frac{\lambda_W}{n}}\sqrt{(1+\nu)\left[1-\left(\frac{\nu}{1+\nu}\right)^{(\kappa+1)^K}\right]}\right)T\right]. \tag{7}$$

The condition $\mathrm{sr}(M_t) \geq 1+\nu$ requires the squared singular values beyond the largest to sum to at least ν times the largest squared singular value at every iteration. For $\nu > 0$ and any integer $\kappa \geq 1$, the square-root factor multiplying $\eta\sqrt{\lambda_W/n}$ in (7) increases strictly from 1 at $K = 0$ to $\sqrt{1+\nu}$ as $K \to \infty$. Thus, additional Newton-Schulz iterations increase the square-root multiplier of the contribution of Muon.

4.2 GENERALIZATION

We establish a generalization bound in terms of the empirical loss and a data-dependent complexity associated with $\mathbf{K}_a$. The detailed proof is deferred to Appendix C.

Let $y := (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ denote the vector of labels in the training set S. We define the data-dependent complexity associated with the Gram matrix $\mathbf{K}_a$ by

$$\mathfrak{C}_a := y^\top \mathbf{K}_a^{-1} y.$$

To establish a generalization guarantee, we make the following assumption on the data distribution $\mathcal{D}$.

Assumption 4.1. Let $\lambda \in (0, 1/2]$ and $\rho \in (0, 1)$. We assume that a training set $S \sim \mathcal{D}^n$ satisfies

$$\mathbb{P}(\lambda_0(X) \geq \lambda) \geq 1 - \rho.$$

Assumption 4.1 relaxes the requirement of Assumption 3.1 by allowing the Gram condition to fail on an event of probability at most ρ. Similar high-probability nondegeneracy conditions appear in Arora et al. (2019) and Yang et al. (2024).

The next theorem provides a generalization guarantee for the neural network trained by Algorithm 1.

Theorem 4.2. *Fix $\delta \in (0, 1)$ and suppose Assumptions 3.2 and 4.1 hold with $\lambda \in (0, 1/2]$ and $\rho = \delta/3$. We assume $\beta_1^2 < \beta_2$, and let $(a_t, W_t)_{t\geq 0}$ be generated by Algorithm 1 with step sizes $\gamma_t \equiv \gamma = \Theta\Big(\frac{\varepsilon_0\sqrt{\lambda}}{(1+\sqrt{\mathcal{B}})\sqrt{n}}\Big)$ and $\eta_t := \eta\sqrt{2\mathcal{L}(\Theta_t)}$, where $\eta = \Theta\Big(\frac{\gamma\lambda^{3/2}}{\mathcal{B}(1+\varepsilon_0)n^{3/2}}\Big)$. If $m \gtrsim \Big(d + \frac{n^4}{\lambda^4}\Big)\log^2\big(\frac{n+d}{\lambda\delta}\big)$, then with probability at least $1-\delta$ over the training sample S and the random initialization, we have for every $T \geq 0$,*

$$\mathcal{L}_{\mathcal{D}}(\Theta_T) \lesssim \left[\frac{\mathcal{L}(\Theta_T)}{\lambda} + \frac{1 + \mathfrak{C}_a}{n}\right] \log^3\left(\frac{(n+d)m}{\lambda\delta}\right). \tag{8}$$

Generalization bound. The bound in (8) separates the empirical optimization error from the data-dependent complexity term. Under the conditions of Theorem 4.2, Theorem 4.1 shows that $\mathcal{L}(\Theta_T)$ decays exponentially to 0. As $T \to \infty$, the bound is therefore governed by $(1 + \mathfrak{C}_a)/n$, up to the logarithmic factor.

Data-dependent complexity. The statistical term in (8) scales as $(1 + \mathfrak{C}_a)/n$, up to logarithmic factors. On the event in Assumption 4.1, bounded labels imply $\mathfrak{C}_a \leq \|y\|_2^2/\lambda \lesssim n/\lambda$, which does not ensure that the statistical term in (8) vanishes as n grows. To bound the statistical term by $\widetilde{O}(1/n)$, we consider data whose labels satisfy $y_i = f_\star(x_i)$. For unit norm inputs with $\mathbf{K}_a \succ 0$, the following choices of $f_\star$ admit explicit bounds on $\mathfrak{C}_a$:

- *Quadratic functions.* If $f_\star(x_i) = c + b^\top x_i + x_i^\top B x_i$, where $c \in \mathbb{R}$, $b \in \mathbb{R}^d$, and $B = B^\top \in \mathbb{R}^{d\times d}$, then $\mathfrak{C}_a \leq 2\pi c^2 + 4\|b\|_2^2 + 4\pi\|B\|_{\mathrm{F}}^2$ by Lemma C.2. $|c| + \|b\|_2 + \|B\|_{\mathrm{F}} \leq C$ ensures bounded labels and $\mathfrak{C}_a = O(1)$.
- *Linear functions.* The preceding quadratic form reduces to $f_\star(x) = b^\top x$ when $c = B = 0$. Its complexity bound then becomes $\mathfrak{C}_a \leq 4\|b\|_2^2 = O(1)$.
- *Finite sums of ridge monomials.* Let $f_\star(x) = \sum_{r=1}^R \alpha_r (u_r^\top x)^{p_r}$, where $\alpha_r \in \mathbb{R}$, $u_r \in \mathbb{R}^d$, and $p_r \in \{0, 1, 2, 4, 6, \dots\}$. If R and the degrees p_r are fixed and $|\alpha_r|\,\|u_r\|_2^{p_r} = O(1)$ for all r, then $\mathfrak{C}_a = O(1)$ by Lemma C.1.

Uniform spherical inputs. One example satisfying Assumption 4.1 is a distribution with uniform spherical inputs under the dimensional condition below. Consider a training set $S = \{(x_i, y_i)\}_{i=1}^n$ consisting of n i.i.d. samples from a distribution $\mathcal{D}$ over $\mathbb{S}^{d-1} \times [-C, C]$, where the inputs are uniformly distributed on the unit sphere $\mathbb{S}^{d-1}$. If $d \gtrsim n + \log(1/\delta)$, then Assumption 4.1 holds with a numerical $\lambda > 0$ and $\rho = \delta/3$, see Proposition C.2. Consequently, we obtain the following generalization guarantee for uniform spherical inputs.

Proposition 4.2. *Fix $\delta \in (0, 1)$. Let $S = \{(x_i, y_i)\}_{i=1}^n$ be the training sample with $(x_i, y_i) \overset{\text{i.i.d.}}{\sim} \mathcal{D}$, where $x \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and $|y| \leq C$ for $(x, y) \sim \mathcal{D}$. Suppose Assumption 3.2 holds. We assume $\beta_1^2 <$*

β_2, and let $(a_t, W_t)_{t\geq 0}$ be generated by Algorithm 1 with step sizes $\gamma_t \equiv \gamma = \Theta\Big(\frac{\varepsilon_0}{(1+\sqrt{\mathcal{B}})\sqrt{n}}\Big)$ and $\eta_t := \eta\sqrt{2\mathcal{L}(\Theta_t)}$, where $\eta = \Theta\Big(\frac{\gamma}{\mathcal{B}(1+\varepsilon_0)n^{3/2}}\Big)$. If $d \gtrsim n + \log(1/\delta)$ and $m \gtrsim (d+n^4)\log^2\big(\frac{n+d}{\delta}\big)$, then with probability at least $1-\delta$ over the training sample S and the random initialization, we have for every $T \geq 0$,

$$\mathcal{L}_{\mathcal{D}}(\Theta_T) \lesssim \left[\mathcal{L}(\Theta_0)\exp\left[-\Omega\left(\frac{\gamma T}{n} + \frac{\eta}{\sqrt{n}}\sum_{t=0}^{T-1}\Phi_{\kappa+1,K}(M_t)\right)\right] + \frac{1+\mathfrak{C}_a}{n}\right]\log^3\left(\frac{(n+d)m}{\delta}\right).$$

The proof is given in Appendix C.3.

5 Experiments

In this section, we study two settings: uniform spherical inputs and real datasets. All training experiments use the two-layer ReLU neural network introduced in Section 3. For uniform spherical inputs, we measure the margins of the ReLU Gram matrices in Assumption 3.1 and track the empirical loss under joint training. On MNIST digits 3 versus 5 and CIFAR-10 cats versus dogs, we compare finite Newton-Schulz updates with exact SVD in terms of training loss and update time. We then assess how test performance on MNIST changes with the training set size. Further details and additional results appear in Appendix A.

Uniform Spherical Inputs. For each $n \in \{32, 64, 96, 128, 192, 256, 384\}$ and $d/n \in \{1, 2, 4\}$, we generate 100 independent input matrices by normalizing Gaussian vectors to the unit sphere. We compute the two limiting ReLU Gram matrices using (4). Table 1 in Appendix A reports the medians of their smallest eigenvalues. All 2100 draws satisfy $0 < \lambda_a < \lambda_W$, so $\lambda_0 = \lambda_a$ throughout. Across the grid, the median values of λ_0 range from 0.0768 to 0.1628. When $d = n$, the median smallest eigenvalue of $XX^\top$ is near zero, whereas the ReLU Gram matrices have substantially larger minimum eigenvalues.

We also train the network on uniform spherical inputs with $(n, d, m) = (64, 128, 512)$. Figure 3 in Appendix A shows approximately geometric empirical loss decay within the prespecified relative-loss interval $[10^{-4}, 10^{-1}]$. The observed margins and loss decay are consistent with the analysis.

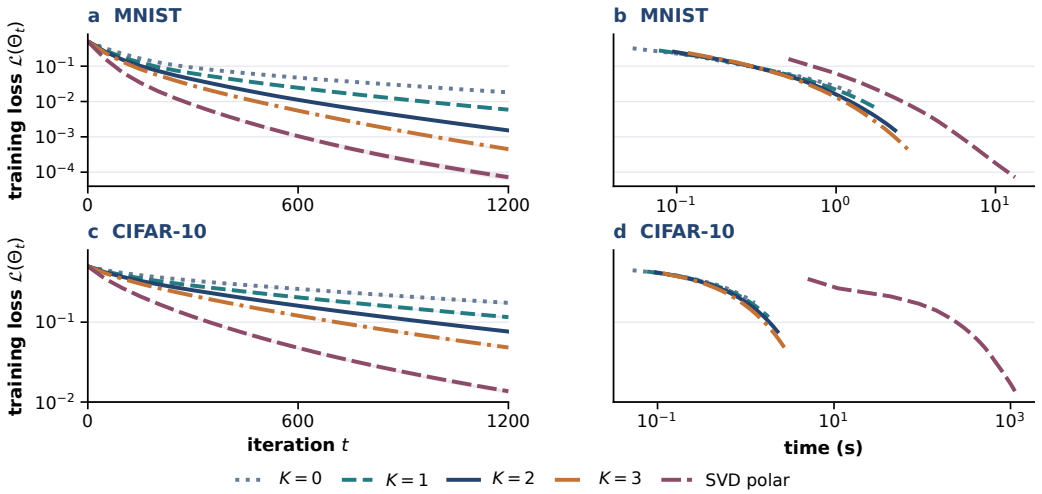

Figure 1: Training loss on MNIST 3 versus 5 (a,b) and CIFAR-10 cats versus dogs (c,d), with $n = 256$, $m = 512$, and three paired seeds. Panels (a,c) plot median loss against completed updates, whereas panels (b,d) plot median loss against median synchronized update time at matched checkpoints. All methods complete 1200 updates.

Real Datasets. We next consider two binary image tasks: MNIST digits 3 versus 5 (LeCun et al., 1998) and CIFAR-10 cats versus dogs (Krizhevsky, 2009). For the optimization comparison, we use $n = 256$, width $m = 512$, Newton-Schulz depths $K \in \{0, 1, 2, 3\}$, and an exact polar direction computed by SVD. Within each of three seeds, all methods share the same training subset and initialization and complete 1200 updates. Figure 1 plots training loss against both update count and synchronized update time. At matched update counts, larger K generally leads to lower training loss, and exact SVD has the lowest loss on both datasets. Its advantage per update does not necessarily translate into shorter update time. On MNIST, the median first recorded crossing of 10^{-2} times the initial training loss occurs at 650 updates and 1.51 s for $K = 3$. Exact SVD reaches the same threshold in 400 updates but takes 3.89 s.

To assess test performance on MNIST, we fix $K = 2$ and vary the training size over $n \in \{128, 256, 512\}$. Each run completes 1200 updates and uses the same test partition. Figure 2 reports test square loss and classification error at the final checkpoint. As n grows from 128 to 512, the median test loss falls from 0.110 to 0.0666, while classification error falls from 4.94% to 2.31%.

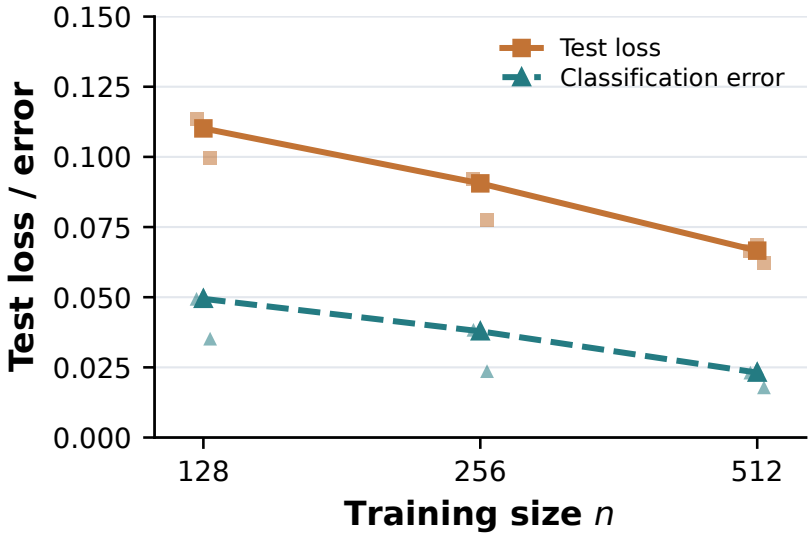

Figure 2: Test performance on MNIST 3 versus 5 after 1200 updates with $K = 2$ and $m = 512$. Test loss uses the square-loss normalization in (1), and classification error is reported as a fraction. Each line traces the median over three seeds as training size varies, while lighter markers show individual runs.

6 Conclusion

We proved that joint Muon–Adam training of sufficiently wide two-layer ReLU neural networks drives the empirical loss to zero at a linear rate. Our analysis separates the contributions of Adam and Muon while retaining Muon's momentum and finite Newton-Schulz update. We also provided a generalization bound on the population loss throughout training. When the data-dependent complexity remains bounded, this bound reaches $\widetilde{O}(1/n)$ after sufficient optimization. Furthermore, we experimentally validate our theoretical insights on synthetic data with uniform spherical inputs and real-world data from MNIST and CIFAR-10. A future direction is to reduce the quartic dependence on n in our width condition, motivated by the quadratic dependence proved for gradient descent with coupled initialization (Munteanu et al., 2022). Besides, extending our theory to deep ReLU networks (Nguyen, 2021) and Transformer architectures (Vaswani et al., 2017; Liu et al., 2025), while covering stochastic updates, remains an open challenge, which we will explore in future work.

AI USE STATEMENT

In this work, we used generative AI tools to aid manuscript writing, including organization and language editing, and to retrieve and identify related work. We also consulted these tools while refining mathematical claims, proofs, and the experimental design. The authors reviewed all AI-assisted material, and take responsibility for the final manuscript.

REFERENCES

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 322–332. PMLR, 2019.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher Complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Jeremy Bernstein and Laker Newhouse. Old Optimizer, New Norm: An Anthology. In *OPT 2024: Workshop on Optimization for Machine Learning*, 2024.

Guillaume Braun, Han Bao, Wei Huang, and Masaaki Imaizumi. Spectral Gradient Descent Mitigates Anisotropy-Driven Misalignment: A Case Study in Phase Retrieval. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

David Carlson, Volkan Cevher, and Lawrence Carin. Stochastic Spectral Descent for Restricted Boltzmann Machines. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 111–119. PMLR, 2015a.

David E. Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned Spectral Descent for Deep Learning. In *Advances in Neural Information Processing Systems 28*, pp. 2971–2979, 2015b.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On Lazy Training in Differentiable Programming. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.

Amit Daniely. Neural Networks Learning and Memorization with (Almost) No Over-Parameterization. In *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc., 2020.

DeepSeek-AI. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence, 2026. URL `https://arxiv.org/abs/2606.19348`. arXiv preprint arXiv:2606.19348.

Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient Descent Provably Optimizes Over-Parameterized Neural Networks. In *The Seventh International Conference on Learning Representations*, 2019.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit Bias of Spectral Descent and Muon on Multiclass Separable Data. In *Advances in Neural Information Processing Systems 38*, pp. 39622–39669. Curran Associates, Inc., 2025.

Gemini Team, Petko Georgiev, et al. Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context, 2024. URL `https://arxiv.org/abs/2403.05530`. arXiv preprint arXiv:2403.05530.

Aaron Grattafiori et al. The Llama 3 Herd of Models, 2024. URL `https://arxiv.org/abs/2407.21783`. arXiv preprint arXiv:2407.21783.

Eitan Gronich and Gal Vardi. The Implicit Bias of Adam and Muon on Smooth Homogeneous Neural Networks. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems 31*, pp. 8571–8580. Curran Associates, Inc., 2018.

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.

Varun Kanade, Patrick Rebeschini, and Tomas Vaškevičius. Exponential Tail Local Rademacher Complexity Risk Bounds Without the Bernstein Condition. *Journal of Machine Learning Research*, 25(388):1–43, 2024.

Changmin Kang, Jihun Yun, Baekrok Shin, Yeseul Cho, and Chulhee Yun. Uniform Spectral Growth and Convergence of Muon in LoRA-Style Matrix Factorization, 2026. URL https://arxiv.org/abs/2602.06385. arXiv preprint arXiv:2602.06385.

Mikail Khona, Aditya Vavre, Boxiang Wang, Deyu Fu, Hao Wu, Mike Chrzanowski, Bryan Catanzaro, Dheevatsa Mudigere, Jeff Pool, Michael Lightstone, Mohammad Shoeybi, Mostofa Patwary, Nima Tajbakhsh, and Tijmen Blankevoort. SOAP, Muon, and Beyond: Pushing LLM Pretraining Scales, 2026. URL https://arxiv.org/abs/2607.20548. arXiv preprint arXiv:2607.20548.

Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton-Schulz. In *The Fourteenth International Conference on Learning Representations*, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *The Third International Conference on Learning Representations*, 2015.

Walid Krichene, Kenneth F. Caluya, and Abhishek Halder. Global Convergence of Second-Order Dynamics in Two-Layer Neural Networks, 2020. URL https://arxiv.org/abs/2007.06852. arXiv preprint arXiv:2007.06852.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, Department of Computer Science, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. In *Advances in Neural Information Processing Systems 32*, 2019.

Jiaxiang Li and Mingyi Hong. A Note on the Convergence of Muon, 2025. URL https://arxiv.org/abs/2502.02900. arXiv preprint arXiv:2502.02900.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with Square Loss: Localization through Offset Rademacher Complexity. In *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1260–1285. PMLR, 2015.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is Scalable for LLM Training, 2025. URL https://arxiv.org/abs/2502.16982. arXiv preprint arXiv:2502.16982.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning Benefits of Spectral Orthogonalization in Muon, 2026. URL https://arxiv.org/abs/2601.13474. arXiv preprint arXiv:2601.13474.

Alexander Munteanu, Simon Omlor, Zhao Song, and David Woodruff. Bounding the width of neural networks via coupled initialization: A worst case analysis. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16083–16122. PMLR, 2022.

Quynh Nguyen. On the Proof of Global Convergence of Gradient Descent for Deep ReLU Networks with Linear Widths. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8056–8062. PMLR, 2021.

OpenAI et al. GPT-4 Technical Report, 2023. URL https://arxiv.org/abs/2303.08774. arXiv preprint arXiv:2303.08774.

Samet Oymak and Mahdi Soltanolkotabi. Toward Moderate Overparameterization: Global Convergence Guarantees for Training Shallow Neural Networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Alexander Razborov. Improved Convergence Guarantees for Shallow Neural Networks, 2022. URL https://arxiv.org/abs/2212.02323. arXiv preprint arXiv:2212.02323.

Tianyu Ruan, Fengzhuo Zhang, Shuche Wang, and Shihua Zhang. Muon Learns More Robust and Transferable Features than Adam, 2026. URL https://arxiv.org/abs/2606.09658. arXiv preprint arXiv:2606.09658.

Ishaan Shah, Anthony M. Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J. Shah, Khoi Nguyen, Kurt Smith, Michael Callahan, Michael Pust, Mohit Parmar, Peter Rushton, Platon Mazarakis, Ritvik Kapila, Saurabh Srivastava, Somanshu Singla, Tim Romanski, Yash Vanjani, and Ashish Vaswani. Practical Efficiency of Muon for Pretraining, 2025. URL https://arxiv.org/abs/2505.02222. arXiv preprint arXiv:2505.02222.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the Convergence Analysis of Muon. *Transactions on Machine Learning Research*, 2026.

ChaeHwan Song, Ali Ramezani-Kebrya, Thomas Pethick, Armin Eftekhari, and Volkan Cevher. Subquadratic Overparameterization for Shallow Neural Networks. In *Advances in Neural Information Processing Systems 34*, pp. 11247–11259. Curran Associates, Inc., 2021.

Zhao Song and Xin Yang. Quadratic Suffices for Over-Parametrization via Matrix Chernoff Bound, 2019. URL https://arxiv.org/abs/1906.03593. arXiv preprint arXiv:1906.03593.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, Low Noise and Fast Rates. In *Advances in Neural Information Processing Systems 23*, pp. 2199–2207, 2010.

Berk Tinaz, Changzhi Xie, and Mahdi Soltanolkotabi. When Both Layers Learn: Training Dynamics of Representing Linear Models via ReLU Networks. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 6310–6371. PMLR, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Number 47 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. ISBN 9781108231596.

Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic Pretraining Optimizers and Where to Find Them. In *International Conference on Learning Representations*, 2026.

An Yang et al. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388. arXiv preprint arXiv:2505.09388.

Greg Yang and Etai Littwin. Tensor Programs IIb: Architectural Universality of Neural Tangent Kernel Training Dynamics. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11762–11772. PMLR, 2021.

Hongru Yang, Ziyu Jiang, Ruizhe Zhang, Yingbin Liang, and Zhangyang Wang. Neural Networks with Sparse Activation Induced by Large Bias: Tighter Analysis with Bias-Generalized NTK. *Journal of Machine Learning Research*, 25(321):1–51, 2024.

APPENDIX

A Experimental Details **13**

B Proofs of Convergence Results **14**

B.1 Proofs of Theorem 4.1 and Proposition 4.1 14

B.2 Convergence with a Geometrically Decaying Muon Learning Rate 26

C Proofs of Generalization Results **28**

C.1 Proof of Theorem 4.2 28

C.2 Examples with Bounded Data-Dependent Complexity 30

C.3 Uniform Spherical Inputs 32

D Technical Lemmas **32**

D.1 Newton-Schulz Iterations 32

D.2 Gram Matrices and Optimizer Updates 34

D.3 Rademacher Complexity and Generalization Bounds 36

A EXPERIMENTAL DETAILS

Table 1: Smallest eigenvalues for independent uniform spherical samples. Each entry is the median over 100 draws at the indicated (n, d). The $XX^\top$ row for $d/n = 1$ is multiplied by 10^5 for readability.

d/n	Quantity	$n = 32$	64	96	128	192	256	384
1	λ_a	.0768	.0821	.0838	.0849	.0863	.0870	.0879
	λ_W	.2135	.2273	.2317	.2351	.2385	.2403	.2426
	λ_0	.0768	.0821	.0838	.0849	.0863	.0870	.0879
	$10^5\lambda_{\min}(XX^\top)$	34.45	7.679	3.927	2.003	0.6475	0.5865	0.2959
2	λ_a	.1172	.1159	.1143	.1137	.1135	.1137	.1130
	λ_W	.2705	.2710	.2713	.2710	.2715	.2719	.2715
	λ_0	.1172	.1159	.1143	.1137	.1135	.1137	.1130
	$\lambda_{\min}(XX^\top)$	.1121	.1012	.0959	.0931	.0922	.0921	.0892
4	λ_a	.1628	.1583	.1579	.1566	.1560	.1552	.1546
	λ_W	.3190	.3156	.3158	.3149	.3147	.3140	.3135
	λ_0	.1628	.1583	.1579	.1566	.1560	.1552	.1546
	$\lambda_{\min}(XX^\top)$	.2858	.2691	.2662	.2619	.2600	.2574	.2546

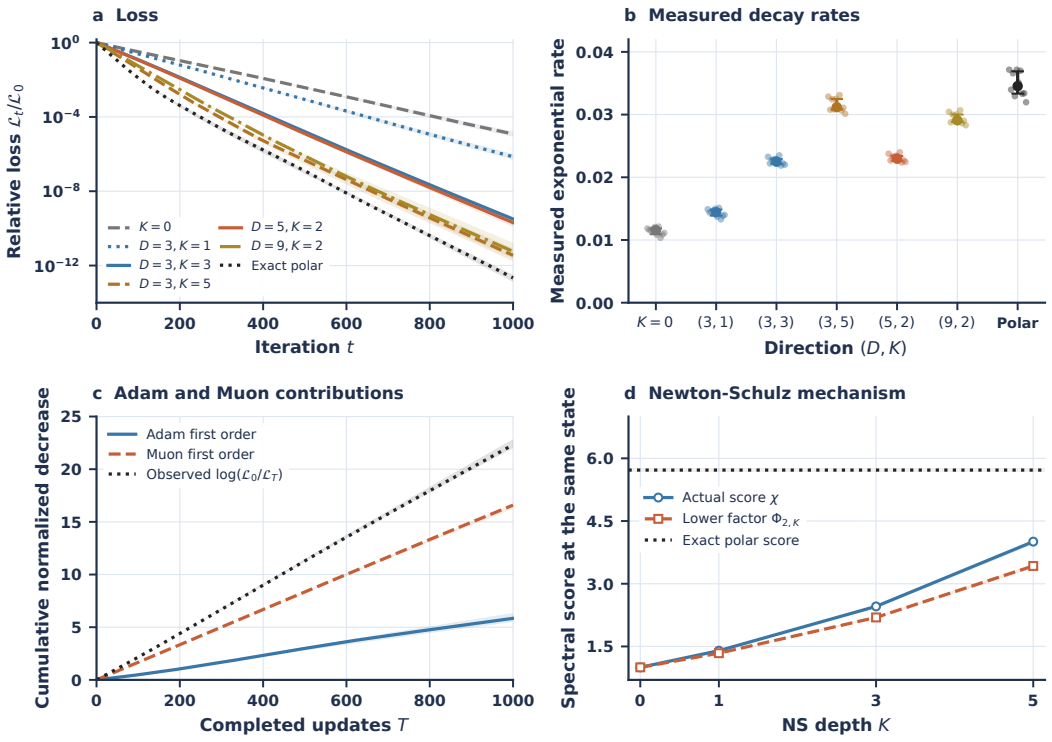

Figure 3: Convergence and optimizer contributions over ten paired seeds. (a) Relative empirical loss. (b) Exponential decay rates fitted over the prespecified relative-loss interval $[10^{-4}, 10^{-1}]$, with individual seed estimates, medians, and interquartile ranges. (c) First-order contributions from Adam and Muon, normalized by the loss at each iteration and then accumulated, alongside the observed logarithmic loss reduction for $D = 5, K = 2$. (d) Spectral scores and lower factors from the cubic Newton-Schulz recursion at depths $K \in \{0, 1, 3, 5\}$, evaluated at the same $D = 5, K = 2$ states at $t = 100$, with the exact polar score as a reference. Shaded bands show interquartile ranges across seeds.

B Proofs of Convergence Results

In this section, we first prove Theorem 4.1 and Proposition 4.1. We then state and prove the convergence result for a geometrically decaying Muon learning rate.

B.1 Proofs of Theorem 4.1 and Proposition 4.1

The main difficulty is that the update directions of Adam and Muon need not align with the current gradients. We analyze a Lyapunov function that combines the empirical loss with a weighted squared norm of the Adam momentum. For Muon, we bound the difference between its corrected momentum and the current hidden gradient to establish a descent bound for the Newton-Schulz direction.

The proof of Theorem 4.1 is organized into five steps: The first four establish local gradient and loss bounds, analyze the Adam and Muon updates separately, and prove contraction of the Lyapunov function. In the final step, we combine these estimates with trajectory bounds to verify that the local conditions hold throughout training and complete the proof by induction. The required Gram matrix properties and Newton-Schulz inequalities are collected in Appendix D. We then derive Proposition 4.1 from the resulting convergence rate.

B.1.1 Local Gradient and Loss Bounds

Let $Z_X(W)_{ij} := \sigma(w_j^\top x_i)$ and $D_{X,j}(W) := \mathrm{diag}\{\mathbf{1}(w_j^\top x_i > 0) : i \in [n]\}$. The empirical Gram matrices for the output and hidden weights take the form

$$\widehat{\mathbf{K}}_a(a, W) := \frac{1}{m} Z_X(W) Z_X(W)^\top,$$
$$\widehat{\mathbf{K}}_W(a, W) := \frac{1}{m} \sum_{j=1}^{m} a_j^2 D_{X,j}(W) X X^\top D_{X,j}(W).$$

Let $r = (f_\Theta(x_i) - y_i)_{i=1}^n$. We obtain

$$\|\nabla_a \mathcal{L}(\Theta)\|_2^2 = \frac{1}{n^2} r^\top \widehat{\mathbf{K}}_a(a, W) r, \qquad \|\nabla_W \mathcal{L}(\Theta)\|_{\mathrm{F}}^2 = \frac{1}{n^2} r^\top \widehat{\mathbf{K}}_W(a, W) r. \tag{B.1}$$

We bound the output gradient through its kernel and obtain the hidden gradient lower bound by controlling the deviation of the output weights from initialization.

Let $R, \rho \geq 0$, and define the parameter region

$$\mathcal{T}(R, \rho) := \left\{(a, W) : \|W - W_0\|_{\mathrm{F}} \leq R,\ \|a - a_0\|_2 \leq \rho\sqrt{m}\right\}.$$

Set

$$\epsilon_X(R) := \left(\frac{R^2}{m}\right)^{1/3} + \frac{\log(8n/\delta)}{m}, \qquad \Delta_X(R) := n\left[\epsilon_X(R) + \frac{R}{\sqrt{m}} + \frac{R^2}{m}\right]. \tag{B.2}$$

The following lemma controls the empirical Gram matrices under perturbations of the hidden weights.

Lemma B.1. *Fix $\delta \in (0, 1)$, suppose $\|x_i\|_2 = 1$ and $|y_i| \leq C$ for every $i \in [n]$, and let Assumptions 3.1 and 3.2 hold. If*

$$m \gtrsim \frac{n^2}{\lambda_0^2} \log^2\left(\frac{8n}{\lambda_0 \delta}\right), \tag{B.3}$$

then, with probability at least $1 - \delta$ over the paired initialization, for every $R \geq 0$ satisfying $\Delta_X(R) \lesssim \lambda_0$ and every $\Theta = (W, a)$ with $\|W - W_0\|_{\mathrm{F}} \leq R$, we have

$$\widehat{\mathbf{K}}_a(a, W) \succeq \frac{\lambda_a}{2} I_n, \qquad \widehat{\mathbf{K}}_W(a_0, W) \succeq \frac{\lambda_W}{2} I_n, \qquad \left\|\widehat{\mathbf{K}}_a(a, W)\right\|_2 \leq Cn. \tag{B.4}$$

Proof. We first bound the minimum eigenvalues of the empirical Gram matrices and the operator norm of the feature matrix at initialization. By the paired initialization in Assumption 3.2, we have

$$\widehat{\mathbf{K}}_a(a_0, W_0) = \frac{1}{p} \sum_{j=1}^{p} \sigma(X u_j) \sigma(X u_j)^\top, \quad \widehat{\mathbf{K}}_W(a_0, W_0) = \frac{1}{p} \sum_{j=1}^{p} D(u_j) X X^\top D(u_j). \tag{B.5}$$

By (4), the two averages have expectations $\mathbf{K}_a$ and $\mathbf{K}_W$, respectively. For $i, l \in [n]$, we have

$$0 \leq \sigma(u_j^\top x_i)\sigma(u_j^\top x_l) \leq \frac{(u_j^\top x_i)^2 + (u_j^\top x_l)^2}{2}.$$

Since $u_j^\top x_i \sim \mathcal{N}(0, \|x_i\|_2^2) = \mathcal{N}(0, 1)$, we have, for $i, l \in [n]$ and $j \in [p]$,

$$\begin{aligned} \left\|\sigma(u_j^\top x_i)\sigma(u_j^\top x_l)\right\|_{\psi_1} &\leq \frac{1}{2}\left(\left\|(u_j^\top x_i)^2\right\|_{\psi_1} + \left\|(u_j^\top x_l)^2\right\|_{\psi_1}\right) \leq C, \\ \left|[D(u_j) X X^\top D(u_j)]_{il}\right| &= |x_i^\top x_l| \mathbf{1}\{u_j^\top x_i > 0,\ u_j^\top x_l > 0\} \leq 1, \end{aligned} \tag{B.6}$$

where $\|Z\|_{\psi_1} := \inf\{s > 0 : \mathbb{E}\exp(|Z|/s) \leq 2\}$ is the subexponential norm and C is an absolute constant. The final inequality follows from $|x_i^\top x_l| \leq \|x_i\|_2 \|x_l\|_2 = 1$. Under (B.3), applying Bernstein's inequality (Vershynin, 2018) entrywise with the bounds (B.6) and taking a union bound over the entries of both Gram matrices, we obtain, with probability at least $1 - \delta/2$,

$$\begin{aligned} \left\|\widehat{\mathbf{K}}_b(a_0, W_0) - \mathbf{K}_b\right\|_2 &\leq n \max_{i,l \in [n]} \left|[\widehat{\mathbf{K}}_b(a_0, W_0) - \mathbf{K}_b]_{il}\right| \leq \frac{\lambda_0}{4}, \\ \widehat{\mathbf{K}}_b(a_0, W_0) &\succeq \left(\lambda_b - \frac{\lambda_0}{4}\right) I_n \succeq \frac{3\lambda_b}{4} I_n, \qquad b \in \{a, W\}. \end{aligned} \tag{B.7}$$

The second line follows from the first and $\lambda_0 \le \lambda_b$. On the event where (B.7) holds, we also obtain

$$\|Z_X(W_0)\|_2^2 = m \left\| \widehat{\mathbf{K}}_a(a_0, W_0) \right\|_2 \le m \left(\|\mathbf{K}_a\|_2 + \frac{\lambda_0}{4} \right) \le m \left(\frac{n}{2} + \frac{\lambda_0}{4} \right) \le Cmn, \tag{B.8}$$

where the equality follows from $\widehat{\mathbf{K}}_a(a_0, W_0) = m^{-1} Z_X(W_0) Z_X(W_0)^\top$, the first inequality follows from (B.7), the second follows from $\mathbf{K}_a \succeq 0$ and $\mathrm{tr}(\mathbf{K}_a) = n/2$, and the last follows from $\lambda_0 \le 1/2$.

Next, we bound the deviations of the empirical Gram matrices from their initial values when $\|W - W_0\|_{\mathrm{F}} \le R$. By independence of $u_1, \dots, u_p$ and the Gaussian bound $\mathbb{P}(|u_j^\top x_i| \le u) \le u$, we obtain, for $k \in [p]$,

$$\mathbb{P}\left(\sum_{j=1}^{p} \mathbf{1}\left\{ |u_j^\top x_i| \le \frac{k}{8p} \right\} \ge k \right) \le \binom{p}{k} \left(\frac{k}{8p} \right)^k \le e^{-k}. \tag{B.9}$$

Applying a union bound to (B.9), we obtain

$$\mathbb{P}\left(\bigcup_{\substack{i \in [n],\, k \in [p] \\ k \ge \log(8n/\delta)}} \left\{ \sum_{j=1}^{p} \mathbf{1}\left\{ |u_j^\top x_i| \le \frac{k}{8p} \right\} \ge k \right\} \right) \le n \sum_{k = \lceil \log(8n/\delta) \rceil}^{\infty} e^{-k} \le \frac{\delta}{8(1 - e^{-1})} < \frac{\delta}{2}. \tag{B.10}$$

Let $N_i(u) := \sum_{j=1}^{p} \mathbf{1}\{|u_j^\top x_i| \le u\}$. By (B.10), we have $N_i(k/(8p)) < k$ for all $i \in [n]$ and $k \in [p]$ satisfying $k \ge \log(8n/\delta)$, with probability at least $1 - \delta/2$. Fix an initialization satisfying these count inequalities, and let $i \in [n]$ and $u > 0$. If $N_i(u) \ge \log(8n/\delta)$, we set $k := N_i(u) \in [p]$. If $u \le k/(8p)$, we would have $k = N_i(u) \le N_i(k/(8p)) < k$, which is a contradiction. Here, the first inequality follows from monotonicity and the second from the uniform count bound derived from (B.10) on the chosen event. Hence, $u > k/(8p)$. Thus, either $N_i(u) < \log(8n/\delta)$ or $N_i(u) < 8pu$. Since each initial weight appears twice and $m = 2p$, we obtain

$$\frac{1}{m} \#\{j : |w_{0j}^\top x_i| \le u\} = \frac{N_i(u)}{p} \le 8u + \frac{2\log(8n/\delta)}{m}, \qquad i \in [n], \quad u > 0. \tag{B.11}$$

For the remainder of the proof, we work on the event where (B.7) and (B.11) hold. Its probability is at least $1 - \delta$ by a union bound.

Since $\|x_i\|_2 = 1$, a change in the activation indicator of neuron j at x_i requires either $|w_{0j}^\top x_i| \le u$ or $\|w_j - w_{0j}\|_2 > u$. We also have $\#\{j : \|w_j - w_{0j}\|_2 > u\} \le R^2/u^2$, since $\|W - W_0\|_{\mathrm{F}} \le R$. Combining this row count with (B.11) and taking the infimum over $u > 0$, we obtain

$$\frac{1}{m} \#\{j : \mathbf{1}\{w_j^\top x_i > 0\} \ne \mathbf{1}\{w_{0j}^\top x_i > 0\}\} \le \inf_{u > 0} \left\{ 8u + \frac{2\log(8n/\delta)}{m} + \frac{R^2}{mu^2} \right\} \lesssim \epsilon_X(R). \tag{B.12}$$

The last comparison follows from $\inf_{u>0}\{8u + R^2/(mu^2)\} = 3 \cdot 2^{4/3}(R^2/m)^{1/3}$.

Since $a_{0j}^2 = 1$ and $|x_i^\top x_l| \le 1$, the absolute change in the (i,l)-entry of $\widehat{\mathbf{K}}_W(a_0, W)$ is at most the sum of the activation change fractions for x_i and x_l. Applying $\|A\|_2 \le n \max_{i,l} |A_{il}|$ and (B.12), we obtain

$$\left\| \widehat{\mathbf{K}}_W(a_0, W) - \widehat{\mathbf{K}}_W(a_0, W_0) \right\|_2 \le Cn\epsilon_X(R). \tag{B.13}$$

For the output Gram matrix, we obtain

$$\begin{aligned} \left\| \widehat{\mathbf{K}}_a(a, W) - \widehat{\mathbf{K}}_a(a_0, W_0) \right\|_2 &\le \frac{\{\|Z_X(W)\|_2 + \|Z_X(W_0)\|_2\} \, \|Z_X(W) - Z_X(W_0)\|_{\mathrm{F}}}{m} \\ &\le Cn \left(\frac{R}{\sqrt{m}} + \frac{R^2}{m} \right), \end{aligned} \tag{B.14}$$

where the first inequality follows from $AA^\top - BB^\top = (A - B)A^\top + B(A - B)^\top$, the triangle inequality, $\|UV^\top\|_2 \le \|U\|_2 \, \|V\|_2$, and $\|U\|_2 \le \|U\|_{\mathrm{F}}$, with $A = Z_X(W)$ and $B = Z_X(W_0)$. The second follows from (B.8), the triangle inequality, and the ReLU Lipschitz

bound $\|Z_X(W) - Z_X(W_0)\|_{\mathrm{F}} \le \sqrt{n}R$. By (B.13), (B.14), and (B.2), we bound both perturbations by $C\Delta_X(R) \le \lambda_0/4$ for a sufficiently small implicit constant in $\Delta_X(R) \lesssim \lambda_0$. Applying (B.7) and $\|\mathbf{K}_a\|_2 \le n/2$, we obtain all three bounds (B.4) simultaneously for all $R \ge 0$ and W satisfying $\Delta_X(R) \lesssim \lambda_0$ and $\|W - W_0\|_{\mathrm{F}} \le R$. Since the initialization event has probability at least $1 - \delta$, the proof is complete. □

We now derive the gradient bounds, with the hidden gradient requiring control of $a - a_0$.

Proposition B.1. *Under the assumptions of Lemma B.1, with probability at least $1 - \delta$ over the paired initialization, the following bounds hold for every $R, \rho \ge 0$ satisfying $\rho \lesssim \sqrt{\lambda_W/n}$ and $\Delta_X(R) \lesssim \lambda_0$, and every $\Theta = (W, a)$ with $(a, W) \in \mathcal{T}(R, \rho)$:*

$$\frac{\lambda_a}{n}\mathcal{L}(\Theta) \lesssim \|\nabla_a \mathcal{L}(\Theta)\|_2^2 \lesssim \mathcal{L}(\Theta), \qquad \|\nabla_a \mathcal{L}(\Theta)\|_\infty \lesssim \sqrt{\mathcal{L}(\Theta)}, \tag{B.15}$$

$$\|\nabla_W \mathcal{L}(\Theta)\|_{\mathrm{F}} \gtrsim \sqrt{\frac{\lambda_W}{n}}\sqrt{2\mathcal{L}(\Theta)}. \tag{B.16}$$

Proof. The conditions $\Delta_X(R) \lesssim \lambda_0$ and $\|W - W_0\|_{\mathrm{F}} \le R$ ensure that Lemma B.1 can be applied. The additional conditions $\|a - a_0\|_2 \le \rho\sqrt{m}$ and $\rho \lesssim \sqrt{\lambda_W/n}$ are needed only for the hidden gradient lower bound. Substituting (B.4) and $\|r\|_2^2 = 2n\mathcal{L}(\Theta)$ into (B.1), we obtain the Euclidean norm estimates in (B.15). The infinity norm estimate follows from $\|\nabla_a \mathcal{L}(\Theta)\|_\infty \le \|\nabla_a \mathcal{L}(\Theta)\|_2$.

Finally, we establish the hidden gradient lower bound by controlling the effect of $a - a_0$. Define the auxiliary matrix $G^{(0)} \in \mathbb{R}^{m\times d}$ by $(G^{(0)})_{j,:} := \frac{a_{0,j}}{n\sqrt{m}} r^\top D_{X,j}(W)X$, $j \in [m]$, where r is the residual vector at $\Theta = (W, a)$. By the triangle inequality and Cauchy–Schwarz, we have

$$\|\nabla_W \mathcal{L}(\Theta)\|_{\mathrm{F}} \ge \left\|G^{(0)}\right\|_{\mathrm{F}} - \frac{\|a - a_0\|_2 \|X\|_2 \|r\|_2}{n\sqrt{m}} \ge \left(c\sqrt{\frac{\lambda_W}{n}} - \rho\right)\sqrt{2\mathcal{L}(\Theta)},$$

where the second inequality follows from $\left\|G^{(0)}\right\|_{\mathrm{F}}^2 = n^{-2} r^\top \widehat{\mathbf{K}}_W(a_0, W) r$, (B.4), $\|r\|_2^2 = 2n\mathcal{L}(\Theta)$, $\|a - a_0\|_2 \le \rho\sqrt{m}$, and $\|X\|_2 \le \sqrt{n}$. Taking the constant in $\rho \lesssim \sqrt{\lambda_W/n}$ sufficiently small, we obtain (B.16), which completes the proof. □

We next establish an upper bound on the loss after a joint update of a and W.

Proposition B.2. *Under the assumptions of Lemma B.1, with probability at least $1 - \delta$ over the paired initialization, the following bound holds for every $R, \rho \ge 0$ satisfying $\rho \lesssim \sqrt{\lambda_W/n}$ and $\Delta_X(R) \lesssim \lambda_0$, and every $\Theta = (W, a)$ and (v, V) such that $(a + sv, W + sV) \in \mathcal{T}(R, \rho)$ for all $s \in [0, 1]$:*

$$\begin{aligned}\mathcal{L}(\Theta + (V, v)) \le \mathcal{L}(\Theta) + \langle \nabla_a \mathcal{L}(\Theta), v\rangle + \langle \nabla_W \mathcal{L}(\Theta), V\rangle + K_R\sqrt{2\mathcal{L}(\Theta)}\,\|V\|_{\mathrm{F}} \\ + C_1\|v\|_2^2 + C_2\|v\|_2\|V\|_{\mathrm{F}} + C_3\|V\|_{\mathrm{F}}^2,\end{aligned} \tag{B.17}$$

where C_1, C_2, C_3 are numerical constants and $K_R \le C\{\rho + \sqrt{\epsilon_X(R)}\}$.

Proof. Let $\Delta f := (f_{\Theta+(V,v)}(x_i) - f_\Theta(x_i))_{i=1}^n$. Expanding the network output and the square loss, we obtain

$$\mathcal{L}(\Theta + (V, v)) - \mathcal{L}(\Theta) = \langle \nabla_a \mathcal{L}(\Theta), v\rangle + \langle \nabla_W \mathcal{L}(\Theta), V\rangle + \frac{1}{n}\langle r, e_1 + e_2\rangle + \frac{1}{2n}\|\Delta f\|_2^2, \tag{B.18}$$

where

$$e_1 := \frac{1}{\sqrt{m}}\{Z_X(W + V) - Z_X(W)\}a - \frac{1}{\sqrt{m}}\sum_{j=1}^m a_j D_{X,j}(W)XV_{j,:}^\top,$$

$$e_2 := \frac{1}{\sqrt{m}}\{Z_X(W + V) - Z_X(W)\}v.$$

We next bound the two error terms. For each $i \in [n]$, we have

$$\#\big\{j \in [m] : \mathbf{1}\{w_j^\top x_i > 0\} \ne \mathbf{1}\{w_j^\top x_i + V_{j,:}x_i > 0\}\big\} \le Cm\epsilon_X(R),$$

where the inequality follows from (B.12) applied to W and $W+V$. We therefore obtain

$$\frac{\|e_1\|_2}{\sqrt{n}} \le C\{\sqrt{\epsilon_X(R)}+\rho\}\,\|V\|_{\mathrm{F}}\,, \qquad \frac{\|e_2\|_2}{\sqrt{n}} \le \frac{\|v\|_2\,\|V\|_{\mathrm{F}}}{\sqrt{m}},$$

where the first bound follows from Cauchy–Schwarz and $\|a-a_0\|_2 \le \rho\sqrt{m}$, and the second follows from the Lipschitz property of ReLU. By paired initialization, (B.8), and the radius conditions, we obtain

$$\sqrt{\frac{2\mathcal{L}(\Theta)}{m}} \le C\left\{\frac{1}{\sqrt{m}}+\rho+\frac{(1+\rho)R}{\sqrt{m}}\right\} \le C.$$

Applying Cauchy–Schwarz and $\|r\|_2 = \sqrt{2n\mathcal{L}(\Theta)}$, we obtain

$$\frac{1}{n}\big|\langle r, e_1+e_2\rangle\big| \le C\{\rho+\sqrt{\epsilon_X(R)}\}\sqrt{2\mathcal{L}(\Theta)}\,\|V\|_{\mathrm{F}} + C\,\|v\|_2\,\|V\|_{\mathrm{F}}\,. \tag{B.19}$$

Finally, we bound the quadratic term in (B.18). By the definition of Δf, we have

$$\begin{aligned}\Delta f &= \frac{1}{\sqrt{m}}\{Z_X(W+V)(a+v) - Z_X(W)a\}\\ &= \frac{1}{\sqrt{m}}Z_X(W)v + \frac{1}{\sqrt{m}}\{Z_X(W+V)-Z_X(W)\}(a+v).\end{aligned}$$

Taking norms, we obtain

$$\frac{\|\Delta f\|_2}{\sqrt{n}} \le \frac{\|Z_X(W)\|_2}{\sqrt{mn}}\,\|v\|_2 + \frac{\|Z_X(W+V)-Z_X(W)\|_{\mathrm{F}}}{\sqrt{mn}}\,\|a+v\|_2\,, \tag{B.20}$$

where the inequality follows from the triangle inequality and $\|Az\|_2 \le \|A\|_{\mathrm{F}}\,\|z\|_2$.

We estimate the three factors separately. By (B.4), we have

$$\|Z_X(W)\|_2^2 = m\left\|\widehat{\mathbf{K}}_a(a,W)\right\|_2 \le Cmn.$$

By the Lipschitz property of ReLU and $\|x_i\|_2 = 1$, we obtain

$$\|Z_X(W+V)-Z_X(W)\|_{\mathrm{F}}^2 \le \sum_{i=1}^n \|Vx_i\|_2^2 \le n\,\|V\|_{\mathrm{F}}^2\,.$$

Since $(a+v, W+V) \in \mathcal{T}(R,\rho)$, we also have

$$\|a+v\|_2 \le \|a_0\|_2 + \|a+v-a_0\|_2 \le (1+\rho)\sqrt{m}.$$

Substituting these estimates into (B.20) and noting that $\rho \lesssim 1$, we obtain

$$\frac{\|\Delta f\|_2}{\sqrt{n}} \le C(\|v\|_2 + \|V\|_{\mathrm{F}}), \qquad \frac{\|\Delta f\|_2^2}{2n} \le C(\|v\|_2^2 + \|V\|_{\mathrm{F}}^2).$$

Substituting the quadratic bound and (B.19) into (B.18), we obtain (B.17), completing the proof. □

B.1.2 Analysis of the Adam Update

For the Adam update, define

$$D_t := \mathrm{diag}\Big((\sqrt{\widehat{v}_t}+\varepsilon_0)^{-1}\Big), \quad S_t := \widehat{m}_t^\top D_t \widehat{m}_t,$$
$$b_t^a := \frac{\beta_1(1-\beta_1^t)}{2(1-\beta_1)}.$$

We set $S_{-1} := 0$ and $\delta_\beta := \frac{1}{2} - \frac{\beta_1(2\beta_1-1)_+}{4}$.

The next lemma bounds the inner product $\langle g_t, D_t\widehat{m}_t\rangle$ and controls the current loss in terms of $S_t + S_{t-1}$. We use these estimates below to quantify Adam's contribution to the contraction of the Lyapunov function. The condition $\beta_1^2 \le \beta_2$ is sufficient here, whereas the self normalization bound in Lemma D.6 requires $\beta_1^2 < \beta_2$.

Lemma B.2. *Suppose* $0 \le \beta_1 < 1$, $0 < \beta_2 < 1$, $\beta_1^2 \le \beta_2$, *and* $\varepsilon_0 > 0$. *For every* $t \ge 0$, *we have*

$$\langle g_t, D_t \widehat{m}_t \rangle \ge (b_{t+1}^a + \delta_\beta) S_t - b_t^a S_{t-1}. \tag{B.21}$$

If $2\mu\mathcal{L}(\Theta_t) \le \|g_t\|_2^2$ *and* $\|g_s\|_\infty \le H$ *for every* $0 \le s \le t$, *then we have*

$$\mathcal{L}(\Theta_t) \le C_L(S_t + S_{t-1}), \tag{B.22}$$

where $C_L := \frac{(1+\beta_1)(H+\varepsilon_0)}{2\mu(1-\beta_1)^2}$.

Proof. First, we prove the inner product bound in (B.21). At $t = 0$, it follows from $\widehat{m}_0 = g_0$, $S_{-1} = 0$, and $b_1^a + \delta_\beta \le 1$. For $t \ge 1$, set

$$r := \frac{\beta_1(1-\beta_1^t)}{1-\beta_1^{t+1}}, \qquad s := \left(\frac{\beta_2(1-\beta_2^t)}{1-\beta_2^{t+1}}\right)^{1/2}.$$

By the update rules in Algorithm 1, we have

$$\widehat{m}_t = r\widehat{m}_{t-1} + (1-r)g_t, \qquad \widehat{v}_t \ge s^2 \widehat{v}_{t-1},$$

where the inequality implies $D_t \preceq s^{-1} D_{t-1}$ since $0 < s < 1$. We therefore obtain

$$\langle \widehat{m}_{t-1}, D_t \widehat{m}_t \rangle \le \sqrt{S_t S_{t-1}/s} \le \frac{1}{2s} S_t + \frac{1}{2} S_{t-1},$$

where the second inequality follows from Young's inequality. Substituting $g_t = (\widehat{m}_t - r\widehat{m}_{t-1})/(1-r)$, we have

$$\langle g_t, D_t \widehat{m}_t \rangle \ge \frac{1 - r/(2s)}{1-r} S_t - \frac{r}{2(1-r)} S_{t-1} = \frac{1 - r/(2s)}{1-r} S_t - b_t^a S_{t-1}.$$

Next, we show that $(1 - r/(2s))/(1-r) - b_{t+1}^a \ge \delta_\beta$. If $\beta_1 = 0$, the left hand side equals one. It remains to consider $0 < \beta_1 < 1$. Let $a = \beta_1^t \in (0, \beta_1]$. The function $x \mapsto x(1-x^t)/(1-x^{t+1})$ is nondecreasing on $[0, 1)$, since its derivative has the sign of $1 - (t+1)x^t + tx^{t+1} \ge 0$. By $\beta_1^2 \le \beta_2$, we obtain

$$s \ge \beta_1 \sqrt{\frac{1-a^2}{1-\beta_1^2 a^2}}, \qquad \frac{r}{s} \le z_a := \sqrt{\frac{(1-a)(1+\beta_1 a)}{(1-\beta_1 a)(1+a)}} \le 1.$$

Substituting these estimates and the definition of b_{t+1}^a, we obtain

$$\frac{1 - r/(2s)}{1-r} - b_{t+1}^a = \frac{1-\beta_1 a}{2(1-\beta_1)} \left(2 - \beta_1 - \frac{r}{s}\right) \ge \frac{1-\beta_1 a}{2(1-\beta_1)} (2 - \beta_1 - z_a).$$

Since $1 - z_a^2 = 2(1-\beta_1)a/\{(1-\beta_1 a)(1+a)\}$ and $1 - z_a \ge (1-z_a^2)/2$, we obtain

$$\frac{1 - r/(2s)}{1-r} - b_{t+1}^a \ge \frac{1}{2} + \frac{a}{2}\left(\frac{1}{1+a} - \beta_1\right).$$

By concavity of $a \mapsto a/(1+a) - \beta_1 a$ on $[0, \beta_1]$, we have

$$\frac{1 - r/(2s)}{1-r} - b_{t+1}^a \ge \frac{1}{2} + \frac{1}{2} \min\left\{0, \beta_1 \left(\frac{1}{1+\beta_1} - \beta_1\right)\right\} \ge \frac{1}{2} - \frac{\beta_1(2\beta_1 - 1)_+}{4} = \delta_\beta,$$

where the last inequality follows from $(1+\beta_1)^{-1} \ge 1/2$. Combining these bounds proves (B.21).

Finally, we prove (B.22). The assumption $\|g_s\|_\infty \le H$ implies $\widehat{v}_s \le H^2 \mathbf{1}$ for $0 \le s \le t$, and hence $D_s \succeq (H + \varepsilon_0)^{-1} I_m$. For $t \ge 1$, apply Young's inequality to $g_t = (\widehat{m}_t - r\widehat{m}_{t-1})/(1-r)$. Using $0 \le r \le \beta_1 < 1$ and $\|\widehat{m}_s\|_2^2 \le (H + \varepsilon_0) S_s$ for $s \in \{t-1, t\}$, we obtain

$$\|g_t\|_2^2 \le \frac{(1+\beta_1)(H+\varepsilon_0)}{(1-\beta_1)^2} (S_t + S_{t-1}).$$

At $t = 0$, the same estimate follows from $\widehat{m}_0 = g_0$ and $S_{-1} = 0$. Combining this estimate with $2\mu\mathcal{L}(\Theta_t) \le \|g_t\|_2^2$, we obtain (B.22). □

B.1.3 Analysis of the Muon Update

Let $\overline{M}_t := M_t/(1-\beta^{t+1})$ denote the corrected hidden momentum. Frobenius normalization cancels this scaling, so we have $Q_K(\overline{M}_t) = Q_K(M_t)$. Let $b := 1 - \max\{\beta_1, \beta\}$.

The next lemma bounds $\left\|\overline{M}_t - G_t\right\|_{\mathrm{F}}$. In Section B.1.5, we combine this estimate with Lemma D.5 to obtain a lower bound on $\langle G_t, Q_K(M_t)\rangle$.

Lemma B.3. *Under the assumptions of Proposition B.1, let Algorithm 1 use $\gamma_j \equiv \gamma > 0$ and $\eta_j = \eta\sqrt{2\mathcal{L}(\Theta_j)}$, where $\eta > 0$. Let $R, \rho \geq 0$ satisfy $\rho \lesssim \sqrt{\lambda_W/n}$ and $\Delta_X(R) \lesssim \lambda_0$, and fix $t \geq 0$ with $\mathcal{L}(\Theta_t) > 0$. With probability at least $1-\delta$ over the paired initialization, if its iterates satisfy*

$$(a_j, W_j) \in \mathcal{T}(R,\rho) \quad (0 \leq j \leq t+1), \qquad \mathcal{L}(\Theta_j) \leq \mathcal{L}(\Theta_0) \quad (0 \leq j \leq t)$$

and, for some $s \in (0, b/2]$,

$$\|r_{j+1} - r_j\|_2 \leq s \|r_j\|_2, \qquad 0 \leq j < t, \tag{B.23}$$

then

$$\|a_{t+1} - a_t\|_2 \leq d_0\sqrt{\mathcal{L}(\Theta_t)}, \qquad \frac{\|r_{t+1} - r_t\|_2}{\|r_t\|_2} \leq C(d_0 + \eta\mathcal{B}),$$

$$\left\|\overline{M}_t - G_t\right\|_{\mathrm{F}} \leq \frac{C\beta\|X\|_2}{(1-s-\beta)\sqrt{\lambda_W}}\left\{s + \sqrt{\epsilon_X(R)} + \rho + \frac{d_0\sqrt{\mathcal{L}(\Theta_0)}}{\sqrt{m}}\right\}\|G_t\|_{\mathrm{F}}.$$

Here $d_0 := C\gamma\sqrt{L_0}/(\varepsilon_0 b)$, with numerical constants $C \geq 1$ and $L_0 > 0$.

Proof. By Lemma B.1, the local Gram matrix bounds (B.4) and the bound (B.12) hold throughout the parameter region with probability at least $1-\delta$ over the paired initialization. We prove the claim under these bounds.

First, we bound the parameter and residual increments. By (B.23), we have $\|r_j\|_2 \leq (1-s)^{-(t-j)}\|r_t\|_2$ for $0 \leq j \leq t$. Expanding $\widehat{m}_t$ and applying (B.15), we obtain

$$\|\widehat{m}_t\|_2 \leq \frac{1-\beta_1}{1-\beta_1^{t+1}}\sum_{j=0}^{t}\beta_1^{t-j}\|g_j\|_2 \leq \sqrt{2L_0\mathcal{L}(\Theta_t)}\sum_{k=0}^{t}\left(\frac{\beta_1}{1-s}\right)^k \leq \frac{2\sqrt{2L_0}}{b}\sqrt{\mathcal{L}(\Theta_t)},$$

where the first inequality follows from the triangle inequality, the second inequality follows from (B.15), (B.23), and $(1-\beta_1)/(1-\beta_1^{t+1}) \leq 1$, and the last from the geometric sum formula and $1 - s - \beta_1 \geq b/2$. Applying $D_t \preceq \varepsilon_0^{-1} I_m$ to the Adam update and the bound $\|Q_K(M_t)\|_{\mathrm{F}} \leq \mathcal{B}$ from Lemma D.1 to the Muon update, we obtain

$$\|a_{t+1} - a_t\|_2 \leq \frac{\gamma}{\varepsilon_0}\|\widehat{m}_t\|_2 \leq d_0\sqrt{\mathcal{L}(\Theta_t)}, \qquad \|W_{t+1} - W_t\|_{\mathrm{F}} \leq \eta\mathcal{B}\sqrt{2\mathcal{L}(\Theta_t)}. \tag{B.24}$$

Since $\mathcal{T}(R,\rho)$ is convex, the current update segment lies in this region. Applying (B.20) to the current update, together with (B.4), the Lipschitz property of ReLU, and $\rho \leq 1$, we obtain

$$\frac{\|r_{t+1} - r_t\|_2}{\|r_t\|_2} \leq \frac{C\{\|a_{t+1} - a_t\|_2 + \|W_{t+1} - W_t\|_{\mathrm{F}}\}}{\sqrt{2\mathcal{L}(\Theta_t)}} \leq C(d_0 + \eta\mathcal{B}).$$

Next, we will bound $G_{j+1} - G_j$ for $0 \leq j < t$. Denote by $(G_j)_{l,:}$ the l-th row of G_j. So we have $(G_j)_{l,:} = a_{j,l} r_j^\top D_{X,l}(W_j)X/(n\sqrt{m})$. Then we obtain

$$\|(G_{j+1} - G_j)_{l,:}\|_2 \leq \frac{\|X\|_2}{n\sqrt{m}}\Big[|a_{j+1,l}|\,\|r_{j+1} - r_j\|_2 + |a_{j+1,l} - a_{j,l}|\,\|r_j\|_2 \\ + |a_{j,l}|\,\|\{D_{X,l}(W_{j+1}) - D_{X,l}(W_j)\}r_j\|_2\Big], \qquad l \in [m],$$

where the inequality follows from the triangle inequality and $\|D_{X,l}(W)\|_2 \leq 1$. Applying (B.12) at W_j and W_{j+1}, we obtain

$$\sum_{l=1}^{m}\|\{D_{X,l}(W_{j+1}) - D_{X,l}(W_j)\}r_j\|_2^2 \leq Cm\epsilon_X(R)\|r_j\|_2^2.$$

Decomposing $a_j = a_0 + (a_j - a_0)$ and applying $|a_{0,l}| = 1$, we obtain

$$\left(\sum_{l=1}^{m} a_{j,l}^2 \left\|\{D_{X,l}(W_{j+1}) - D_{X,l}(W_j)\}r_j\right\|_2^2\right)^{1/2} \leq \left(\sqrt{Cm\epsilon_X(R)} + \|a_j - a_0\|_2\right)\|r_j\|_2 \leq C\sqrt{m}\left(\sqrt{\epsilon_X(R)} + \rho\right)\|r_j\|_2,$$

where the first inequality follows from $\|D_{X,l}(W_{j+1}) - D_{X,l}(W_j)\|_2 \leq 1$, and the second inequality follows from $\|a_j - a_0\|_2 \leq \rho\sqrt{m}$. By $\|G_{j+1} - G_j\|_{\mathrm{F}}^2 = \sum_{l=1}^{m} \|(G_{j+1} - G_j)_{l,:}\|_2^2$ we obtain

$$\|G_{j+1} - G_j\|_{\mathrm{F}} \leq \frac{C\|X\|_2}{n}\left[\|r_{j+1} - r_j\|_2 + \left(\sqrt{\epsilon_X(R)} + \rho + \frac{\|a_{j+1} - a_j\|_2}{\sqrt{m}}\right)\|r_j\|_2\right],$$

where $\|a_{j+1}\|_2 \leq (1+\rho)\sqrt{m} \leq 2\sqrt{m}$. And by (B.23) and (B.24), we have

$$\|r_{j+1} - r_j\|_2 \leq s\|r_j\|_2, \qquad \|a_{j+1} - a_j\|_2 \leq d_0\sqrt{\mathcal{L}(\Theta_j)} \leq d_0\sqrt{\mathcal{L}(\Theta_0)}.$$

Substituting these bounds into the Frobenius norm estimate and applying $\|r_j\|_2 = \sqrt{2n\mathcal{L}(\Theta_j)}$, we obtain

$$\|G_{j+1} - G_j\|_{\mathrm{F}} \leq \frac{C\|X\|_2}{\sqrt{n}}\left\{s + \sqrt{\epsilon_X(R)} + \rho + \frac{d_0\sqrt{\mathcal{L}(\Theta_0)}}{\sqrt{m}}\right\}\sqrt{2\mathcal{L}(\Theta_j)}. \tag{B.25}$$

Finally, we bound $\left\|\overline{M}_t - G_t\right\|_{\mathrm{F}}$. Expanding the recursion for M_t and rearranging the finite sums, we have

$$\overline{M}_t - G_t = -\sum_{j=0}^{t-1} \frac{\beta^{t-j}(1-\beta^{j+1})}{1-\beta^{t+1}}(G_{j+1} - G_j).$$

Since $(1-\beta^{j+1})/(1-\beta^{t+1}) \leq 1$, we obtain

$$\left\|\overline{M}_t - G_t\right\|_{\mathrm{F}} \leq \sum_{j=0}^{t-1} \beta^{t-j}\|G_{j+1} - G_j\|_{\mathrm{F}}.$$

Substituting (B.25) and applying the residual comparison from (B.23), we obtain

$$\begin{aligned}\left\|\overline{M}_t - G_t\right\|_{\mathrm{F}} &\leq \frac{C\|X\|_2}{\sqrt{n}}\left\{s + \sqrt{\epsilon_X(R)} + \rho + \frac{d_0\sqrt{\mathcal{L}(\Theta_0)}}{\sqrt{m}}\right\}\sqrt{2\mathcal{L}(\Theta_t)}\sum_{k=1}^{t}\left(\frac{\beta}{1-s}\right)^k \\ &\leq \frac{C\beta\|X\|_2}{(1-s-\beta)\sqrt{n}}\left\{s + \sqrt{\epsilon_X(R)} + \rho + \frac{d_0\sqrt{\mathcal{L}(\Theta_0)}}{\sqrt{m}}\right\}\sqrt{2\mathcal{L}(\Theta_t)},\end{aligned} \tag{B.26}$$

where the last inequality follows from $\sum_{k=1}^{t}(\beta/(1-s))^k \leq \beta/(1-s-\beta)$, since $1-s-\beta \geq b/2 > 0$. By (B.16), we have

$$\sqrt{2\mathcal{L}(\Theta_t)} \lesssim \sqrt{\frac{n}{\lambda_W}}\|G_t\|_{\mathrm{F}}. \tag{B.27}$$

Substituting (B.27) into (B.26), we obtain the desired momentum bound in Lemma B.3. At $t = 0$, we have $\overline{M}_0 = M_0/(1-\beta) = G_0$, so the bound also holds, which completes the proof. □

B.1.4 Contraction of the Lyapunov Function

In this step of the proof, we establish a contraction inequality for a Lyapunov function by combining the estimates for Adam and Muon, with separate contributions from the two optimizers. We consider the learning rates $\gamma_t \equiv \gamma$ and $\eta_t = \eta\sqrt{2\mathcal{L}(\Theta_t)}$. To control the contribution of Adam's momentum in (B.21), we introduce the Lyapunov function

$$\mathcal{E}_t := \mathcal{L}(\Theta_t) + \gamma(b_t^a + \omega_\beta)S_{t-1}, \qquad t \geq 0, \tag{B.28}$$

where $\omega_\beta := \delta_\beta/4$ and S_{t-1} is defined in Section B.1.2. It is straightforward to verify that $\mathcal{E}_0 = \mathcal{L}(\Theta_0)$ and $\mathcal{L}(\Theta_t) \le \mathcal{E}_t$.

For convenience, we set $\eta_\beta := \delta_\beta/8$. We define

$$A_1 := C_1 + \frac{C_2}{2}, \qquad R_1 := 2C_3 + C_2, \qquad C_m := \frac{\beta_1}{2(1-\beta_1)} + \omega_\beta, \qquad C_E := C_L + \gamma C_m. \tag{B.29}$$

In the following proposition, we establish a contraction bound for $\mathcal{E}_t$ by absorbing the errors from hidden weight updates into the Muon descent term.

Proposition B.3. *Fix an iteration $t \ge 0$ of Algorithm 1 with $\mathcal{L}(\Theta_t) > 0$ and $M_t \ne 0$. Assume* (B.21) *and* (B.22) *hold at t, and* (B.17) *holds on the segment from Θ_t to Θ_{t+1}. If, for some $\mu_0 > 0$, we have $\langle G_t, Q_K(M_t)\rangle \ge \frac{\mu_0}{2}\chi_K(M_t)\sqrt{2\mathcal{L}(\Theta_t)}$, $\gamma A_1/\varepsilon_0 \le \delta_\beta/4$, and $\underline{\Lambda} := \mu_0 - 2K_R\sqrt{\mathcal{B}} - \eta R_1 \mathcal{B} > 0$, then*

$$\mathcal{E}_{t+1} \le (1 - c_t)\mathcal{E}_t, \tag{B.30}$$

where $c_t := \min\left\{\frac{\eta_\beta}{C_m}, \frac{\gamma\eta_\beta + \eta\underline{\Lambda} C_L \chi_K(M_t)}{C_E}\right\} \in (0, 1/2]$.

Proof. First, we derive an upper bound on $\mathcal{L}(\Theta_{t+1}) - \mathcal{L}(\Theta_t)$ for the joint Adam and Muon update. Let $u_t := a_{t+1} - a_t$ and $V_t := W_{t+1} - W_t$. By the Adam update in Algorithm 1, we have

$$\|u_t\|_2^2 = \gamma^2 \widehat{m}_t^\top D_t^2 \widehat{m}_t \le \frac{\gamma^2}{\varepsilon_0} S_t, \tag{B.31}$$

where the inequality follows from $0 \prec D_t \preceq \varepsilon_0^{-1} I_m$. By Lemma D.1, we also have

$$\|Q_K(M_t)\|_{\mathrm{F}}^2 \le \mathcal{B}\chi_K(M_t), \qquad \|Q_K(M_t)\|_{\mathrm{F}} \le \sqrt{\mathcal{B}}\chi_K(M_t).$$

Substituting $u_t = -\gamma D_t \widehat{m}_t$ and $V_t = -\eta\sqrt{2\mathcal{L}(\Theta_t)} Q_K(M_t)$ into (B.17), we obtain

$$\begin{aligned}
\mathcal{L}(\Theta_{t+1}) - \mathcal{L}(\Theta_t) &\le -\gamma \langle g_t, D_t \widehat{m}_t\rangle - \eta\sqrt{2\mathcal{L}(\Theta_t)} \langle G_t, Q_K(M_t)\rangle + \frac{\gamma^2 A_1}{\varepsilon_0} S_t \\
&\quad + 2\eta K_R \mathcal{L}(\Theta_t) \|Q_K(M_t)\|_{\mathrm{F}} + \eta^2 R_1 \mathcal{L}(\Theta_t) \|Q_K(M_t)\|_{\mathrm{F}}^2 \\
&\le -\gamma \langle g_t, D_t \widehat{m}_t\rangle + \frac{\gamma^2 A_1}{\varepsilon_0} S_t - \eta\underline{\Lambda}\chi_K(M_t)\mathcal{L}(\Theta_t),
\end{aligned} \tag{B.32}$$

where the first inequality follows from (B.31), Young's inequality and (B.29), while the second inequality follows from the assumed lower bound on $\langle G_t, Q_K(M_t)\rangle$, the two bounds from Lemma D.1, and the definition of $\underline{\Lambda}$.

Next, we bound the change in the Lyapunov function:

$$\begin{aligned}
\mathcal{E}_{t+1} - \mathcal{E}_t &= \mathcal{L}(\Theta_{t+1}) - \mathcal{L}(\Theta_t) + \gamma(b_{t+1}^a + \omega_\beta) S_t - \gamma(b_t^a + \omega_\beta) S_{t-1} \\
&\le -\gamma\left(\delta_\beta - \omega_\beta - \frac{\gamma A_1}{\varepsilon_0}\right) S_t - \gamma\omega_\beta S_{t-1} - \eta\underline{\Lambda}\chi_K(M_t)\mathcal{L}(\Theta_t) \\
&\le -\gamma\eta_\beta(S_t + S_{t-1}) - \eta\underline{\Lambda}\chi_K(M_t)\mathcal{L}(\Theta_t),
\end{aligned} \tag{B.33}$$

where the equality follows from (B.28), the first inequality follows from (B.32) and (B.21), and the second follows from $\gamma A_1/\varepsilon_0 \le \delta_\beta/4$, $\omega_\beta = \delta_\beta/4$, and $\eta_\beta = \delta_\beta/8$.

Finally, we derive the contraction factor. By (B.22) and $\mathcal{L}(\Theta_t) > 0$, we have $S_t + S_{t-1} > 0$. Let $x := \mathcal{L}(\Theta_t)/(S_t + S_{t-1})$, so that $0 < x \le C_L$. By (B.28) and $b_t^a + \omega_\beta \le C_m$, we have

$$\mathcal{E}_t \le \mathcal{L}(\Theta_t) + \gamma C_m (S_t + S_{t-1}). \tag{B.34}$$

We then obtain

$$\begin{aligned}
\frac{\gamma\eta_\beta(S_t + S_{t-1}) + \eta\underline{\Lambda}\chi_K(M_t)\mathcal{L}(\Theta_t)}{\mathcal{E}_t} &\ge \frac{\gamma\eta_\beta + \eta\underline{\Lambda}\chi_K(M_t)x}{\gamma C_m + x} \\
&\ge \min\left\{\frac{\eta_\beta}{C_m}, \frac{\gamma\eta_\beta + \eta\underline{\Lambda} C_L \chi_K(M_t)}{C_E}\right\} = c_t,
\end{aligned} \tag{B.35}$$

where the first inequality follows from (B.34), the second follows from the fact that the scalar ratio attains its minimum over $[0, C_L]$ at $x = 0$ or $x = C_L$, and the equality follows from the definition of c_t. Combining (B.35) and (B.33), we obtain (B.30). Since both entries in the minimum defining c_t are positive and $C_m \ge \omega_\beta = 2\eta_\beta > 0$ by (B.29), we have $0 < c_t \le \eta_\beta/C_m \le 1/2$. □

B.1.5 Completion of the Proof

We combine the preceding estimates to complete the proof of Theorem 4.1. For the remainder of the appendix, write $q := \kappa + 1$.

Proof of Theorem 4.1. We consider learning rates satisfying

$$0 < \gamma \le \frac{c_h \varepsilon_0 b^2 \sqrt{\lambda_0}}{(1+\sqrt{\mathcal{B}})\sqrt{n}}, \qquad 0 < \eta \le \frac{c_h b \sqrt{\lambda_0}}{\mathcal{B}(1+\sqrt{\mathcal{B}})\sqrt{n}}, \tag{B.36}$$

and set

$$s_\star := \frac{c_0 b \sqrt{\lambda_0}}{(1+\sqrt{\mathcal{B}})\sqrt{n}}, \qquad \rho_\star := \frac{s_\star}{\sqrt{\mathcal{B}}}, \qquad R := C\sqrt{\frac{n}{\lambda_0}}, \tag{B.37}$$

where $c_0 := 2^{-8} C_0^{-4}$, and $C_0 \ge 2$, $c_h > 0$, and $C > 0$ are constants chosen below, independently of the learning rates within (B.36). In particular, $s_\star \le b/2$. By the width requirement in Theorem 4.1 and (B.2), we can choose the implicit width constant so that (B.3) holds and

$$\frac{n}{\varepsilon_0 \lambda_0 \sqrt{m}} \le c\rho_\star, \qquad \epsilon_X(R) \le \rho_\star^2, \qquad \Delta_X(R) \le c\lambda_0. \tag{B.38}$$

By Assumption 3.2, we have $\mathcal{L}(\Theta_0) = (2n)^{-1}\sum_i y_i^2 \le C^2/2$. We assume $\mathcal{L}(\Theta_0) > 0$, since otherwise all updates vanish. Let $c_{\mathrm{A}} := c\gamma\lambda_a/n$ and $c_{\mathrm{M}} := c\eta\sqrt{\lambda_W/n}$, where $c > 0$ is chosen below with $c_{\mathrm{A}} + c_{\mathrm{M}} \le 1/2$. We prove by induction on $T \ge 0$ that

$$\begin{aligned} &(a_j, W_j) \in \mathcal{T}(R, \rho_\star), && 0 \le j \le T, \\ &0 < \mathcal{L}(\Theta_j) \le \mathcal{E}_j \le \mathcal{L}(\Theta_0) e^{-(c_{\mathrm{A}}+c_{\mathrm{M}})j}, && 0 \le j \le T, \\ &\|r_{j+1} - r_j\|_2 \le s_\star \|r_j\|_2, && 0 \le j < T. \end{aligned}$$

The last bound is (B.23) with $s = s_\star$, required by Lemma B.3. The case $T = 0$ follows from initialization and $\mathcal{E}_0 = \mathcal{L}(\Theta_0)$. Assume the claims hold through T. We prove them at $T+1$ in three steps.

First, we show that the update from Θ_T to Θ_{T+1} remains in $\mathcal{T}(R, \rho_\star)$. By the Adam update in Algorithm 1, we obtain

$$\begin{aligned} \sum_{t=0}^{T} \|a_{t+1} - a_t\|_2 &\le \frac{\gamma}{\varepsilon_0} \sum_{t=0}^{T} \frac{1-\beta_1}{1-\beta_1^{t+1}} \sum_{j=0}^{t} \beta_1^{t-j} \|g_j\|_2 \\ &\le \frac{\gamma}{\varepsilon_0} \sum_{j=0}^{T} \|g_j\|_2 \sum_{t=j}^{T} \beta_1^{t-j} \le \frac{\gamma}{\varepsilon_0(1-\beta_1)} \sum_{j=0}^{T} \|g_j\|_2, \end{aligned} \tag{B.39}$$

where the first inequality follows from $D_t \preceq \varepsilon_0^{-1} I_m$, and the second inequality follows from $(1-\beta_1)/(1-\beta_1^{t+1}) \le 1$. We further obtain

$$\begin{aligned} \sum_{t=0}^{T} \|a_{t+1} - a_t\|_2 &\lesssim \frac{\gamma\sqrt{\mathcal{L}(\Theta_0)}}{\varepsilon_0(1-\beta_1)} \sum_{j=0}^{T} e^{-(c_{\mathrm{A}}+c_{\mathrm{M}})j/2} \\ &\le \frac{\gamma\sqrt{\mathcal{L}(\Theta_0)}}{\varepsilon_0(1-\beta_1)\{1 - e^{-(c_{\mathrm{A}}+c_{\mathrm{M}})/2}\}} \lesssim \frac{\gamma\sqrt{\mathcal{L}(\Theta_0)}}{\varepsilon_0(1-\beta_1)(c_{\mathrm{A}}+c_{\mathrm{M}})}, \end{aligned} \tag{B.40}$$

where the first inequality follows from (B.15) and the induction hypothesis, the second inequality follows from the geometric sum formula, and the last inequality follows from $1 - e^{-(c_{\mathrm{A}}+c_{\mathrm{M}})/2} \ge (c_{\mathrm{A}}+c_{\mathrm{M}})/4$ for $0 < c_{\mathrm{A}} + c_{\mathrm{M}} \le 1/2$. Similarly, applying $\|Q_K(M_t)\|_{\mathrm{F}} \le \mathcal{B}$ from Lemma D.1 to the Muon updates, we obtain

$$\sum_{t=0}^{T} \|W_{t+1} - W_t\|_{\mathrm{F}} \le \eta\mathcal{B}\sqrt{2\mathcal{L}(\Theta_0)} \sum_{t=0}^{T} e^{-(c_{\mathrm{A}}+c_{\mathrm{M}})t/2} \lesssim \frac{\eta\mathcal{B}\sqrt{\mathcal{L}(\Theta_0)}}{c_{\mathrm{A}}+c_{\mathrm{M}}}. \tag{B.41}$$

By (B.40), (B.41), and the definitions of c_{A} and c_{M}, we obtain

$$\sum_{t=0}^{T} \|a_{t+1} - a_t\|_2 \le \frac{C'n}{\varepsilon_0 \lambda_a}, \qquad \sum_{t=0}^{T} \|W_{t+1} - W_t\|_{\mathrm{F}} \le C'\mathcal{B}\sqrt{\frac{n}{\lambda_W}}, \tag{B.42}$$

where the first inequality follows from $c_{\mathrm{A}} + c_{\mathrm{M}} \ge c\gamma\lambda_a/n$, and the second inequality follows from $c_{\mathrm{A}} + c_{\mathrm{M}} \ge c\eta\sqrt{\lambda_W/n}$. The constant $C' > 0$ is independent of T, γ, and η. Choosing $C \ge 2C'\mathcal{B}$ in (B.37) and enlarging the implicit width constant in (B.38), we obtain

$$\sum_{t=0}^{T} \|a_{t+1} - a_t\|_2 \le \frac{\rho_\star\sqrt{m}}{2}, \qquad \sum_{t=0}^{T} \|W_{t+1} - W_t\|_{\mathrm{F}} \le \frac{R}{2}.$$

Thus $(a_{T+1}, W_{T+1}) \in \mathcal{T}(R, \rho_\star)$, and the update segment also lies in this region by convexity.

Next, we extend (B.23) to $0 \le j < T+1$ and establish a lower bound on $\langle G_T, Q_K(M_T)\rangle$. By the first step and the induction hypothesis, we have $(a_j, W_j) \in \mathcal{T}(R, \rho_\star)$ for $0 \le j \le T+1$. We also have $0 < \mathcal{L}(\Theta_j) \le \mathcal{L}(\Theta_0)$ for $0 \le j \le T$ and (B.23) with $s = s_\star$ for $0 \le j < T$ by the induction hypothesis. Applying Lemma B.3 with $t = T$, $\rho = \rho_\star$, and $s = s_\star$, we obtain

$$\|r_{T+1} - r_T\|_2 \le C(d_0 + \eta\mathcal{B}) \|r_T\|_2 \le s_\star \|r_T\|_2 ,$$

where $d_0 \lesssim \gamma/(\varepsilon_0 b)$, and the second inequality follows from (B.36) and (B.37). This extends (B.23) to $0 \le j < T+1$. Since $s_\star < 1$, we also have $\|r_{T+1}\|_2 \ge (1 - s_\star) \|r_T\|_2 > 0$, so $\mathcal{L}(\Theta_{T+1}) > 0$.

To apply Lemma D.5, we next bound $\left\|\overline{M}_T - G_T\right\|_{\mathrm{F}}$. The learning rates (B.36) and the width requirement imply $d_0\sqrt{\mathcal{L}(\Theta_0)/m} \le s_\star$. Applying the bound on $\left\|\overline{M}_t - G_t\right\|_{\mathrm{F}}$ in Lemma B.3 at $t = T$, we obtain

$$\begin{aligned}\left\|\overline{M}_T - G_T\right\|_{\mathrm{F}} &\le \frac{C\beta \|X\|_2}{(1 - s_\star - \beta)\sqrt{\lambda_W}} \left\{ s_\star + \sqrt{\epsilon_X(R)} + \rho_\star + d_0\sqrt{\frac{\mathcal{L}(\Theta_0)}{m}} \right\} \|G_T\|_{\mathrm{F}} \\ &\le \frac{2C\sqrt{n}}{b\sqrt{\lambda_W}}(2s_\star + 2\rho_\star) \|G_T\|_{\mathrm{F}} = \frac{4C\sqrt{n}(s_\star + \rho_\star)}{b\sqrt{\lambda_W}} \|G_T\|_{\mathrm{F}} ,\end{aligned}$$

where the second inequality follows from $\sqrt{\epsilon_X(R)} \le \rho_\star$, $d_0\sqrt{\mathcal{L}(\Theta_0)/m} \le s_\star$, $\|X\|_2 \le \sqrt{n}$, and $1 - s_\star - \beta \ge b/2$. We next consider the coefficient of $\|G_T\|_{\mathrm{F}}$:

$$\frac{4C\sqrt{n}(s_\star + \rho_\star)}{b\sqrt{\lambda_W}} = \frac{4Cc_0}{\sqrt{\mathcal{B}}}\sqrt{\frac{\lambda_0}{\lambda_W}} \le \frac{8Cc_0}{1 + \sqrt{\mathcal{B}}} \le \frac{1}{2(1 + \sqrt{\mathcal{B}})},$$

where the equality follows from (B.37), the first inequality follows from $1 + \sqrt{\mathcal{B}} \le 2\sqrt{\mathcal{B}}$, and the second inequality follows from $16Cc_0 \le 1$. We therefore obtain $\left\|\overline{M}_T - G_T\right\|_{\mathrm{F}} \le \|G_T\|_{\mathrm{F}} / [2(1 + \sqrt{\mathcal{B}})]$. Applying Lemmas D.5 and D.1, we obtain

$$\begin{gathered}\langle G_T, Q_K(M_T)\rangle \ge \frac{1}{2}\chi_K(M_T) \|G_T\|_{\mathrm{F}} \gtrsim \sqrt{\frac{\lambda_W}{n}}\chi_K(M_T)\sqrt{2\mathcal{L}(\Theta_T)}, \\ 1 \le \Phi_{q,K}(M_T) \le \chi_K(M_T) \le \mathcal{B}.\end{gathered} \tag{B.43}$$

Finally, we prove that $\mathcal{E}_{T+1} \le e^{-c_{\mathrm{A}} - c_{\mathrm{M}}\Phi_{q,K}(M_T)}\mathcal{E}_T$. We first verify the conditions of Proposition B.3 at $t = T$. By (B.15) and the induction hypothesis, we have $\|g_j\|_\infty \le C_0$ for $0 \le j \le T$. Applying Lemma B.2 with $H = C_0$, we obtain (B.21) and (B.22) at $t = T$. By (B.43), we can choose $\mu_0 = \Theta(\sqrt{\lambda_W/n})$, independently of T, so that

$$\langle G_T, Q_K(M_T)\rangle \ge \frac{\mu_0}{2}\chi_K(M_T)\sqrt{2\mathcal{L}(\Theta_T)}.$$

It remains to verify the learning rate conditions. By (B.29), we have $A_1, R_1 \lesssim 1$. For the Adam learning rate, we obtain

$$\frac{\gamma A_1}{\varepsilon_0} \le Cc_h \le \frac{\delta_\beta}{4},$$

where the first inequality follows from (B.36) and $A_1 \lesssim 1$, and the second inequality follows from the choice of c_h.

Next, we verify $\underline{\Lambda} \geq \mu_0/2$ for the Muon step. By (B.17) and (B.38), we have $K_R \lesssim \rho_\star$. We then obtain

$$2K_R\sqrt{\mathcal{B}} + \eta R_1 \mathcal{B} \leq C\{\rho_\star\sqrt{\mathcal{B}} + \eta\mathcal{B}\} \leq C(c_0 + c_h)\frac{b\sqrt{\lambda_0}}{(1+\sqrt{\mathcal{B}})\sqrt{n}},$$

where the first inequality follows from $K_R \lesssim \rho_\star$ and $R_1 \lesssim 1$, and the second inequality follows from (B.36) and (B.37). Since $\mu_0 = \Theta(\sqrt{\lambda_W/n})$, we can choose C_0 and c_h so that $2K_R\sqrt{\mathcal{B}} + \eta R_1\mathcal{B} \leq \mu_0/2$. Thus, $\mu_0/2 \leq \underline{\Lambda} \leq \mu_0$. We can therefore apply Proposition B.3 and obtain (B.30) at $t = T$.

We next prove that c_T in (B.30) satisfies $c_T \geq c_{\mathrm{A}} + c_{\mathrm{M}}\Phi_{q,K}(M_T)$. By (B.15) and (B.22), we have $C_L = \Theta(n/\lambda_a)$. Together with (B.29) and (B.36), we obtain $C_E = C_L + \gamma C_m = \Theta(C_L)$. By (B.29) and $\eta_\beta = \delta_\beta/8$, we also have $\eta_\beta = \Theta(1)$ and $\eta_\beta/C_m \gtrsim 1 - \beta_1 \geq b$. Substituting these estimates and $\underline{\Lambda} = \Theta(\sqrt{\lambda_W/n})$ into c_T, we obtain

$$c_T = \min\left\{\frac{\eta_\beta}{C_m}, \frac{\gamma\eta_\beta + \eta\underline{\Lambda}C_L\chi_K(M_T)}{C_E}\right\} \gtrsim \min\left\{b,\ \gamma\frac{\lambda_a}{n} + \eta\sqrt{\frac{\lambda_W}{n}}\Phi_{q,K}(M_T)\right\},$$

where the inequality also follows from $\chi_K(M_T) \geq \Phi_{q,K}(M_T)$ in (B.43). To compare the rate term with b, we obtain

$$\begin{aligned}\gamma\frac{\lambda_a}{n} + \eta\sqrt{\frac{\lambda_W}{n}}\Phi_{q,K}(M_T) &\leq \gamma\frac{\lambda_a}{n} + \eta\mathcal{B}\sqrt{\frac{\lambda_W}{n}} \\ &\leq \frac{c_h b\sqrt{\lambda_0}}{(1+\sqrt{\mathcal{B}})\sqrt{n}}\left(\varepsilon_0 b\frac{\lambda_a}{n} + \sqrt{\frac{\lambda_W}{n}}\right) \lesssim c_h b,\end{aligned}$$

where the first inequality follows from (B.43), the second inequality follows from (B.36), and the last inequality follows from $b \leq 1$ and $\lambda_a, \lambda_W \leq 1/2$. Consequently, we can choose the constant c in c_{A} and c_{M}, independently of T, so that $c_T \geq c_{\mathrm{A}} + c_{\mathrm{M}}\Phi_{q,K}(M_T)$. Applying (B.30), $1 - x \leq e^{-x}$, and the lower bound on c_T, we obtain

$$\mathcal{E}_{T+1} \leq e^{-c_{\mathrm{A}} - c_{\mathrm{M}}\Phi_{q,K}(M_T)}\mathcal{E}_T. \tag{B.44}$$

Since $\Phi_{q,K}(M_T) \geq 1$ by (B.43), this also gives $\mathcal{E}_{T+1} \leq e^{-(c_{\mathrm{A}}+c_{\mathrm{M}})}\mathcal{E}_T$.

Combining (B.44) with the induction hypothesis, we obtain

$$\mathcal{L}(\Theta_{T+1}) \leq \mathcal{E}_{T+1} \leq \mathcal{L}(\Theta_0)e^{-(c_{\mathrm{A}}+c_{\mathrm{M}})(T+1)},$$

where the first inequality follows from (B.28). This completes all induction claims at $T + 1$.

We therefore obtain, for every $T \geq 0$,

$$\mathcal{L}(\Theta_T) \leq \mathcal{E}_T \leq \mathcal{L}(\Theta_0)\exp\left\{-c_{\mathrm{A}}T - c_{\mathrm{M}}\sum_{t=0}^{T-1}\Phi_{q,K}(M_t)\right\}.$$

Taking $T \to \infty$ in (B.40) and (B.41), we also obtain

$$\begin{aligned}\sum_{t=0}^{\infty}\|a_{t+1} - a_t\|_2 &\lesssim \frac{\gamma/\varepsilon_0}{\gamma\lambda_a/n + \eta\sqrt{\lambda_W/n}}, \\ \sum_{t=0}^{\infty}\|W_{t+1} - W_t\|_{\mathrm{F}} &\lesssim \frac{\eta\mathcal{B}}{\gamma\lambda_a/n + \eta\sqrt{\lambda_W/n}}.\end{aligned} \tag{B.45}$$

In particular, Θ_t converges to a finite limit Θ_∞, with $\mathcal{L}(\Theta_\infty) = 0$ by continuity. This completes the proof. □

Proof of Proposition 4.1. By Lemma D.1 and $\mathrm{sr}(M_t) \geq 1 + \nu$, we have

$$\Phi_{\kappa+1,K}(M_t) \geq \Psi_{\kappa+1,K}(\mathrm{sr}(M_t)) \geq \Psi_{\kappa+1,K}(1+\nu) = \sqrt{(1+\nu)\left[1 - \left(\frac{\nu}{1+\nu}\right)^{(\kappa+1)^K}\right]},$$

where the second inequality follows from the monotonicity of $\Psi_{\kappa+1,K}(r)$ in $r \geq 1$. Summing this lower bound over $0 \leq t < T$ and substituting into (6), we obtain (7). □

B.2 Convergence with a Geometrically Decaying Muon Learning Rate

The next theorem shows that a Muon learning rate chosen before training also ensures convergence to zero empirical loss.

Theorem B.1. *Fix $\delta \in (0,1)$ and suppose Assumptions 3.1 and 3.2 hold. We assume $\beta_1^2 < \beta_2$. Let $(a_t, W_t)_{t\geq 0}$ be generated by Algorithm 1 with step sizes $\gamma_t \equiv \gamma = \Theta\Big(\varepsilon_0\sqrt{\lambda_0}/[(1+\sqrt{\mathcal{B}})\sqrt{n}]\Big)$ and $\eta_t := \eta_0\rho^t$, where $\eta_0 = O(\gamma\lambda_a/(\mathcal{B}n))$ and $\rho \in (0,1)$. If $m \gtrsim \frac{n^3}{\lambda_0^3(1-\rho)^2}\log^2\Big(\frac{n}{\lambda_0\delta}\Big)$, then with probability at least $1-\delta$ over the random initialization, we have for every $T \geq 1$,*

$$\mathcal{L}(\Theta_T) \lesssim \exp\left[-\Omega\left(\frac{\gamma\lambda_a}{n}T\right)\right]\mathcal{L}(\Theta_0) + \frac{n\mathcal{B}^2\eta_0^2}{\gamma\lambda_a}\sum_{t=0}^{T-1}\exp\left[-\Omega\left(\frac{\gamma\lambda_a}{n}(T-1-t)\right)\right]\rho^{2t}. \quad \text{(B.46)}$$

Consequently, $\mathcal{L}(\Theta_T) \to 0$ as $T \to \infty$.

Linear convergence rate. Theorem B.1 guarantees that the empirical loss converges to zero exponentially in T. The first term in (B.46) comes from the Adam updates. The second term bounds the cumulative effect of the Muon updates on $\mathcal{L}(\Theta_T)$. Each summand is a product of two geometric factors: ρ^{2t} from the squared learning rate of Muon and the exponential factor over the remaining $T-1-t$ iterations. The sum of these products also decays exponentially in T. Combining the two terms in (B.46), we obtain a guaranteed exponential decay rate of $\Omega(\min\{\gamma\lambda_a/n, -2\log\rho\})$ for the empirical loss $\mathcal{L}(\Theta_T)$.

We first bound the Lyapunov function in the local parameter region.

Proposition B.4. *Let $\delta \in (0,1)$, and suppose Assumptions 3.1 and 3.2 hold and $\beta_1^2 \leq \beta_2 \in (0,1)$. Let $(a_t, W_t)_{t\geq 0}$ be generated by Algorithm 1 with $\gamma_t \equiv \gamma = \Theta\Big(\varepsilon_0\sqrt{\lambda_0}/[(1+\sqrt{\mathcal{B}})\sqrt{n}]\Big)$ and $\eta_t := \eta_0\rho^t$, where $\eta_0 \geq 0$ and $\rho \in (0,1)$. Suppose m and $R_a, R_W \geq 0$ satisfy*

$$m \gtrsim \frac{n^2}{\lambda_0^2}\log^2\left(\frac{8n}{\lambda_0\delta}\right), \qquad \frac{R_a}{\sqrt{m}} \lesssim \sqrt{\frac{\lambda_W}{n}}, \qquad \Delta_X(R_W) \lesssim \lambda_0.$$

With probability at least $1-\delta$ over the paired initialization, if $(a_t, W_t) \in \mathcal{T}(R_W, R_a/\sqrt{m})$ and $\mathcal{L}(\Theta_t) \leq \mathcal{L}(\Theta_0) + 1$ for $0 \leq t \leq T$, then we have, for every $T \geq 1$,

$$\mathcal{E}_T \leq r_{\mathrm{A}}^T\mathcal{L}(\Theta_0) + \frac{C\mathcal{B}^2}{\alpha_{\mathrm{A}}}\sum_{t=0}^{T-1} r_{\mathrm{A}}^{T-1-t}\eta_t^2. \quad \text{(B.47)}$$

Here $r_{\mathrm{A}} := 1 - \alpha_{\mathrm{A}}/2 \in (0,1)$ for some $\alpha_{\mathrm{A}} = \Theta(\gamma\lambda_a/n)$.

Proof. By Lemma B.1, the estimates in Propositions B.1 and B.2 hold with probability at least $1-\delta$ over the paired initialization. First, applying (B.17) at $\Theta = \Theta_t$ with $v = -\gamma D_t\widehat{m}_t$ and $V = -\eta_t Q_K(M_t)$, we obtain, for $0 \leq t < T$,

$$\begin{aligned}\mathcal{L}(\Theta_{t+1}) - \mathcal{L}(\Theta_t) \leq\ & -\gamma\left\langle g_t, D_t\widehat{m}_t\right\rangle - \eta_t\left\langle G_t, Q_K(M_t)\right\rangle \\ & + K_R\eta_t\sqrt{2\mathcal{L}(\Theta_t)}\left\|Q_K(M_t)\right\|_{\mathrm{F}} + C_1\gamma^2\left\|D_t\widehat{m}_t\right\|_2^2 \\ & + C_2\gamma\eta_t\left\|D_t\widehat{m}_t\right\|_2\left\|Q_K(M_t)\right\|_{\mathrm{F}} + C_3\eta_t^2\left\|Q_K(M_t)\right\|_{\mathrm{F}}^2.\end{aligned}$$

For the quadratic terms, we obtain

$$\begin{aligned}& C_1\gamma^2\left\|D_t\widehat{m}_t\right\|_2^2 + C_2\gamma\eta_t\left\|D_t\widehat{m}_t\right\|_2\left\|Q_K(M_t)\right\|_{\mathrm{F}} + C_3\eta_t^2\left\|Q_K(M_t)\right\|_{\mathrm{F}}^2 \\ & \leq A_1\gamma^2\left\|D_t\widehat{m}_t\right\|_2^2 + \frac{R_1}{2}\eta_t^2\left\|Q_K(M_t)\right\|_{\mathrm{F}}^2 \\ & \leq \frac{\gamma^2 A_1}{\varepsilon_0}S_t + \frac{R_1}{2}\eta_t^2\left\|Q_K(M_t)\right\|_{\mathrm{F}}^2,\end{aligned} \quad \text{(B.48)}$$

where the first inequality follows from Young's inequality and (B.29), and the second inequality follows from (B.31).

For the hidden update, we have

$$\|G_t\|_{\mathrm{F}} \le \frac{\|a_t\|_2 \|X\|_2 \|r_t\|_2}{n\sqrt{m}} \lesssim \sqrt{\mathcal{L}(\Theta_t)},$$

where the first inequality follows from $(G_t)_{j,:} = a_{t,j} r_t^\top D_{X,j}(W_t) X/(n\sqrt{m})$ and $\|D_{X,j}(W_t)\|_2 \le 1$ for $j \in [m]$, and the second inequality follows from $\|a_t\|_2 \le \sqrt{m} + R_a \lesssim \sqrt{m}$, $\|X\|_2 \le \sqrt{n}$, and $\|r_t\|_2^2 = 2n\mathcal{L}(\Theta_t)$. We therefore obtain

$$\begin{aligned} &-\eta_t \langle G_t, Q_K(M_t)\rangle + K_R \eta_t \sqrt{2\mathcal{L}(\Theta_t)} \|Q_K(M_t)\|_{\mathrm{F}} \\ &\quad \le \eta_t \left(\|G_t\|_{\mathrm{F}} + K_R \sqrt{2\mathcal{L}(\Theta_t)} \right) \|Q_K(M_t)\|_{\mathrm{F}} \le C\mathcal{B}\eta_t \sqrt{\mathcal{L}(\Theta_t)}, \end{aligned} \tag{B.49}$$

where the first inequality follows from Cauchy–Schwarz inequality, and the second inequality follows from $\|G_t\|_{\mathrm{F}} \lesssim \sqrt{\mathcal{L}(\Theta_t)}$, $\|Q_K(M_t)\|_{\mathrm{F}} \le \mathcal{B}$ from Lemma D.1, and $K_R \le C\{R_a/\sqrt{m} + \sqrt{\epsilon_X(R_W)}\} \lesssim 1$ from Proposition B.2. Combining (B.48) and (B.49), we obtain

$$\mathcal{L}(\Theta_{t+1}) - \mathcal{L}(\Theta_t) \le -\gamma \langle g_t, D_t \widehat{m}_t \rangle + \frac{\gamma^2 A_1}{\varepsilon_0} S_t + C\mathcal{B}\eta_t \sqrt{\mathcal{L}(\Theta_t)} + C\mathcal{B}^2 \eta_t^2. \tag{B.50}$$

We next bound $\mathcal{E}_{t+1} - \mathcal{E}_t$. We have

$$\begin{aligned} \mathcal{E}_{t+1} - \mathcal{E}_t &= \mathcal{L}(\Theta_{t+1}) - \mathcal{L}(\Theta_t) + \gamma(b_{t+1}^a + \omega_\beta) S_t - \gamma(b_t^a + \omega_\beta) S_{t-1} \\ &\le -\gamma \left(\delta_\beta - \omega_\beta - \frac{\gamma A_1}{\varepsilon_0} \right) S_t - \gamma \omega_\beta S_{t-1} + C\mathcal{B}\eta_t \sqrt{\mathcal{L}(\Theta_t)} + C\mathcal{B}^2 \eta_t^2 \\ &\le -\gamma \eta_\beta (S_t + S_{t-1}) + C\mathcal{B}\eta_t \sqrt{\mathcal{E}_t} + C\mathcal{B}^2 \eta_t^2, \end{aligned}$$

where the equality follows from (B.28), the first inequality follows from (B.50) and (B.21), and the second inequality follows from (B.36) and $\mathcal{L}(\Theta_t) \le \mathcal{E}_t$.

By (B.15), we have

$$\frac{\lambda_a}{n} \mathcal{L}(\Theta_t) \lesssim \|g_t\|_2^2, \qquad \max_{0 \le s \le t} \|g_s\|_\infty \lesssim 1,$$

where the second inequality also follows from $\mathcal{L}(\Theta_s) \le \mathcal{L}(\Theta_0) + 1 \lesssim 1$ for $0 \le s \le t$. By Lemma B.2, we obtain (B.22) with $C_L = \Theta(n/\lambda_a)$. Combining (B.22) and (B.34), we obtain

$$\mathcal{E}_t \le (C_L + \gamma C_m)(S_t + S_{t-1}) = C_E (S_t + S_{t-1}).$$

Set $\alpha_{\mathrm{A}} := \gamma \eta_\beta / C_E = \Theta(\gamma \lambda_a / n)$. We then obtain

$$\mathcal{E}_{t+1} \le (1 - \alpha_{\mathrm{A}}) \mathcal{E}_t + C\mathcal{B}\eta_t \sqrt{\mathcal{E}_t} + C\mathcal{B}^2 \eta_t^2 \le r_{\mathrm{A}} \mathcal{E}_t + \frac{C\mathcal{B}^2}{\alpha_{\mathrm{A}}} \eta_t^2, \tag{B.51}$$

where the first inequality follows from $\mathcal{E}_t \le C_E(S_t + S_{t-1})$, and the second inequality follows from Young's inequality, $r_{\mathrm{A}} = 1 - \alpha_{\mathrm{A}}/2$, and $0 < \alpha_{\mathrm{A}} \le \eta_\beta / C_m \le 1/2$. Iterating (B.51) with $\mathcal{E}_0 = \mathcal{L}(\Theta_0)$ from (B.28), we obtain (B.47), completing the proof. □

Now we apply Proposition B.4 to complete the proof of Theorem B.1.

Proof of Theorem B.1. First, we prove by induction that, for every $t \ge 0$, Θ_t lies in the parameter region

$$\|a - a_0\|_2 \le \frac{Cn}{\lambda_0 (1 - \rho)}, \qquad \|W - W_0\|_{\mathrm{F}} \le \frac{C}{1 - \rho}, \tag{B.52}$$

and satisfies $\mathcal{E}_t \le \mathcal{L}(\Theta_0) + 1$, where $C > 0$ is a constant. At $t = 0$, both bounds follow from initialization and $\mathcal{E}_0 = \mathcal{L}(\Theta_0)$ in (B.28). Assume they hold for $0 \le t \le T$. We verify them at $t = T + 1$.

By (B.39), (B.15), and $\mathcal{L}(\Theta_t) \le \mathcal{E}_t$, we obtain

$$\sum_{t=0}^{T} \|a_{t+1} - a_t\|_2 \le \frac{C\gamma}{\varepsilon_0 (1 - \beta_1)} \sum_{t=0}^{T} \sqrt{\mathcal{E}_t}, \tag{B.53}$$

It remains to bound $\sum_{t=0}^{T}\sqrt{\mathcal{E}_t}$. Under the induction hypothesis, it is easy to check that (B.51) holds for $0 \le t < T$. Then we obtain

$$\sqrt{\mathcal{E}_{t+1}} \le \sqrt{r_{\mathrm{A}}}\sqrt{\mathcal{E}_t} + \frac{C\mathcal{B}}{\sqrt{\alpha_{\mathrm{A}}}}\eta_t, \qquad 0 \le t < T.$$

Writing $H_T := \sum_{t=0}^{T-1}\eta_t$, summing over $0 \le t < T$, and rearranging, we obtain

$$(1-\sqrt{r_{\mathrm{A}}})\sum_{t=0}^{T}\sqrt{\mathcal{E}_t} \le \sqrt{\mathcal{E}_0} - \sqrt{r_{\mathrm{A}}}\sqrt{\mathcal{E}_T} + \frac{C\mathcal{B}H_T}{\sqrt{\alpha_{\mathrm{A}}}} \le \sqrt{\mathcal{L}(\Theta_0)} + \frac{C\mathcal{B}\eta_0}{(1-\rho)\sqrt{\alpha_{\mathrm{A}}}}, \tag{B.54}$$

where the second inequality follows from $\mathcal{E}_0 = \mathcal{L}(\Theta_0)$, and $H_T \le \eta_0/(1-\rho)$. Combining (B.53) and (B.54) with the step size choices, we obtain $\sum_{t=0}^{T}\|a_{t+1}-a_t\|_2 \lesssim n/[\lambda_0(1-\rho)]$. By the Muon update and Lemma D.1, we also have $\sum_{t=0}^{T}\|W_{t+1}-W_t\|_{\mathrm{F}} \le \mathcal{B}\eta_0/(1-\rho) \lesssim 1/(1-\rho)$. Therefore, by the triangle inequality and the two movement bounds, we obtain

$$\begin{aligned}\|a_{T+1}-a_0\|_2 &\le \sum_{t=0}^{T}\|a_{t+1}-a_t\|_2 \le \frac{Cn}{\lambda_0(1-\rho)},\\ \|W_{T+1}-W_0\|_{\mathrm{F}} &\le \sum_{t=0}^{T}\|W_{t+1}-W_t\|_{\mathrm{F}} \le \frac{C}{1-\rho}.\end{aligned} \tag{B.55}$$

Thus, Θ_{T+1} satisfies (B.52). The segment joining Θ_T and Θ_{T+1} therefore lies in this region by convexity. Next, we take

$$0 \le \eta_0 \le \frac{\alpha_{\mathrm{A}}}{\sqrt{2C}\mathcal{B}} = \Theta\left(\frac{\gamma\lambda_a}{\mathcal{B}n}\right),$$

where C is the constant in (B.51). Applying (B.51) at $t = T$, we obtain

$$\mathcal{E}_{T+1} \le r_{\mathrm{A}}\mathcal{E}_T + \frac{C\mathcal{B}^2\eta_0^2}{\alpha_{\mathrm{A}}}\rho^{2T} \le r_{\mathrm{A}}\big(\mathcal{L}(\Theta_0)+1\big) + \frac{\alpha_{\mathrm{A}}}{2} = \mathcal{L}(\Theta_0) + 1 - \frac{\alpha_{\mathrm{A}}}{2}\mathcal{L}(\Theta_0) \le \mathcal{L}(\Theta_0)+1,$$

where the second inequality follows from the induction hypothesis, and $C\mathcal{B}^2\eta_0^2/\alpha_{\mathrm{A}} \le \alpha_{\mathrm{A}}/2$, and the equality follows from $r_{\mathrm{A}} = 1-\alpha_{\mathrm{A}}/2$. This completes the induction.

Finally, applying (B.47) with $\eta_t = \eta_0\rho^t$, we obtain, for every $T \ge 1$,

$$\begin{aligned}\mathcal{L}(\Theta_T) &\le e^{-\alpha_{\mathrm{A}}T/2}\mathcal{L}(\Theta_0) + \frac{C\mathcal{B}^2\eta_0^2}{\alpha_{\mathrm{A}}}\sum_{t=0}^{T-1}e^{-\alpha_{\mathrm{A}}(T-1-t)/2}\rho^{2t}\\ &\le e^{-c\gamma\lambda_a T/n}\mathcal{L}(\Theta_0) + \frac{Cn\mathcal{B}^2\eta_0^2}{\gamma\lambda_a}\sum_{t=0}^{T-1}e^{-c\gamma\lambda_a(T-1-t)/n}\rho^{2t},\end{aligned}$$

where the first inequality follows from (B.28) and $r_{\mathrm{A}}^j = (1-\alpha_{\mathrm{A}}/2)^j \le e^{-\alpha_{\mathrm{A}}j/2}$ for $j \ge 0$, and the second inequality follows from $\alpha_{\mathrm{A}} = \Theta(\gamma\lambda_a/n)$, with constants $c, C > 0$ independent of T. This proves (B.46). Since $r_{\mathrm{A}}, \rho^2 < 1$, we have $\sum_{t=0}^{T-1} r_{\mathrm{A}}^{T-1-t}\rho^{2t} \le T\max\{r_{\mathrm{A}}, \rho^2\}^{T-1} \longrightarrow 0$, so $\mathcal{L}(\Theta_T) \to 0$. This completes the proof. □

C Proofs of Generalization Results

We prove Theorem 4.2 by combining the parameter path bounds with a decomposition of the output weight displacement relative to the feature space at each iteration. We use the statistical lemmas in Appendix D to obtain the population loss bound. We then bound the data-dependent kernel complexity of polynomial targets and verify the kernel condition for uniform spherical inputs.

C.1 Proof of Theorem 4.2

For each $T \ge 0$, let $Z_T := Z_X(W_T)$, $b_T := a_T - a_0$, and $P_T := Z_T^\top(Z_T Z_T^\top)^{-1}Z_T$ whenever $Z_T Z_T^\top$ is invertible. The next proposition controls the components of b_T inside and outside the row space of Z_T. Its proof also verifies that P_T is well defined along the trajectory.

Proposition C.1. *Suppose the assumptions and width requirement of Theorem 4.2 hold. Take γ as in* (B.36)*, with λ_0 replaced by λ, and set $\eta = \tau\gamma\lambda^{3/2}/[\mathcal{B}(1+\varepsilon_0)n^{3/2}]$ for a sufficiently small $\tau > 0$. Assume the input matrix X satisfies $\lambda_0(X) \geq \lambda$. With probability at least $1 - \delta/3$ over the paired initialization, the following bounds hold for every $T \geq 0$:*

$$\|W_T - W_0\|_{\mathrm{F}} \lesssim \sqrt{\frac{\lambda}{n}}, \qquad \frac{\|W_T\|_2}{\sqrt{m}} \lesssim 1,$$
$$\|P_T b_T\|_2^2 \lesssim 1 + \mathfrak{C}_a + \frac{n}{\lambda}\mathcal{L}(\Theta_T), \qquad \|(I_m - P_T)b_T\|_2^2 \lesssim \frac{1}{n}\log\left(\frac{2e(n+d)m}{\lambda\delta}\right).$$

Proof. The chosen learning rates satisfy the conditions of Theorem 4.1 on the sample event $\lambda_0(X) \geq \lambda$. We apply that theorem with failure probability $\delta/9$, absorbing the constant change in the logarithm into the width requirement. By (6), we may take $r := \exp\{-c\gamma\lambda/[(1+\varepsilon_0)n]\} \in (0,1)$ so that

$$\mathcal{L}(\Theta_t) \leq \mathcal{L}(\Theta_0) r^{2t}, \qquad 1 - r \gtrsim \frac{\gamma\lambda}{(1+\varepsilon_0)n}. \tag{C.1}$$

Since $\mathcal{L}(\Theta_0) \leq C^2/2$, the Muon update and Lemma D.1 imply

$$\sum_{t=0}^{\infty} \|W_{t+1} - W_t\|_{\mathrm{F}} \lesssim \frac{\eta\mathcal{B}}{1-r} \lesssim \tau\sqrt{\frac{\lambda}{n}}. \tag{C.2}$$

Gaussian matrix concentration also bounds $\|W_0\|_2/\sqrt{m}$ by a constant. The first two bounds in the proposition follow.

We next control the part of b_T seen by the current feature matrix. Let $\mathbf{K}_T := m^{-1}Z_T Z_T^\top$ and $r_t := m^{-1/2}Z_t a_t - y$. The initialization estimates (B.7) and (B.8) imply $\mathbf{K}_0 \succeq 3\mathbf{K}_a/4$ and $\|Z_0\|_2 \lesssim \sqrt{mn}$. By the Lipschitz property of ReLU and (C.2), we obtain

$$\|\mathbf{K}_T - \mathbf{K}_0\|_2 \lesssim \frac{n\|W_T - W_0\|_{\mathrm{F}}}{\sqrt{m}} + \frac{n\|W_T - W_0\|_{\mathrm{F}}^2}{m} \leq \frac{\lambda}{4}.$$

The width requirement ensures the last inequality. Thus $\mathbf{K}_T \succeq \mathbf{K}_a/2$, so P_T is well defined. Since $Z_0 a_0 = 0$ and $\|a_0\|_2 = \sqrt{m}$, we also have $\|Z_T a_0\|_2/\sqrt{m} \leq \sqrt{n}\|W_T - W_0\|_{\mathrm{F}} \lesssim \tau\sqrt{\lambda}$. The minimum norm identity for P_T now gives

$$\|P_T b_T\|_2 = \left\|\mathbf{K}_T^{-1/2}(y + r_T - m^{-1/2}Z_T a_0)\right\|_2 \lesssim \sqrt{\mathfrak{C}_a} + \sqrt{\frac{n\mathcal{L}(\Theta_T)}{\lambda}} + \tau.$$

Here we used $\mathbf{K}_T^{-1} \preceq 2\mathbf{K}_a^{-1}$ and $\|r_T\|_2^2 = 2n\mathcal{L}(\Theta_T)$. Squaring proves the bound on $\|P_T b_T\|_2^2$.

It remains to control the part of the Adam update outside the row space of Z_T. Set $E_t := D_t - \varepsilon_0^{-1} I_m$ and write the corrected Adam momentum as $\widehat{m}_t = \sum_{s=0}^{t} \omega_{ts} g_s$, where $\omega_{ts} := (1-\beta_1)\beta_1^{t-s}/(1-\beta_1^{t+1})$. Because $(I_m - P_T)Z_T^\top = 0$ and $g_s = Z_s^\top r_s/(n\sqrt{m})$, we have, for $s < T$,

$$\|(I_m - P_T)g_s\|_2 \leq \frac{\sqrt{2\mathcal{L}(\Theta_s)}}{\sqrt{m}} \sum_{u=s}^{T-1} \|W_{u+1} - W_u\|_{\mathrm{F}}.$$

Consequently, the Adam update satisfies

$$\|(I_m - P_T)b_T\|_2 \leq \frac{\gamma}{\varepsilon_0} \sum_{s=0}^{T-1} \left(\sum_{t=s}^{T-1} \omega_{ts}\right) \|(I_m - P_T)g_s\|_2 + \gamma \sum_{t=0}^{T-1} \|E_t \widehat{m}_t\|_2. \tag{C.3}$$

The first term in (C.3) measures changes in the feature matrix. By the bound on $\|(I_m - P_T)g_s\|_2$, (C.1), the Muon update, and $\sum_{t=s}^{T-1} \omega_{ts} \leq (1-\beta_1)^{-1}$, we bound this term by $\gamma\eta\mathcal{B}/[\varepsilon_0(1-r)^2\sqrt{m}] \lesssim \tau(1+\varepsilon_0)\sqrt{n/(\lambda m)}/\varepsilon_0$.

For the second term, Gaussian concentration and a union bound imply $\max_j \|(Z_0)_{:j}\|_2 \lesssim \sqrt{n\log(2m/\delta)}$. The hidden path bound (C.2) extends the column estimate to every Z_t. Together with $\|Z_0\|_2 \lesssim \sqrt{mn}$ and (C.1), we obtain $\sup_{t\geq 0}\|g_t\|_\infty \lesssim \sqrt{\log(2m/\delta)/m}$ and

$\|g_t\|_2 \lesssim r^t$. Since $|(E_t)_{jj}| = \sqrt{(\widehat{v}_t)_j}/[\varepsilon_0(\sqrt{(\widehat{v}_t)_j} + \varepsilon_0)]$, Lemma D.6 implies $\|E_t\widehat{m}_t\|_2 \lesssim \varepsilon_0^{-2}\sqrt{\log(2m/\delta)/m}\sqrt{\|\widehat{v}_t\|_1}$. The weight of $g_s^{\odot 2}$ in $\widehat{v}_t$ is at most β_2^{t-s}, so $\sum_{t\geq 0}\sqrt{\|\widehat{v}_t\|_1} \leq (1-\sqrt{\beta_2})^{-1}\sum_{s\geq 0}\|g_s\|_2 \lesssim (1-r)^{-1}$. Consequently, by (C.1), we obtain

$$\gamma\sum_{t=0}^{T-1}\|E_t\widehat{m}_t\|_2 \lesssim \frac{(1+\varepsilon_0)n}{\varepsilon_0^2\lambda\sqrt{m}}\sqrt{\log\left(\frac{2m}{\delta}\right)}.$$

We substitute the two bounds into (C.3) and use the width requirement to obtain

$$\|(I_m - P_T)b_T\|_2^2 \lesssim \tau^2\frac{(1+\varepsilon_0)^2 n}{\varepsilon_0^2\lambda m} + \frac{(1+\varepsilon_0)^2 n^2}{\varepsilon_0^4\lambda^2 m}\log\left(\frac{2m}{\delta}\right) \lesssim \frac{1}{n}\log\left(\frac{2e(n+d)m}{\lambda\delta}\right).$$

We also assign failure probability at most $\delta/9$ each to the initial Gram matrix and Gaussian norm estimates. Together with the convergence event, these bounds hold for every T with conditional probability at least $1-\delta/3$, completing the proof. □

Proof of Theorem 4.2. The kernel condition in Assumption 4.1 fails with probability at most $\delta/3$. On its complementary event, we apply Proposition C.1. For each T, let $\widetilde{a}_T := a_0 + P_T b_T$. Since $Z_T(I_m - P_T)b_T = 0$, the projected predictor has the same empirical loss as Θ_T.

Conditional on the initialization, consider $\mathcal{G}_{\mathrm{ctr}}(A, R, R)$ for a deterministic hidden radius $R \asymp \sqrt{\lambda/n}$ and output radii A on a dyadic grid from 1 to $C(1+\sqrt{n/\lambda})$. The range covers $\|\widetilde{a}_T - a_0\|_2$ for every T, since Proposition C.1 gives

$$\|\widetilde{a}_T - a_0\|_2^2 \lesssim 1 + \mathfrak{C}_a + \frac{n}{\lambda}\mathcal{L}(\Theta_T),$$

while $\mathfrak{C}_a \leq C^2 n/\lambda$ and $\mathcal{L}(\Theta_T) \leq \mathcal{L}(\Theta_0) \leq C^2/2$. The smallest grid radius covering $\widetilde{a}_T$ is at most $C(1+\sqrt{\mathfrak{C}_a + n\mathcal{L}(\Theta_T)/\lambda})$. For each grid radius, Lemma D.7 and the bound on $\|W_0\|_2/\sqrt{m}$ imply $\mathfrak{R}_n\{\mathcal{G}_{\mathrm{ctr}}(A,R,R)\} \lesssim A/\sqrt{n}$ and $|f(x)| \lesssim A$. We apply Lemma D.8 to the entire grid, allocating total failure probability $\delta/3$. Since the statistical event covers every function in these classes, it holds simultaneously for all T without a union bound over iterations. Using Young's inequality to absorb the square root term in that lemma, we obtain

$$\mathcal{L}_{\mathcal{D}}((W_T, \widetilde{a}_T)) \lesssim \mathcal{L}(\Theta_T) + \frac{1 + \mathfrak{C}_a + n\mathcal{L}(\Theta_T)/\lambda}{n}\log^3\left(\frac{(n+d)m}{\lambda\delta}\right). \tag{C.4}$$

The population predictions of $(W_T, \widetilde{a}_T)$ and Θ_T differ by $m^{-1/2}[(I_m - P_T)b_T]^\top\sigma(W_T x)$. By Proposition C.1 and $\|x\|_2 \leq 1$, we have

$$\frac{1}{2}\mathbb{E}_x\left|f_{\Theta_T}(x) - f_{(W_T,\widetilde{a}_T)}(x)\right|^2 \lesssim \|(I_m - P_T)b_T\|_2^2 \lesssim \frac{1}{n}\log\left(\frac{2e(n+d)m}{\lambda\delta}\right).$$

Applying the squared triangle inequality in $L^2(\mathcal{D})$ to this estimate and (C.4), we obtain

$$\mathcal{L}_{\mathcal{D}}(\Theta_T) \lesssim \left[\frac{\mathcal{L}(\Theta_T)}{\lambda} + \frac{1+\mathfrak{C}_a}{n}\right]\log^3\left(\frac{(n+d)m}{\lambda\delta}\right), \qquad T \geq 0. \tag{C.5}$$

This proves (8). A union bound over the kernel condition, initialization, and statistical events gives total failure probability at most δ. The event is uniform in T, so the bound also holds at stopping times selected from the training history. Finally, (B.45) implies $\Theta_T \to \Theta_\infty$. By (C.1), continuity, Fatou's lemma, and (C.5), we obtain $\mathcal{L}(\Theta_\infty) = 0$ and

$$\mathcal{L}_{\mathcal{D}}(\Theta_\infty) \lesssim \frac{1+\mathfrak{C}_a}{n}\log^3\left(\frac{(n+d)m}{\lambda\delta}\right).$$

□

C.2 Examples with Bounded Data-Dependent Complexity

We give target functions for which the data-dependent complexity $\mathfrak{C}_a$ is bounded, independently of the sample size.

Polynomial Targets. We first consider $f(x) := \alpha(u^\top x)^p$ with $p \in \{0, 1, 2, 4, 6, \ldots\}$. To state the general bound, define the Hermite polynomial H_p by $H_p(z) := (-1)^p e^{z^2/2} \frac{\mathrm{d}^p}{\mathrm{d}z^p} e^{-z^2/2}$, and let $\widehat{\sigma}_p := \mathbb{E}[H_p(G)\sigma(G)]$ for $G \sim \mathcal{N}(0, 1)$.

Lemma C.1. *The function f on $\mathbb{S}^{d-1}$ satisfies $\|f\|_{\mathcal{H}_\mathrm{R}} \le \sqrt{p!}|\alpha| \|u\|_2^p / |\widehat{\sigma}_p|$. Consequently, any sample with $\|x_i\|_2 = 1$, $y_i = f(x_i)$, and $\mathbf{K}_a \succ 0$ obeys $\mathfrak{C}_a \le \frac{p!}{\widehat{\sigma}_p^2} \alpha^2 \|u\|_2^{2p}$.*

Proof. By Gaussian integration, we have

$$\widehat{\sigma}_0 = \widehat{\sigma}_2 = \frac{1}{\sqrt{2\pi}}, \qquad \widehat{\sigma}_1 = \frac{1}{2}, \qquad \widehat{\sigma}_{2k} = -(2k-3)\widehat{\sigma}_{2k-2} \quad (k \ge 2), \tag{C.6}$$

so $\widehat{\sigma}_p \neq 0$ for every allowed degree. For $p = 0$, the bound follows by taking the constant coefficient $c(w) = \alpha/\widehat{\sigma}_0$. If $p \ge 1$ and $u = 0$, then $f = 0$. It remains to consider $p \ge 1$ and $u \neq 0$. Set $\bar{u} = u/\|u\|_2$. Since $\bar{u}^\top w$ and $x^\top w$ are standard Gaussian with correlation $\bar{u}^\top x$, the Hermite correlation identity implies

$$\mathbb{E}_w\big[H_p(\bar{u}^\top w)\sigma(w^\top x)\big] = \widehat{\sigma}_p(\bar{u}^\top x)^p.$$

Multiplying this identity by $\alpha \|u\|_2^p / \widehat{\sigma}_p$ and using $u^\top x = \|u\|_2 \, \bar{u}^\top x$, we obtain

$$f(x) = \mathbb{E}_w[c(w)\sigma(w^\top x)],$$

where $c(w) := \frac{\alpha\|u\|_2^p}{\widehat{\sigma}_p} H_p(\bar{u}^\top w)$. By the feature representation of the RKHS and $\mathbb{E}[H_p(G)^2] = p!$, we have

$$\|f\|_{\mathcal{H}_\mathrm{R}} \le \{\mathbb{E}[c(w)^2]\}^{1/2} = \frac{\sqrt{p!}}{|\widehat{\sigma}_p|}|\alpha| \|u\|_2^p .$$

The empirical bound follows from $y^\top \mathbf{K}_a^{-1} y = \min\{\|g\|_{\mathcal{H}_\mathrm{R}}^2 : g(x_i) = y_i\} \le \|f\|_{\mathcal{H}_\mathrm{R}}^2$. □

The same bound extends to finite sums of these targets. Consider the function

$$f_\star(x) := \sum_{r=1}^R \alpha_r (u_r^\top x)^{p_r}, \qquad p_r \in \{0, 1, 2, 4, 6, \ldots\}, \tag{C.7}$$

and define

$$S_\star := \sum_{r=1}^R \frac{\sqrt{p_r!}}{|\widehat{\sigma}_{p_r}|}|\alpha_r| \|u_r\|_2^{p_r} .$$

Applying Lemma C.1 and the triangle inequality, we obtain

$$\|f_\star\|_{\mathcal{H}_\mathrm{R}} \le S_\star, \qquad f_{\star,X}^\top \mathbf{K}_a^{-1} f_{\star,X} \le S_\star^2.$$

Moreover, $\sup_{x \in \mathbb{S}^{d-1}} |f_\star(x)| \le \sum_{r=1}^R |\alpha_r| \|u_r\|_2^{p_r}$. These examples parallel Arora et al. (2019, Section 6), with the output activation kernel in place of the hidden tangent kernel.

Quadratic Targets. We next consider target functions of the form $f_\star(x) = c + b^\top x + x^\top B x$, where B is symmetric. The following lemma bounds $\mathfrak{C}_a$ in terms of c, b, and B.

Lemma C.2. *For any sample with $\|x_i\|_2 = 1$, $y_i = f_\star(x_i)$, and $\mathbf{K}_a \succ 0$, we have*

$$\mathfrak{C}_a \le \|f_\star\|_{\mathcal{H}_\mathrm{R}}^2 \le 2\pi c^2 + 4\|b\|_2^2 + 4\pi \|B\|_\mathrm{F}^2 .$$

Moreover, $|f_\star(x)| \le |c| + \|b\|_2 + \|B\|_2$ for every $x \in \mathbb{S}^{d-1}$.

Proof. For $w \sim \mathcal{N}(0, I_d)$, define

$$h(w) := \sqrt{2\pi}\, c + 2b^\top w + \sqrt{2\pi}\big(w^\top B w - \operatorname{tr} B\big).$$

For a unit vector x and $G = w^\top x$, we have $\mathbb{E}[w \mid G] = Gx$ and $\mathbb{E}[w^\top B w - \operatorname{tr} B \mid G] = (G^2 - 1)x^\top B x$. Using these conditional expectations and the first three Hermite coefficients in (C.6), we obtain

$$\mathbb{E}_w[h(w)\sigma(w^\top x)] = c + b^\top x + x^\top B x = f_\star(x).$$

The three terms in $h(w)$ are pairwise orthogonal under $w \sim \mathcal{N}(0, I_d)$. Thus

$$\begin{aligned}\|f_\star\|_{\mathcal{H}_{\mathrm{R}}}^2 &\le \mathbb{E}[h(w)^2] = 2\pi c^2 + 4\mathbb{E}[(b^\top w)^2] + 2\pi\mathbb{E}[(w^\top B w - \operatorname{tr} B)^2] \\ &= 2\pi c^2 + 4\|b\|_2^2 + 4\pi\|B\|_{\mathrm{F}}^2,\end{aligned}$$

where the last equality follows from the Gaussian moments $\mathbb{E}[(b^\top w)^2] = \|b\|_2^2$ and $\mathbb{E}[(w^\top B w - \operatorname{tr} B)^2] = 2\|B\|_{\mathrm{F}}^2$. The interpolation argument in Lemma C.1 implies $\mathfrak{C}_a \le \|f_\star\|_{\mathcal{H}_{\mathrm{R}}}^2$. Finally, the label bound follows from $|b^\top x| \le \|b\|_2$ and $|x^\top B x| \le \|B\|_2$ for $\|x\|_2 = 1$. □

C.3 Uniform Spherical Inputs

Proposition C.2. *Suppose the input marginal of $\mathcal{D}$ is uniform on $\mathbb{S}^{d-1}$, and the labels satisfy $|y| \le C$. There are numerical constants $c \in (0, 1/2]$ and $C_1 > 0$ such that, if $d \ge C_1\{n + \log(1/\rho)\}$ for $\rho \in (0,1)$, then*

$$\mathbb{P}_X(\lambda_0(X) \ge c) \ge 1 - \rho.$$

Consequently, Assumption 4.1 holds.

Proof. The columns of $\sqrt{d}\,X^\top$ are independent, isotropic sub-Gaussian vectors in $\mathbb{R}^d$, each of norm $\sqrt{d}$. Applying the singular value bound for matrices with such columns (Vershynin, 2018, Chapter 4), we obtain $XX^\top \succeq c_0 I_n$ with probability at least $1 - \rho$ under the stated dimension condition, where $c_0 > 0$ is numerical. Applying Lemma D.3, we obtain $\lambda_0(X) \ge \lambda_{\min}(XX^\top)/4 \ge c_0/4$. The event depends only on the inputs, so the conditional distribution of the labels does not affect the conclusion. □

Proof of Proposition 4.2. Apply Proposition C.2 with $\rho = \delta/6$. Under the stated dimension condition, we have $\lambda_0(X) \ge c$ with probability at least $1 - \delta/6$ for a numerical constant $c > 0$. Taking $\lambda = c$ in Theorem 4.2 and using confidence parameter $\delta/2$, we obtain (8) with a numerical kernel margin. On the sample event $\lambda_0(X) \ge c$, the stated width and step sizes also satisfy the conditions of Theorem 4.1. Applying it conditionally with failure probability $\delta/3$, we obtain (6) with $\lambda_a, \lambda_W \ge c$. The union bound gives joint success probability at least $1 - \delta/2 - \delta/3 - \delta/6 = 1 - \delta$. Substituting the resulting empirical loss bound into (8) and absorbing numerical constants into the asymptotic notation, we obtain (9). □

D Technical Lemmas

In this section, we gather several technical lemmas used in our analysis.

D.1 Newton-Schulz Iterations

We begin with basic properties of the Newton-Schulz iteration.

Proposition D.1. *For every integer $\kappa \ge 1$, the polynomial P_κ (2) satisfies*

$$1 \le P_\kappa(s) \le \Lambda_\kappa, \qquad 0 \le 1 - sP_\kappa^2(s) \le (1-s)^{\kappa+1} \quad \textit{for every } s \in [0,1],$$

where $\Lambda_\kappa = (2\kappa+1)\frac{\binom{2\kappa}{\kappa}}{4^\kappa}$.

Proof. Let $u = 1 - s$. By the definition of P_κ, we have $1 \le P_\kappa(1-u) \le P_\kappa(0)$ for $u \in [0,1]$. For $0 \le u < 1$, we also have

$$\frac{1-u^{\kappa+1}}{1-u} = \sum_{j=0}^{\kappa} u^j \le P_\kappa^2(1-u) \le \frac{1}{1-u},$$

where the last inequality follows from $P_\kappa(1-u) \le (1-u)^{-1/2}$. Rearranging and using continuity at $u = 1$, we obtain the desired bound. The identity $P_\kappa(0) = \Lambda_\kappa$ follows by induction on κ. □

Write $\Lambda := \Lambda_\kappa = P_\kappa(0)$. We define the spectral score and the associated Frobenius cost ratio by

$$\chi_K(B) := \frac{\langle B, Q_K(B)\rangle}{\|B\|_{\mathrm{F}}}, \qquad \zeta_K(B) := \frac{\|Q_K(B)\|_{\mathrm{F}}^2}{\chi_K(B)}.$$

By convention, we set $\chi_K(0) = 0$ and $\zeta_K(0) = 0$. For $q = \kappa + 1$, define

$$\Psi_{q,K}(r) := \sqrt{r\{1 - (1 - r^{-1})^{q^K}\}}, \qquad r \geq 1.$$

The following lemma bounds the spectral score $\chi_K(B)$ and Frobenius cost ratio $\zeta_K(B)$ of the Newton-Schulz direction.

Lemma D.1. *Let $\kappa \geq 1$ and $K \geq 0$ be integers. For every nonzero matrix B, we have*

$$\|Q_K(B)\|_2 \leq 1, \qquad \zeta_K(B) \leq \min\left\{\Lambda^K, \frac{\mathrm{rank}(B)}{\chi_K(B)}\right\},$$
$$\Lambda^K \geq \chi_K(B) \geq \Phi_{q,K}(B) \geq \Psi_{q,K}(\mathrm{sr}(B)) \geq 1.$$

Consequently, we have

$$\frac{\|Q_K(B)\|_{\mathrm{F}}}{\chi_K(B)} = \sqrt{\frac{\zeta_K(B)}{\chi_K(B)}} \leq \sqrt{\frac{\Lambda^K}{\chi_K(B)}} \leq \Lambda^{K/2}.$$

Proof. For an initial singular value $s \in [0, 1]$, let $h_0(s) = s$ and $h_{k+1}(s) = h_k(s)P_\kappa(h_k^2(s))$. By Proposition D.1, we have

$$s \leq h_K(s) \leq 1, \qquad 1 - h_K^2(s) \leq (1 - s^2)^{q^K}, \qquad h_K(s) \leq \Lambda^K s. \tag{D.1}$$

For simplicity, we write $s_i = s_i(B)$. Since the iteration preserves singular vectors, we have $\chi_K(B) = \sum_i s_i h_K(s_i)$ and $\|Q_K(B)\|_{\mathrm{F}}^2 = \sum_i h_K^2(s_i)$. By (D.1) and $\sum_i s_i^2 = 1$, we obtain $\|Q_K(B)\|_2 \leq 1$ and $\chi_K(B) \leq \Lambda^K$. Summing $h_K^2(s_i) \leq \Lambda^K s_i h_K(s_i)$ and $h_K^2(s_i) \leq 1$, we obtain both bounds on $\zeta_K(B)$.

Set $N = q^K$. We have $\Psi_{q,K}^2(r) = \sum_{j=0}^{N-1}(1 - r^{-1})^j \geq 1$, which is nondecreasing in $r \geq 1$, strictly so when $K \geq 1$. By (D.1) and (5), we obtain

$$\chi_K(B) \geq \Phi_{q,K}(B) = \sum_i s_i^2 \Psi_{q,K}(s_i^{-2}) \geq \Psi_{q,K}(\mathrm{sr}(B)) \geq 1.$$

By $\chi_K(B) \geq 1$, $\zeta_K(B) \leq \Lambda^K$, and the definition of $\zeta_K(B)$, we obtain the stated bound on $\|Q_K(B)\|_{\mathrm{F}}/\chi_K(B)$. □

We next characterize how $\chi_K(B)$, $\Phi_{q,K}(B)$, and $\Psi_{q,K}(r)$ vary with the polynomial degree D and the number K of Newton-Schulz iterations, where $q = \kappa + 1 = (D + 1)/2$.

Lemma D.2. *Let $B \neq 0$ and $r \geq 1$. Let $D = 2\kappa + 1 \in \{3, 5, \ldots\}$ denote the degree of each Newton-Schulz update, and set $q = (D + 1)/2$. Then the following properties hold:*

(i) For each D, $\chi_K(B)$ is nondecreasing in K. If $\mathrm{sr}(B) > 1$, both $\chi_K(B)$ and $\Phi_{q,K}(B)$ increase strictly with K. If $r > 1$, $\Psi_{q,K}(r)$ also increases strictly with K.

(ii) For each $K \geq 1$, $\chi_K(B)$ and $\Phi_{q,K}(B)$ increase strictly with D when $\mathrm{sr}(B) > 1$, and $\Psi_{q,K}(r)$ increases strictly with D when $r > 1$.

(iii) For each D and $K \geq 0$, we have

$$0 \leq \frac{\|B\|_*}{\|B\|_{\mathrm{F}}} - \chi_K(B) \leq \frac{\|B\|_*}{\|B\|_{\mathrm{F}}} - \Phi_{q,K}(B) \leq \sum_i s_i(B) e^{-q^K s_i^2(B)}.$$

Consequently, for each D, we have

$$\Phi_{q,K}(B) \longrightarrow \frac{\|B\|_*}{\|B\|_{\mathrm{F}}}, \qquad \Psi_{q,K}(r) \longrightarrow \sqrt{r} \qquad \text{as } K \longrightarrow \infty.$$

Proof. Write $s_i = s_i(B)$, set $T_\kappa(s) := sP_\kappa(s^2)$, and let $T_\kappa^{\circ K}$ denote its K-fold composition, with $T_\kappa^{\circ 0}(s) = s$. Since each Newton-Schulz step preserves singular vectors, we have $\chi_K(B) = \sum_i s_i T_\kappa^{\circ K}(s_i)$.

(i) Dependence on K. By Proposition D.1, we have $s \le T_\kappa(s) \le 1$ on $[0,1]$. Thus each scalar iterate is nondecreasing in K, and so is $\chi_K(B)$. For $0 < s < 1$, $P_\kappa(s^2)$ in (2) is a partial sum of the series for s^{-1}, whose terms are all positive. We therefore have $1 < P_\kappa(s^2) < s^{-1}$, and hence $s < T_\kappa(s) < 1$. If $\mathrm{sr}(B) > 1$, then $0 < s_i \le \mathrm{sr}(B)^{-1/2} < 1$, so every $T_\kappa^{\circ K}(s_i)$ remains in $(0,1)$ and increases strictly with K. Summing with the positive weights s_i, we obtain strict monotonicity of $\chi_K(B)$. For $0 < u < 1$, the function $N \mapsto \sqrt{1-u^N}$ is strictly increasing on $N > 0$. Since $q > 1$, the exponent q^K increases strictly with K. Applying this fact in (5) and the definition of $\Psi_{q,K}$, we obtain strict monotonicity of $\Phi_{q,K}(B)$ when $\mathrm{sr}(B) > 1$ and of $\Psi_{q,K}(r)$ when $r > 1$.

(ii) Dependence on D. By (2), we have

$$T_{\kappa+1}(s) - T_\kappa(s) = \frac{\binom{2\kappa+2}{\kappa+1}}{4^{\kappa+1}} s(1-s^2)^{\kappa+1} > 0, \qquad 0 < s < 1.$$

Differentiating (2) and using the expression for Λ_κ in Proposition D.1, we also have

$$T_\kappa'(s) = \Lambda_\kappa (1-s^2)^\kappa > 0, \qquad 0 < s < 1. \tag{D.2}$$

Since both maps preserve $(0,1)$, we combine their pointwise comparison with (D.2) and obtain by induction

$$T_{\kappa+1}^{\circ K}(s) > T_\kappa^{\circ K}(s), \qquad 0 < s < 1, \quad K \ge 1.$$

When $\mathrm{sr}(B) > 1$, summing over s_i with their positive weights, we obtain strict monotonicity of $\chi_K(B)$ in κ, equivalently in $D = 2\kappa + 1$. For each $K \ge 1$, $q^K = ((D+1)/2)^K$ increases strictly with D. Applying the scalar monotonicity used in (i) to (5) and the definition of $\Psi_{q,K}$, we obtain their strict increase in D under $\mathrm{sr}(B) > 1$ and $r > 1$, respectively.

(iii) Approximation error and limits. By (D.1) and Lemma D.1, we obtain

$$\begin{aligned} 0 \le \frac{\|B\|_*}{\|B\|_{\mathrm{F}}} - \chi_K(B) &\le \frac{\|B\|_*}{\|B\|_{\mathrm{F}}} - \Phi_{q,K}(B) \\ &\le \sum_i s_i (1-s_i^2)^{q^K} \le \sum_i s_i e^{-q^K s_i^2}, \end{aligned} \tag{D.3}$$

where the last two inequalities follow from $1 - \sqrt{1-z} \le z$ and $1 - z \le e^{-z}$ for $z \in [0,1]$, respectively. For each D, we have $q > 1$ and $s_i > 0$, so the upper bound (D.3) tends to zero as $K \to \infty$. We therefore obtain the stated limit of $\Phi_{q,K}(B)$. From the definition of $\Psi_{q,K}(r)$ and $0 \le 1 - r^{-1} < 1$, we also obtain $\Psi_{q,K}(r) \to \sqrt{r}$. □

D.2 Gram Matrices and Optimizer Updates

In this subsection, we establish bounds for the ReLU Gram matrices and the update directions of Muon and Adam.

Lemma D.3. *Suppose $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times d}$ satisfies $\|x_i\|_2 = 1$ for all $i \in [n]$. Then*

$$\mathbf{K}_a \succeq \frac{1}{4} XX^\top, \qquad \mathbf{K}_W \succeq \frac{1}{4} XX^\top.$$

In particular, if $XX^\top \succeq \kappa I_n$, then $\lambda_a, \lambda_W \ge \kappa/4$ and $\lambda_0 \ge \kappa/4$.

Proof. By symmetry of $u \sim \mathcal{N}(0, I_d)$, we have $\mathbb{E}[(Xu)|Xu|^\top] = 0$. By the definition of ReLU, we have $\sigma(Xu) = (Xu + |Xu|)/2$. Substituting this identity into (4), we obtain

$$\mathbf{K}_a = \frac{1}{4} XX^\top + \frac{1}{4} \mathbb{E}[|Xu||Xu|^\top] \succeq \frac{1}{4} XX^\top,$$

where the equality follows from $\mathbb{E}[uu^\top] = I_d$, and the inequality follows from $|Xu||Xu|^\top \succeq 0$.

To bound $\mathbf{K}_W$, we define $S_X(u) := \operatorname{diag}\big(\operatorname{sign}(u^\top x_i) : i \in [n]\big)$. By symmetry, we have $\mathbb{E}[S_X(u)] = 0$. Expanding (4), we obtain

$$\mathbf{K}_W = \frac{1}{4} XX^\top + \frac{1}{4}\mathbb{E}[S_X(u)XX^\top S_X(u)] \succeq \frac{1}{4} XX^\top,$$

where the inequality follows from $S_X(u)XX^\top S_X(u) \succeq 0$. If $XX^\top \succeq \kappa I_n$, we have $\mathbf{K}_a, \mathbf{K}_W \succeq (\kappa/4) I_n$. We therefore obtain $\lambda_a, \lambda_W \geq \kappa/4$ and $\lambda_0 = \min\{\lambda_a, \lambda_W\} \geq \kappa/4$. □

We next show that the two ReLU Gram matrices have the same nullspace.

Lemma D.4. *Suppose $\|x_i\|_2 = 1$ for every $i \in [n]$. The Gram matrices in* (4) *satisfy*

$$\ker(\mathbf{K}_a) = \ker(\mathbf{K}_W).$$

Consequently, $\lambda_a > 0$, $\lambda_W > 0$, and $\lambda_0 > 0$ are equivalent.

Proof. Fix $c \in \mathbb{R}^n$ and define $F_c(u) := \sum_{i=1}^n c_i \sigma(u^\top x_i)$. For $u \sim \mathcal{N}(0, I_d)$, we obtain

$$c^\top \mathbf{K}_a c = \mathbb{E}[F_c(u)^2], \qquad c^\top \mathbf{K}_W c = \mathbb{E}\,\|\nabla F_c(u)\|_2^2,$$

where $\nabla F_c(u) = \sum_i c_i x_i \mathbf{1}\{u^\top x_i > 0\}$ almost surely. Since both Gram matrices are positive semidefinite, $c \in \ker(\mathbf{K}_a)$ is equivalent to $c^\top \mathbf{K}_a c = 0$, and the same holds for $\mathbf{K}_W$.

If $c \in \ker(\mathbf{K}_a)$, then $F_c(u) = 0$ almost surely. Continuity of F_c and full support of the Gaussian distribution imply $F_c \equiv 0$. Thus $\nabla F_c(u) = 0$ almost surely, and we obtain $c \in \ker(\mathbf{K}_W)$. Conversely, if $c \in \ker(\mathbf{K}_W)$, then $\nabla F_c(u) = 0$ almost surely. On each activation region, F_c is linear with no constant term, so $F_c(u) = u^\top \nabla F_c(u) = 0$ almost surely. Hence $c \in \ker(\mathbf{K}_a)$. We conclude that the nullspaces coincide, which proves the equivalence of the three positivity conditions. □

Now, we prove Proposition 3.1, which states that pairwise nonparallel inputs guarantee $\lambda_0 > 0$.

Proof of Proposition 3.1. By Du et al. (2019, Theorem 3.1), pairwise nonparallel inputs satisfy $\mathbf{K}_W \succ 0$. Applying Lemma D.4, we obtain $\mathbf{K}_a \succ 0$ as well. Therefore, $\lambda_0 = \min\{\lambda_a, \lambda_W\} > 0$. □

Lemma D.5. *Suppose $M_t \neq 0$, and let $\overline{M}_t := M_t/(1-\beta^{t+1})$. If*

$$\left\|\overline{M}_t - G_t\right\|_{\mathrm{F}} \leq \frac{1}{2(1+\sqrt{\mathcal{B}})} \|G_t\|_{\mathrm{F}},$$

then

$$\langle G_t, Q_K(M_t)\rangle \geq \frac{1}{2}\chi_K(M_t)\,\|G_t\|_{\mathrm{F}}.$$

Proof. It is easy to check that $Q_K(M_t) = Q_K(\overline{M}_t)$. By the definition of $\chi_K(M_t)$, we have

$$\begin{aligned}\left\langle G_t, Q_K(\overline{M}_t)\right\rangle &\geq \chi_K(M_t)\left\|\overline{M}_t\right\|_{\mathrm{F}} - \|Q_K(M_t)\|_{\mathrm{F}}\left\|G_t - \overline{M}_t\right\|_{\mathrm{F}} \\ &\geq \chi_K(M_t)\|G_t\|_{\mathrm{F}} - \{\chi_K(M_t) + \|Q_K(M_t)\|_{\mathrm{F}}\}\left\|G_t - \overline{M}_t\right\|_{\mathrm{F}}, \end{aligned} \tag{D.4}$$

where the first inequality follows from Frobenius duality, and the second follows from the triangle inequality. From (D.4), we obtain the claimed lower bound of $\left\langle G_t, Q_K(\overline{M}_t)\right\rangle$ whenever

$$\left\|\overline{M}_t - G_t\right\|_{\mathrm{F}} \leq \frac{\chi_K(M_t)}{2\{\chi_K(M_t) + \|Q_K(M_t)\|_{\mathrm{F}}\}} \|G_t\|_{\mathrm{F}}. \tag{D.5}$$

By Lemma D.1, we have $\|Q_K(M_t)\|_{\mathrm{F}}/\chi_K(M_t) \leq \sqrt{\mathcal{B}}$, so $\left\|\overline{M}_t - G_t\right\|_{\mathrm{F}} \leq \frac{1}{2(1+\sqrt{\mathcal{B}})}\|G_t\|_{\mathrm{F}}$ implies (D.5), which completes the proof. □

Lemma D.6. *Suppose $0 \leq \beta_1 < 1$, $0 < \beta_2 < 1$, and $\beta_1^2 < \beta_2$. Define $\kappa_\beta^2 := \frac{(1-\beta_1)^2}{(1-\beta_2)(1-\beta_1^2/\beta_2)}$. Then, for every $t \geq 0$ and $j \in [m]$, we have*

$$|(\widehat{m}_t)_j| \leq \kappa_\beta \sqrt{(\widehat{v}_t)_j}.$$

Proof. We fix $t \geq 0$ and $j \in [m]$. Then we have

$$\begin{aligned}|(m_t)_j|^2 &= (1-\beta_1)^2 \left(\sum_{k=0}^{t} \beta_1^k (g_{t-k})_j\right)^2 \\ &\leq (1-\beta_1)^2 \left(\sum_{k=0}^{t} (\beta_1^2/\beta_2)^k\right)\left(\sum_{k=0}^{t} \beta_2^k (g_{t-k})_j^2\right) \\ &= \kappa_\beta^2 \{1-(\beta_1^2/\beta_2)^{t+1}\}(v_t)_j, \end{aligned} \tag{D.6}$$

where the inequality follows from the Cauchy–Schwarz inequality. The last equality follows from the identities

$$\sum_{k=0}^{t} \beta_2^k (g_{t-k})_j^2 = \frac{(v_t)_j}{1-\beta_2}, \qquad \sum_{k=0}^{t} (\beta_1^2/\beta_2)^k = \frac{1-(\beta_1^2/\beta_2)^{t+1}}{1-\beta_1^2/\beta_2},$$

together with the definition of κ_β. Applying the bias corrections to (D.6), we obtain

$$\begin{aligned}|(\widehat{m}_t)_j|^2 &\leq \kappa_\beta^2 \frac{(1-\beta_2^{t+1})\{1-(\beta_1^2/\beta_2)^{t+1}\}}{(1-\beta_1^{t+1})^2}(\widehat{v}_t)_j \\ &\leq \kappa_\beta^2 (\widehat{v}_t)_j, \end{aligned}$$

where the second inequality follows from $\beta_2^{t+1} + (\beta_1^2/\beta_2)^{t+1} \geq 2\beta_1^{t+1}$. Taking square roots, we obtain the desired bound. □

D.3 Rademacher Complexity and Generalization Bounds

In this subsection, we establish the complexity and population loss bounds used in the generalization proofs. Let $S = (x_1, \ldots, x_n)$ be an input sample with data matrix $X = (x_1, \ldots, x_n)^\top$. We define the empirical Rademacher complexity of a function class $\mathcal{H}$ on S and its worst case Rademacher complexity over samples satisfying $\|x_i\|_2 \leq 1$, respectively, by

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) := \frac{1}{n}\mathbb{E}_\epsilon \sup_{f\in\mathcal{H}} \left|\sum_{i=1}^{n} \epsilon_i f(x_i)\right|, \qquad \mathfrak{R}_n(\mathcal{H}) := \sup_{\|x_i\|_2 \leq 1} \widehat{\mathfrak{R}}_S(\mathcal{H}),$$

where the expectation is over independent Rademacher random variables $\epsilon_1, \ldots, \epsilon_n$.

We consider networks whose output and hidden weights remain close to their paired initialization. For $A, R, H \geq 0$, define

$$\mathcal{G}_{\mathrm{ctr}}(A, R, H) := \{f_\Theta : \|a - a_0\|_2 \leq A,\ \|W - W_0\|_2 \leq R,\ \|W - W_0\|_{\mathrm{F}} \leq H\}.$$

Lemma D.7. *Under the paired initialization in Assumption 3.2, every deterministic input sample S satisfies*

$$\widehat{\mathfrak{R}}_S\{\mathcal{G}_{\mathrm{ctr}}(A, R, H)\} \leq \frac{A\left\|W_0 X^\top\right\|_{\mathrm{F}}}{n\sqrt{m}} + \frac{2AR\|X\|_{\mathrm{F}}}{n} + \frac{2H\|X\|_{\mathrm{F}}}{n}.$$

Consequently,

$$\mathfrak{R}_n\{\mathcal{G}_{\mathrm{ctr}}(A, R, H)\} \leq \frac{A\|W_0\|_2}{\sqrt{mn}} + \frac{2(AR+H)}{\sqrt{n}}.$$

For every $f \in \mathcal{G}_{\mathrm{ctr}}(A, R, H)$, we have

$$|f(x)| \leq \left[\frac{A(\|W_0\|_2 + R)}{\sqrt{m}} + H\right]\|x\|_2.$$

Proof. Let $c := a - a_0$. Since the paired initialization satisfies $a_0^\top \sigma(W_0 x) = 0$, we have

$$f_\Theta(x) = \frac{1}{\sqrt{m}} c^\top \sigma(Wx) + \frac{1}{\sqrt{m}} a_0^\top \{\sigma(Wx) - \sigma(W_0 x)\}.$$

We bound the Rademacher contributions of these two terms separately, retaining the operator norm constraint for the first and the shared Frobenius norm constraint for the second.

First, we write $\Delta W := W - W_0$. For the contribution involving c, we have

$$\begin{aligned} I &:= \frac{1}{n\sqrt{m}}\mathbb{E}_\epsilon \sup_{\|\Delta W\|_2 \le R} \sup_{\|c\|_2 \le A} \left| c^\top \sum_i \epsilon_i \sigma((W_0 + \Delta W)x_i) \right| \\ &= \frac{A}{n\sqrt{m}}\mathbb{E}_\epsilon \sup_{\|\Delta W\|_2 \le R} \left\| \sum_i \epsilon_i \sigma((W_0 + \Delta W)x_i) \right\|_2 \\ &\le \frac{A}{n\sqrt{m}}\mathbb{E}_\epsilon \left\| \sum_i \epsilon_i \sigma(W_0 x_i) \right\|_2 \\ &\quad + \frac{A}{n\sqrt{m}}\mathbb{E}_\epsilon \sup_{\|\Delta W\|_2 \le R} \left\| \sum_i \epsilon_i \{\sigma((W_0 + \Delta W)x_i) - \sigma(W_0 x_i)\} \right\|_2 \\ &=: I_0 + I_\Delta, \end{aligned}$$

where the equality follows from $\sup_{\|c\|_2 \le A} |c^\top z| = A \|z\|_2$, the inequality follows from the triangle inequality, and I_0 and I_Δ denote the two terms in order.

For the first term I_0, we obtain

$$\begin{aligned} I_0 &\le \frac{A}{n\sqrt{m}} \left(\mathbb{E}_\epsilon \left\| \sum_i \epsilon_i \sigma(W_0 x_i) \right\|_2^2 \right)^{1/2} \\ &= \frac{A}{n\sqrt{m}} \left(\sum_i \|\sigma(W_0 x_i)\|_2^2 \right)^{1/2} \le \frac{A \left\| W_0 X^\top \right\|_{\mathrm{F}}}{n\sqrt{m}}, \end{aligned} \tag{D.7}$$

where the first inequality follows from Jensen's inequality, the equality follows from independence and the zero means of the Rademacher variables, and the last inequality follows from $\|\sigma(W_0 x_i)\|_2 \le \|W_0 x_i\|_2$.

Let $\Delta w_j^\top$ denote row j of ΔW, and define

$$\begin{aligned} \varphi_{ji}(z) &:= \sigma(w_{0,j}^\top x_i + z) - \sigma(w_{0,j}^\top x_i), \\ S_j(u) &:= \sum_i \epsilon_i \varphi_{ji}(u^\top x_i), \qquad j \in [m]. \end{aligned}$$

The vector inside I_Δ has coordinates $S_j(\Delta w_j)$, and $\|\Delta w_j\|_2 \le \|\Delta W\|_2 \le R$. Each φ_{ji} is Lipschitz with constant one and vanishes at zero. We therefore have

$$\mathbb{E}_\epsilon \sup_{\|u\|_2 \le R} |S_j(u)|^2 \le 4\mathbb{E}_\epsilon \sup_{\|u\|_2 \le R} \left| u^\top \sum_i \epsilon_i x_i \right|^2 = 4R^2 \mathbb{E}_\epsilon \left\| \sum_i \epsilon_i x_i \right\|_2^2 = 4R^2 \|X\|_{\mathrm{F}}^2, \tag{D.8}$$

where the inequality follows from Theorem 4.12 in Ledoux & Talagrand (1991) applied with the convex function $z \mapsto z^2$ on $[0, \infty)$, the first equality follows from $\sup_{\|u\|_2 \le R} |u^\top z|^2 = R^2 \|z\|_2^2$, and the last equality follows from $\mathbb{E}_\epsilon \|\sum_i \epsilon_i x_i\|_2^2 = \sum_i \|x_i\|_2^2 = \|X\|_{\mathrm{F}}^2$.

Summing (D.8) over $j \in [m]$, we obtain

$$\begin{aligned} I_\Delta &= \frac{A}{n\sqrt{m}}\mathbb{E}_\epsilon \sup_{\|\Delta W\|_2 \le R} \left(\sum_j S_j(\Delta w_j)^2 \right)^{1/2} \\ &\le \frac{A}{n\sqrt{m}} \left(\sum_j \mathbb{E}_\epsilon \sup_{\|u\|_2 \le R} |S_j(u)|^2 \right)^{1/2} \le \frac{A}{n\sqrt{m}} \sqrt{4mR^2 \|X\|_{\mathrm{F}}^2} = \frac{2AR\|X\|_{\mathrm{F}}}{n}, \end{aligned}$$

where the first equality follows from $\|z\|_2 = (\sum_{j=1}^m z_j^2)^{1/2}$ for $z \in \mathbb{R}^m$, the first inequality follows from $\|\Delta w_j\|_2 \le R$ and Jensen's inequality, and the second inequality follows from (D.8). Combining

this estimate of I_Δ with the bound on I_0 in (D.7), we obtain

$$I \le \frac{A\left\|W_0 X^\top\right\|_{\mathrm{F}}}{n\sqrt{m}} + \frac{2AR\left\|X\right\|_{\mathrm{F}}}{n}. \tag{D.9}$$

Next, we bound the contribution involving a_0. Let $v_j^\top$ denote row j of $W - W_0$, set $b_j := (a_0)_j/\sqrt{m}$, and write

$$Z(\epsilon) := \sup_{\sum_j \|v_j\|_2^2 \le H^2} \sum_{i,j} \epsilon_i b_j \varphi_{ji}(v_j^\top x_i).$$

Here $\sum_j b_j^2 = 1$. Since $v_j = 0$ is feasible for every j, both $Z(\epsilon)$ and $Z(-\epsilon)$ are nonnegative. Separating the contributions of c and a_0, we obtain

$$\begin{aligned}\widehat{\mathfrak{R}}_S\{\mathcal{G}_{\mathrm{ctr}}(A,R,H)\} &\le I + \frac{1}{n}\mathbb{E}_\epsilon \max\{Z(\epsilon), Z(-\epsilon)\} \\ &\le I + \frac{2}{n}\mathbb{E}_\epsilon Z(\epsilon),\end{aligned} \tag{D.10}$$

where the first inequality follows from the triangle inequality and relaxing the hidden weight constraints: the c term retains only $\|W - W_0\|_2 \le R$, whereas the a_0 term retains only $\|W - W_0\|_{\mathrm{F}} \le H$. The second inequality follows from $\max\{Z(\epsilon), Z(-\epsilon)\} \le Z(\epsilon) + Z(-\epsilon)$ and symmetry of the signs.

To bound $\mathbb{E}_\epsilon Z(\epsilon)$, we introduce an auxiliary parameter $u > 0$ to relax the constraint $\sum_j \|v_j\|_2^2 \le H^2$ with a quadratic penalty, and later minimize the resulting bound over u. For every $u > 0$, we have

$$\mathbb{E}_\epsilon Z(\epsilon) \le uH^2 + \sum_j \mathbb{E}_\epsilon \sup_{v\in\mathbb{R}^d}\left\{ b_j \sum_i \epsilon_i \varphi_{ji}(v^\top x_i) - u\|v\|_2^2 \right\}, \tag{D.11}$$

where the inequality follows from adding the nonnegative term $u(H^2 - \sum_j \|v_j\|_2^2)$ and dropping the norm constraint. To bound each expectation, we fix j and condition on the Rademacher variables $\{\epsilon_k : k \ne i\}$. Define $h(v) := b_j \sum_{k\ne i} \epsilon_k \varphi_{jk}(v^\top x_k) - u\|v\|_2^2$, $z_v := v^\top x_i$. Then we have

$$\begin{aligned}&\mathbb{E}_{\epsilon_i} \sup_v\{h(v) + \epsilon_i b_j \varphi_{ji}(z_v)\} \\ =&\frac{1}{2}\sup_v\{h(v) + b_j\varphi_{ji}(z_v)\} + \frac{1}{2}\sup_v\{h(v) - b_j\varphi_{ji}(z_v)\} \\ \le&\frac{1}{2}\sup_{v,v'}\{h(v) + h(v') + |b_j|\,|z_v - z_{v'}|\} \\ =&\mathbb{E}_{\epsilon_i}\sup_v\{h(v) + \epsilon_i b_j z_v\},\end{aligned} \tag{D.12}$$

where the first equality follows from $\mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = 1/2$, the inequality follows from $|\varphi_{ji}(z_v) - \varphi_{ji}(z_{v'})| \le |z_v - z_{v'}|$, and the last equality follows from interchanging v and v' in the supremum. Since (D.12) holds for any h independent of ϵ_i, we apply it to $\epsilon_1, \ldots, \epsilon_n$ and substitute into (D.11) to obtain

$$\mathbb{E}_\epsilon Z(\epsilon) \le uH^2 + \sum_j \mathbb{E}_\epsilon \sup_{v\in\mathbb{R}^d}\left\{ b_j v^\top \sum_i \epsilon_i x_i - u\|v\|_2^2 \right\}.$$

Completing the square in each supremum, we obtain

$$\mathbb{E}_\epsilon Z(\epsilon) \le uH^2 + \frac{1}{4u}\sum_j b_j^2 \mathbb{E}_\epsilon \left\| \sum_i \epsilon_i x_i \right\|_2^2 = uH^2 + \frac{\|X\|_{\mathrm{F}}^2}{4u},$$

where the equality follows from $\sum_j b_j^2 = 1$ and $\mathbb{E}_\epsilon \|\sum_i \epsilon_i x_i\|_2^2 = \sum_i \|x_i\|_2^2 = \|X\|_{\mathrm{F}}^2$. Taking the infimum over $u > 0$, we obtain

$$\mathbb{E}_\epsilon Z(\epsilon) \le \inf_{u>0}\left\{ uH^2 + \frac{\|X\|_{\mathrm{F}}^2}{4u} \right\} = H\left\|X\right\|_{\mathrm{F}}.$$

Substituting this estimate and (D.9) into (D.10), we obtain

$$\widehat{\mathfrak{R}}_S\{\mathcal{G}_{\mathrm{ctr}}(A,R,H)\} \le \frac{A\left\|W_0X^\top\right\|_{\mathrm{F}}}{n\sqrt{m}} + \frac{2AR\|X\|_{\mathrm{F}}}{n} + \frac{2H\|X\|_{\mathrm{F}}}{n}.$$

By $\left\|W_0X^\top\right\|_{\mathrm{F}} \le \|W_0\|_2\|X\|_{\mathrm{F}}$ and $\|X\|_{\mathrm{F}} \le \sqrt{n}$ for inputs in the unit ball, we obtain the uniform complexity bound.

Finally, by the decomposition of f_Θ, we have

$$\begin{aligned}|f_\Theta(x)| &\le \frac{1}{\sqrt{m}}\left[\|c\|_2\,\|\sigma(Wx)\|_2 + \|a_0\|_2\,\|\sigma(Wx)-\sigma(W_0x)\|_2\right]\\ &\le \left[\frac{A(\|W_0\|_2+R)}{\sqrt{m}} + H\right]\|x\|_2,\end{aligned}$$

where the first inequality follows from Cauchy–Schwarz, and the second follows from the Lipschitz property of ReLU, $\|c\|_2 \le A$, $\|a_0\|_2 = \sqrt{m}$, $\|W\|_2 \le \|W_0\|_2 + R$, and $\|W - W_0\|_2 \le \|W - W_0\|_{\mathrm{F}} \le H$. □

The following lemma specializes Theorem 1 of Srebro et al. (2010) to the square loss. This result builds on local Rademacher complexity (Bartlett et al., 2005).

Lemma D.8. *Let $S = \{(x_i, y_i)\}_{i=1}^n$ be an i.i.d. sample from $\mathcal{D}$, where $\|x\|_2 \le 1$ and $|y| \le C$ almost surely. Suppose $\mathcal{F}$ is a deterministic function class satisfying $\mathfrak{R}_n(\mathcal{F}) \le \Gamma$ and $|f(x)| \le U$ for every $f \in \mathcal{F}$ and $\|x\|_2 \le 1$. Given $\delta \in (0,1)$, let $V := C + U$ and define $\Xi := \log^{3/2}(en)\Gamma + V\sqrt{\frac{\log(2/\delta)}{n}}$. Then, with probability at least $1-\delta$ over S, we have*

$$\mathcal{L}_{\mathcal{D}}(f) \le \mathcal{L}(f) + C\sqrt{\mathcal{L}(f)}\,\Xi + C\Xi^2 \qquad \text{for all } f \in \mathcal{F},$$

where $\mathcal{L}(f)$ and $\mathcal{L}_{\mathcal{D}}(f)$ denote the empirical and population square losses, respectively.

Proof. The square loss $\varphi(z,y) := (z-y)^2/2$ satisfies $\partial_z^2\varphi = 1$ and $0 \le \varphi(z,y) \le V^2/2$ whenever $|z| \le U$ and $|y| \le C$. Applying Theorem 1 in Srebro et al. (2010) with smoothness constant 1, loss bound $V^2/2$, and $\mathfrak{R}_n(\mathcal{F}) \le \Gamma$, we obtain, with probability at least $1-\delta$,

$$\mathcal{L}_{\mathcal{D}}(f) \le \mathcal{L}(f) + C\sqrt{\mathcal{L}(f)}\,\Xi + C\left[\log^3(en)\Gamma^2 + \frac{V^2\log(2/\delta)}{n}\right]$$

for all $f \in \mathcal{F}$. Since $\log^3(en)\Gamma^2 + V^2\log(2/\delta)/n \le \Xi^2$, we obtain the stated bound. □