NONASYMPTOTIC μP THEORY FOR NEURAL NETWORKS WITH APPLICATIONS TO SMALL-TO-LARGE GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Maximal Update Parameterization (μP) provides a principled framework for scaling neural networks while preserving non-degenerate feature-learning dynamics in the width limit. However, existing theory for μP is largely asymptotic, relying on the infinite-width limit and leaving its behavior in *finite-width* regimes poorly understood. In this paper, we use the Tensor Program (TP) framework to provide a non-asymptotic characterization of gradient descent dynamics in canonical scalar-input, width-m shallow settings: two-layer linear and quadratic networks, and three-layer linear networks. Our analysis provides high-probability guarantees for the finite-width approximation over each fixed training horizon. As an application, we analyze *small-to-large generalization* for KL-regularized domain reweighting under two-layer square-loss dynamics, showing that the learned domain weights along the width-m training trajectory remain close to their infinite-width counterparts with discrepancy $\widetilde{O}(1/\sqrt{m})$. Finally, synthetic and real-world experiments provide qualitative evidence of cross-width domain-weight stability.

1 INTRODUCTION

Modern large-scale AI models have unlocked unprecedented capabilities in multiple domains, including natural language processing (NLP) (Brown et al., 2020), Computer Vision (CV) (Dosovitskiy et al., 2021) and scientific discovery (Jumper et al., 2021). For example, Large language models (LLMs) trained on web-scale text have recently made dramatic progress, serving not only as powerful NLP tools but as general-purpose task solvers (Achiam et al., 2023). However, the computational cost of training and tuning these large-scale models is prohibitive. To mitigate this, practitioners increasingly rely on small-to-large generalization (Khaddaj et al., 2025), using inexpensive "proxy" models to guide critical decisions such as dataset selection (Yang et al., 2024b), domain reweighting (Xie et al., 2023; Fan et al., 2024) and hyperparameter tuning (Wortsman et al., 2024; Yang et al., 2021).

Despite the widespread adoption of this strategy, its quantitative reliability at finite width remains less understood. Existing μP/TP theory provides an asymptotic explanation for stable feature learning and hyperparameter transfer across widths. We turn this asymptotic picture into quantitative finite-width guarantees, bounding the error of the infinite-width approximation along the training trajectory with high probability. Furthermore, it is unclear how to choose the size of the small model to enable reliable transfer. A leading framework for addressing this is Maximal Update Parameterization (μP) (Yang & Hu, 2021). Theoretically, μP scales hyperparameters to preserve training dynamics across widths, ensuring that as width increases, the network remains in the "rich" feature-learning regime (Yang et al., 2023; Woodworth et al., 2020) rather than collapsing into the "lazy" kernel regime (Jacot et al., 2018; Chizat et al., 2019). While μP provides a principled basis for analyzing transfer, the majority of the existing theory remains asymptotic in the infinite-width limit (Yang & Hu, 2021; Chen et al., 2025). Recent finite-width analyses include one-hidden-layer networks (Glasgow et al., 2025), two-layer mean-field models (Mahankali et al., 2023; Montanari & Urbani, 2025; Bordelon & Pehlevan, 2023; Bordelon et al., 2024a; 2025), and three-layer linear networks with discrete-time updates (Chizat et al., 2024). Therefore, it remains unclear how to characterize small-to-large generalization under the framework of μP for *finite-width multilayer neural networks*, leaving a critical gap in determining the appropriate proxy model width so that its observed behavior reliably predicts that of a target large-scale model.

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.

Solving this problem demands a *non-asymptotic analysis* of μP. A key technical hurdle in such an analysis is the emergence of statistical dependencies between neurons at finite width across iterations of the learning algorithm (e.g., SGD). Standard μP theory (Yang & Hu, 2021) uses Tensor Program limit laws to identify deterministic limits of empirical observables while accounting for structured dependencies. At finite width, these dependencies evolve across training iterations. Therefore, a non-asymptotic μP theory must explicitly control how initialization fluctuations propagate through the coupled model updates.

In this paper, we provide a non-asymptotic characterization of gradient descent dynamics under μP for a canonical class of shallow neural networks: linear neural networks and neural networks with quadratic activation (Du & Lee, 2018). The main difficulty of the finite-width analysis is controlling training-induced dependencies beyond the infinite-width μP limit laws (Yang & Hu, 2021; Yang & Littwin, 2023). In two-layer networks, we handle this by comparing the finite-width trajectory with a deterministic reference trajectory. The reference process remains i.i.d. across neurons: Gaussian in the linear case and bounded-degree Gaussian-polynomial in the quadratic case. The output error is then decomposed into an i.i.d. empirical fluctuation and a training-induced residual, and the residual is controlled by a coordinate-wise bound. For three-layer linear networks, the matrix layer couples all coordinates over time, so the reference-process argument is no longer sufficient. We instead unroll the dynamics and express every fixed-time observable as a function of the Gaussian initialization. Concentration around the finite-width expectation follows from weighted Gaussian Poincaré inequalities and Gaussian hypercontractivity (Beckner, 1989). Our contributions are summarized as follows.

- We provide a non-asymptotic high probability analysis of μP for scalar-input one-hidden-layer linear and quadratic networks, and two-hidden-layer linear networks. We prove that the deviation between the output of a finite-width network and its infinite-width limit is bounded by $\widetilde{O}\big(m^{-1/2}\big)$, where $\widetilde{O}(\cdot)$ hides horizon-dependent constants and polylogarithmic factors of m and $1/\delta$, and δ denotes the failure probability. We control discrete-time gradient descent trajectories over each fixed training horizon with explicit confidence dependence. A detailed comparison with prior work is given in Table 1 in Appendix A.
- We leverage our non-asymptotic μP theory to study small-to-large generalization (Khaddaj et al., 2025) in a KL-regularized domain reweighting problem. We show that, for two-layer neural networks under square-loss dynamics, the learned domain weights along a width-m trajectory converge to those along the corresponding infinite-width trajectory, with approximation error scaling as $\widetilde{O}\big(m^{-1/2}\big)$. This yields a finite-horizon width–accuracy tradeoff for adaptive domain weights in the analyzed two-layer setting.
- Synthetic and real-world experiments provide qualitative evidence of cross-width domain-weight stability in multidimensional, ReLU, and GPT-style settings beyond the theorem assumptions.

2 Related Work

Maximal Update Parameterization. Maximal update parameterization (μP) is a promising approach for maximizing feature learning while avoiding feature explosion as the neural network width is scaled up (Yang & Hu, 2021; Yang et al., 2021). Its derivation is rooted in the Tensor Program (TP) framework, which provides a systematic method for analyzing infinite-width limits across a broad range of neural network architectures (Yang, 2019; 2020; Yang & Littwin, 2023; Yang & Hu, 2021; Golikov & Yang, 2022). Within this framework, Yang & Hu (2021) identified two qualitatively different regimes for infinite-width multilayer perceptrons: a kernel regime, in which the training dynamics are largely determined by initialization, and a feature-learning regime, in which the training data plays a dominant role in shaping the learned representations. A related line of work further studies depth scaling and infinite-depth limits (Yang et al., 2024a; Hayou & Yang, 2023; Noci et al., 2023). These works primarily focus on asymptotic infinite-width or infinite-depth limits. Related finite-width analyses use ODEs or gradient flow (Vyas et al., 2023; Kunin et al., 2024) or study structured shallow linear models via reparameterization (Noci et al., 2024). Mei et al. (2019) give high-probability risk bounds comparing discrete SGD with a continuous-time mean-field limit, including a step-size error. Chizat et al. (2024) establish quantitative convergence for vector-input three-layer linear networks with discrete updates. Glasgow et al. (2025) obtain longer-time control for projected mean-field dynamics under structural and stability assumptions.

Under our stated assumptions, we obtain high-probability output-trajectory control, an $O(m^{-1})$ expectation-bias bound for the three-layer linear model, and adaptive domain-weight stability; see Table 1.

Hyperparameter Transfer and Small-to-Large Generalization. Building on the μP parameterization (Yang & Hu, 2021), Yang et al. (2021) showed that μP enables reliable hyperparameter transfer across width: hyperparameters tuned on small models remain near-optimal as width scales across diverse neural architectures. Subsequent work used dynamical mean-field theory (DMFT) to analyze hyperparameter transfer from small to large models (Bordelon & Pehlevan, 2022; Bordelon et al., 2024b; Bordelon & Pehlevan, 2025), but these guarantees are asymptotic. Beyond hyperparameter tuning, empirical studies demonstrate small-to-large generalization in broader settings, where small proxy models are used for data selection, data cleaning, and domain reweighting (Engstrom et al., 2024; Xie et al., 2023; Fan et al., 2024; Chen et al., 2023). More recently, Khaddaj et al. (2025) empirically studied small-to-large generalization via data influence, identifying regimes in which perturbations to the training data distribution reliably predict large-model behavior across compute scales. Our contribution is a finite-horizon, high-probability guarantee for adaptive domain-weight trajectories in the stated shallow settings.

Learning Dynamics for Neural Networks. The early works investigating the learning dynamics of neural networks were established under the infinite width limit, including the Neural Tangent Kernel (NTK) (Jacot et al., 2018; Chizat et al., 2019; Arora et al., 2019) and the mean-field approach (Chizat & Bach, 2018; Mei et al., 2018; 2019; Chizat et al., 2024). Recently, there is a line of work going beyond the infinite-width limit. In particular, they focus on feature learning process of optimization algorithms in neural networks under finite width (Woodworth et al., 2020; Allen-Zhu & Li, 2023; Li et al., 2020; Glasgow, 2024; Nichani et al., 2023; Fu et al., 2025; Xu & Liu, 2025; Suzuki et al., 2023; Ren et al., 2023; Berthier et al., 2025; Bordelon & Pehlevan, 2023; Montanari & Urbani, 2025). In this paper, our focus is analyzing the gap of the learning dynamics of gradient descent between the finite and infinite width limit under μP, which also fits into the feature learning regime.

3 Preliminaries and Notations

Computation Model. Given a general multilayer perceptron (MLP) with L hidden layers specified by weight matrices $W^1 \in \mathbb{R}^{m\times 1}$, $\{W^\ell\}_{\ell=2}^{L} \in \mathbb{R}^{m\times m}$, $W^{L+1} \in \mathbb{R}^{1\times m}$, and an activation function $\sigma(\cdot)$, the network computation is formally defined as

$$h^1 = W^1\xi, \quad x^\ell = \sigma(h^\ell), \quad h^{\ell+1} = W^{\ell+1}x^\ell, \quad g_\theta^{(m)}(\xi) = W^{L+1}x^L, \tag{1}$$

where $L \geq 1$ is any positive integer and $\ell \in \{1, \dots, L-1\}$, and θ denotes all parameters $[W^1, \dots, W^{L+1}]$, with one weight matrix for each hidden layer and one for the output layer.

As $m \to \infty$, the corresponding infinite-width feature-learning limit is defined as $g_\theta^{(m)}(\xi) \xrightarrow{\text{a.s.}} \bar{g}_\theta(\xi)$ (Yang & Hu, 2021, Theorem 6.4). The neural network is assumed to be trained using square loss $\ell(\cdot)$ via gradient descent. At iteration t, the training data is denoted as $(\xi_t, y(\xi_t))$. We assume $|\xi_t| \leq 1$ and $y(\xi_t)$ is uniformly bounded as in Yang & Hu (2021). The corresponding parameters in (1) at the t-th iteration are $W_t^1, \{h_t^\ell, W_t^\ell, x_t^\ell\}_{\ell=2}^{L}, W_t^{L+1}$. We run the optimization algorithm (i.e., gradient descent) for T iterations and $t \leq T$. Define the derivative of the loss for the width m model as $\chi^{(m)} := \ell'(g_\theta^{(m)}, y(\xi))$ and the corresponding infinite-width limit (Yang & Hu, 2021, Section 6.1) is defined as $\chi^{(m)} \xrightarrow[m\to\infty]{\text{a.s.}} \bar{\chi}$. Throughout the paper, we omit the superscript (m) and use χ to denote the derivative of the loss for ease of presentation. Unless stated otherwise, we use square loss $\ell(g, y) := \frac{1}{2}(g-y)^2$. The setup of the Maximal Update Parameterization (μP) (Yang & Hu, 2021) is specified by the following initialization, learning-rate, and notation conventions:

1. For layers $\ell = \{2, \dots, L\}$, the initial weight matrices $W_0^\ell \in \mathbb{R}^{m\times m}$ have i.i.d. entries sampled from a Gaussian distribution with variance $1/m$: $(W_0^\ell)_{\alpha\beta} \sim \mathcal{N}\big(0, \frac{1}{m}\big)$, where $\alpha, \beta = 1, \dots, m$.
2. The input-layer matrix $W^1 \in \mathbb{R}^{m\times 1}$ and the output-layer matrix $W^{L+1} \in \mathbb{R}^{1\times m}$ have i.i.d. entries at initialization $(W_0^1)_\alpha \sim \mathcal{N}(0, 1)$, $(W_0^{L+1})_\alpha \sim \mathcal{N}(0, \frac{1}{m^2})$, where $\alpha = 1, \dots, m$.
3. For notational convenience, we set the scalar base learning rate to $\eta = 1$. The layerwise μP learning-rate multipliers remain those specified below. Any fixed $O(1)$ base learning rate can be restored multiplicatively and absorbed into the constants.

4. Here, if h_t^ℓ appears without argument, it means $h_t^\ell(\xi_t)$, $\forall \ell \in [L]$. The same rule holds for x_t^ℓ, dh_t^ℓ, dx_t^ℓ, and χ_t, where t is the iteration counter.
5. Unless stated otherwise, the training horizon T and depth L are fixed independently of the width m. Time-dependent parameters in the infinite-width limit are therefore bounded over $0 \le t \le T$.
6. $O(\cdot)$ allows constants depending on T and other fixed parameters. $\widetilde{O}(\cdot)$ denotes high-probability bounds also hiding logarithmic factors in m and $1/\delta$, whose exponents may depend on the fixed horizon T.
7. For a real-valued random variable X and $p \ge 1$, we write $\|X\|_{L^p} := (\mathbb{E}|X|^p)^{1/p}$. More generally, if X takes values in a normed space $(E, \|\cdot\|_E)$, then $\|X\|_{L^p(E)} := (\mathbb{E}\|X\|_E^p)^{1/p}$.

Gradient descent on two-layer neural networks. Let $U_{t,\alpha} \in \mathbb{R}$ be the hidden weight and $a_{t,\alpha}$ the output weight of neuron α. Define the μP-scaled output weight $V_{t,\alpha} := m a_{t,\alpha}$, so $g_{\theta_t}^{(m)}(\xi) = m^{-1}\sum_{\alpha=1}^m V_{t,\alpha}\sigma(U_{t,\alpha}\xi)$. With μP learning-rate multipliers $\eta_a = m^{-1}$ and $\eta_U = m$, gradient descent gives $V_{t+1,\alpha} = V_{t,\alpha} - \chi_t \sigma(U_{t,\alpha}\xi_t)$ and $U_{t+1,\alpha} = U_{t,\alpha} - \chi_t V_{t,\alpha}\sigma'(U_{t,\alpha}\xi_t)\xi_t$. Thus, for $\sigma(z) = z$, $V_{t+1,\alpha} = V_{t,\alpha} - \chi_t \xi_t U_{t,\alpha}$ and $U_{t+1,\alpha} = U_{t,\alpha} - \chi_t \xi_t V_{t,\alpha}$. For $\sigma(z) = z^2$, $V_{t+1,\alpha} = V_{t,\alpha} - \chi_t \xi_t^2 U_{t,\alpha}^2$ and $U_{t+1,\alpha} = U_{t,\alpha} - 2\chi_t \xi_t^2 V_{t,\alpha} U_{t,\alpha}$. These are the two-block recursions used in Appendices C and D for the linear and pure-quadratic activation functions, respectively.

Gradient descent on Three-Layer Linear Neural Networks. For the scalar-input three-layer linear model, we track the first-layer parameter vector rather than an input-dependent activation. The evolving base variables are $u_t \in \mathbb{R}^m$, $W_t \in \mathbb{R}^{m\times m}$, $m\mathbf{v}_t \in \mathbb{R}^m$. Here $\mathbf{v}_t$ is the actual output-weight vector, while $m\mathbf{v}_t$ is its μP base variable. For an input ξ, the network output is $g_{\theta_t}^{(m)}(\xi) := \xi\, m^{-1}\langle m\mathbf{v}_t, W_t u_t\rangle$. At the training input ξ_t, define $z_t := W_t u_t$, $d_t := W_t^\top(m\mathbf{v}_t)$, $G_t^{(m)} := m^{-1}\langle m\mathbf{v}_t, z_t\rangle$. Thus $g_{\theta_t}^{(m)} := g_{\theta_t}^{(m)}(\xi_t) = \xi_t G_t^{(m)} = \xi_t m^{-1}\langle m\mathbf{v}_t, W_t u_t\rangle$. We write $\rho_t := \xi_t\chi_t$. The input-dependent first-layer activation at input ξ is ξu_t, but it is not treated as a separate evolving state variable. The μP learning-rate multipliers are $\eta_{\mathbf{v}} = m^{-1}$, $\eta_W = 1$, $\eta_u = m$. Gradient descent gives $m\mathbf{v}_{t+1} = m\mathbf{v}_t - \rho_t z_t$. The middle-layer update is $W_{t+1} = W_t - \rho_t m^{-1}(m\mathbf{v}_t)u_t^\top$. The first-layer parameter update is $u_{t+1} = u_t - \rho_t d_t$. Thus the three-layer linear μP recursion is $m\mathbf{v}_{t+1} = m\mathbf{v}_t - \rho_t z_t$, $u_{t+1} = u_t - \rho_t d_t$, $W_{t+1} = W_t - \rho_t m^{-1}(m\mathbf{v}_t)u_t^\top$, with $z_t = W_t u_t$, $d_t = W_t^\top(m\mathbf{v}_t)$, $g_{\theta_t}^{(m)}(\xi) = \xi\, m^{-1}\langle m\mathbf{v}_t, W_t u_t\rangle$. This is the recursion used in Appendix F.

4 Theoretical Results for Nonasymptotic μP

We first establish a high-probability upper bound on the finite-width error of square-loss μP dynamics. We then investigate how training stability affects its dependence on the training horizon: Theorem 4.3 shows that exponential amplification is unavoidable in the worst case allowed by these assumptions, while Theorem 4.4 establishes a time-uniform bound in a stable cross-entropy example.

Theorem 4.1 (Finite-time μP output concentration)**.** *Consider the scalar-input setup in Section 3, with square loss, bounded inputs and labels, fixed depth and learning-rate convention, Gaussian μP initialization with the specified variances, and a training sequence $(\xi_t, y(\xi_t))_{0\le t\le T}$ that is deterministic or independent of the initialization. For each t, write $g_{\theta_t}^{(m)} := g_{\theta_t}^{(m)}(\xi_t)$, $\bar g_{\theta_t} := \bar g_{\theta_t}(\xi_t)$, where $\bar g_{\theta_t}$ denotes the infinite-width output limit, deterministic conditional on the training sequence. For random training sequences, all probabilities and expectations below are conditional on that sequence.*

In the one-hidden-layer linear and pure-quadratic cases this deterministic limit is the reference-particle limit identified in Lemma D.4; in the two-hidden-layer linear case it is the word-closed limit identified with the Tensor Program limit in Lemma F.15. Then, in each of the following settings, there exist constants $C_T, a_T > 0$, depending only on the fixed data bound and on T, but not on m or δ, such that with probability at least $1-\delta$, $\max_{0\le t\le T}\left|g_{\theta_t}^{(m)} - \bar g_{\theta_t}\right| \le C_T m^{-1/2}\left(1 + \log\frac{2m(T+1)}{\delta}\right)^{a_T}$, where $a_T := 2\cdot 7^{T+2} + 1$. The settings are:

1. *one-hidden-layer linear networks, $L = 1$, $\sigma(z) = z$;*
2. *one-hidden-layer pure-quadratic networks, $L = 1$, $\sigma(z) = z^2$;*

3. two-hidden-layer linear networks, $L = 2$, $\sigma(z) = z$.

Remark 4.2. For each fixed training horizon T, the finite-width output stays uniformly close to the deterministic infinite-width limit over $0 \le t \le T$, with deviation $\widetilde{O}(m^{-1/2})$. Here, $\widetilde{O}(\cdot)$ hides the horizon-dependent prefactor C_T and polylogarithmic factors in m and $1/\delta$. The displayed a_T grows exponentially with T and is a sufficient exponent from the concentration analysis. This worst-case upper bound need not be attained along a particular training trajectory. The next two results address when exponential amplification is necessary and when the time dependence can be removed entirely. The proofs are given in Appendices C, D, and F. For two-layer linear networks, Appendix I gives a vector-input refinement with fixed logarithmic powers and explicit learning-rate dependence.

Worst-case necessity under square loss. The square-loss Tensor Program setting considered here permits unstable training dynamics. The following two-layer linear example shows that exponential amplification can occur at the leading $m^{-1/2}$ fluctuation scale.

Theorem 4.3 (Worst-case exponential dependence under square loss)**.** *Consider the two-layer linear μP dynamics in Section 3, with $g_t^{(m)}(\xi) = \xi q_t^{(m)}$, $q_t^{(m)} := m^{-1}\sum_{\alpha=1}^m V_{t,\alpha}U_{t,\alpha}$, and i.i.d. $U_{0,\alpha}, V_{0,\alpha} \sim N(0,1)$. Take $\xi_t = \epsilon_t/\sqrt{2}$ and $y_t = \beta\xi_t$, where $\beta = 2\sqrt{3}$ and the ϵ_t are independent Rademacher random variables, each taking values $+1$ and -1 with probability $1/2$. Then, for every fixed $T \ge 1$, $\sqrt{m}\big(g_T^{(m)}(\xi_T) - \bar{g}_T(\xi_T)\big) \xrightarrow[m\to\infty]{\mathrm{d}} N\big(0, \frac{21}{2}9^{T-1}\big)$, where $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution. Thus, any fixed-confidence $O(m^{-1/2})$ bound for this example requires an $\Omega(3^T)$ prefactor.*

Remark. The limiting fluctuations have scale $3^T/\sqrt{m}$, arising from the unstable linearized dynamics of this construction (Appendix H.1). Thus, exponential amplification is intrinsic to the worst case allowed by the square-loss assumptions. This lower bound concerns the fluctuation prefactor; the sufficient logarithmic exponent in Theorem 4.1 is a separate upper-bound quantity. For this learning-rate choice, the linearized update has an eigenvalue of -3.

A time-uniform cross-entropy example. Our GPT-style experiments use cross-entropy. Motivated by this loss, we next consider a two-layer linear model with label-smoothed cross-entropy. Under the specified data and learning rate, its concentration bound has no T-dependence.

Theorem 4.4 (Time-uniform bound under label-smoothed cross-entropy)**.** *Consider data $\xi_i = y_i v$, $y_i \in \{-1, +1\}$, $\|v\|_2 = 1$, and the two-layer linear μP network $g_t^{(m)}(\xi) = m^{-1}\sum_{\alpha=1}^m V_{t,\alpha}U_{t,\alpha}^\top\xi$, initialized independently with $V_{0,\alpha} \sim N(0,1)$, $U_{0,\alpha} \sim N(0, I_d)$. Use the loss $\ell(g(\xi), y) = -p\log\sigma(yg(\xi)) - (1-p)\log(1-\sigma(yg(\xi)))$, where $\sigma(q) = (1+e^{-q})^{-1}$, $p = 1-\varepsilon/2$, and $\varepsilon = 0.1$. Train both layers by full-batch gradient descent with base rate $\eta_\star = q_\star/[2(p-1/2)]$, where $q_\star = \log[p/(1-p)]$. There exist $C, C_0 < \infty$, independent of m, δ, T, such that for every $\delta \in (0,1)$, all $m \ge m_0(\delta) := \lceil C_0\log(4/\delta)\rceil$ satisfy $\sup_{t\ge 0}\max_i |g_t^{(m)}(\xi_i) - \bar{g}_t(\xi_i)| \le C\sqrt{\log(4/\delta)/m}$ with probability at least $1-\delta$.*

Remark. Both C and $m_0(\delta)$ are independent of T, and the same initialization event controls all iterations; see Appendix H.2. The chosen learning rate sends the limiting trajectory to a stable fixed point after one update. The guarantee uses this stability and the stated data structure; extending it to general cross-entropy training remains open.

5 Key Techniques and Proof Sketch

5.1 Key Challenges

In the infinite-width μP tensor-program analysis of Yang & Hu (2021), empirical averages over neurons converge to deterministic quantities, yielding deterministic output dynamics. At finite width, dependencies introduced by training must be controlled quantitatively. Although the parameters are initialized independently, gradient descent couples all coordinates through the empirical residual $\chi_t := g_{\theta_t}^{(m)} - y(\xi_t)$. Thus, every coordinate update depends on the same random scalar χ_t, and the resulting coordinates become dependent across neurons. Moreover, in deeper networks such as three-layer neural networks, the situation is more delicate because the evolving weight matrix W_t mixes coordinates, so bilinear quantities such as $\frac{1}{m}\langle m\mathbf{v}_t, W_t u_t\rangle$ cannot be treated as averages of independent random variables. In the scalar models studied here, the predictors have the form

$c_t^{(m)}\xi$ or $c_t^{(m)}\xi^2$. Width therefore changes parameter dynamics and fluctuations while keeping the predictor class fixed, isolating training stability from representational capacity. The scalar output coefficient does not determine its next update: the three-layer model also requires the evolving matrix and cross-time kernels.

5.2 Proof Sketch for Linear Networks

Condition on a data sequence that is either deterministic or independent of the Gaussian initialization; all constants may depend on T and the bounded data but not on m or δ. We write $L_\delta := 1 + \log(2/\delta)$ and $L_{m,\delta} := 1 + \log(2m/\delta)$. The notation $\widetilde{O}(m^{-1/2})$ hides fixed powers of these logarithmic factors.

Two-layer neural network. Using the notation from Section 3, we compare the finite-width trajectory with a Gaussian reference trajectory (Appendix C). Its update is a deterministic linear map of the initialized Gaussian pair, so the reference coordinates remain i.i.d. centered Gaussians at every fixed time. Writing $e^v_{t,\alpha}$ and $e^u_{t,\alpha}$ for the coordinate gaps, the output decomposes as

$$g^{(m)}_{\theta_t} - \bar{g}_{\theta_t} = \underbrace{\frac{1}{m}\sum_{\alpha=1}^{m}\big(V^{\mathrm{ref}}_{t,\alpha}X^{\mathrm{ref}}_{t,\alpha} - g^{\mathrm{ref}}_t\big)}_{\text{empirical Gaussian fluctuation}} + \underbrace{\frac{1}{m}\sum_{\alpha=1}^{m}\mathrm{Err}_{t,\alpha}}_{\text{training-induced residual}},$$

where $\bar{g}_{\theta_t} := g^{\mathrm{ref}}_t$ and $\mathrm{Err}_{t,\alpha}$ collects the mixed terms $V^{\mathrm{ref}}_{t,\alpha}e^x_{t,\alpha} + X^{\mathrm{ref}}_{t,\alpha}e^v_{t,\alpha} + e^v_{t,\alpha}e^x_{t,\alpha}$. Lemmas C.1 and C.5 establish this Gaussian reference and decomposition, while Lemma C.2 gives the Gaussian maximum bound and $m^{-1/2}$ concentration of the centered empirical product term. For the training-induced residual, the coordinatewise error recursion in Lemma C.4 and the bootstrap in Lemma C.6 give $\Delta_t \le C_t m^{-1/2}L_\delta + C_t L^{1/2}_{m,\delta}\varepsilon_t + \varepsilon_t^2$, $\varepsilon_{t+1} \le C_t\big((1+\Delta_t)\varepsilon_t + \Delta_t L^{1/2}_{m,\delta}\big)$, which jointly control empirical fluctuation and training-induced drift beyond the initialization law of large numbers. Starting from $\varepsilon_0 = 0$, Theorem C.7 in Appendix C closes this induction and proves the uniform fixed-time bound $|g^{(m)}_{\theta_t} - \bar{g}_{\theta_t}| = \widetilde{O}(m^{-1/2})$ for all $t \le T$.

Three-layer neural network. For the three-layer linear model, the evolving matrix W_t destroys coordinatewise independence, so the two-layer reference-particle argument no longer applies. Appendix F instead unrolls the dynamics as $W_{t+1} = W_0 - \sum_{r=0}^{t}\frac{\rho_r}{m}(m\mathbf{v}_r)u_r^\top$, $z_{t+1} = W_0u_{t+1} - \sum_{r=0}^{t}\rho_r(m\mathbf{v}_r)K^{xx}_{r,t+1}$, which shows that the later iterates are polynomial functions of the Gaussian initialization $\mathcal{G}_0 = (W_0, u_0, m\mathbf{v}_0)$ together with finitely many auxiliary kernels. In Appendix F, Lemma F.1 proves these unrolled identities, and Lemma F.6 proves that every scalar observable needed up to time T is a bounded-degree Gaussian polynomial. This reduces the proof to two generic estimates: $\mathrm{Var}(S) = O(m^{-1})$, $|S - \mathbb{E}S| = \widetilde{O}(m^{-1/2})$ for each observable S of interest. The variance bound follows from a weighted Gaussian Poincaré inequality for normalized inner products $S = m^{-1}\langle U, V\rangle$; Lemma F.7 establishes both estimates given the relevant vector and derivative fourth moments. Theorem F.8 propagates these vector, gradient, and scalar bounds simultaneously. Finally, Proposition F.14 proves that the finite-width expectations satisfy $\mathbb{E}S = S^\infty + O(m^{-1})$, which identifies the deterministic infinite-width limit. Applying these results to $S = g^{(m)}_{\theta_t}$ yields $|g^{(m)}_{\theta_t} - \bar{g}_{\theta_t}| = \widetilde{O}(m^{-1/2})$.

5.3 Proof Sketch for Networks with Quadratic Activation

For $\sigma(z) = z^2$, the reference particles are no longer Gaussian after one update, but remain i.i.d. polynomials of degree at most 2^t in their initial Gaussian pair. Lemma D.1 in Appendix D uses this structure to obtain polylogarithmic bounds on particle magnitudes and $\widetilde{O}(m^{-1/2})$ concentration of the reference empirical fluctuation $S^{\mathrm{ref}}_t := m^{-1}\sum_{\alpha=1}^{m}(Y^{\mathrm{ref}}_{t,\alpha} - g^{\mathrm{ref}}_t)$, where $Y^{\mathrm{ref}}_{t,\alpha} := V^{\mathrm{ref}}_{t,\alpha}\xi^2_t(U^{\mathrm{ref}}_{t,\alpha})^2$ and $g^{\mathrm{ref}}_t = \bar{g}_{\theta_t}$ is the deterministic infinite-width output at iteration t.

To control the finite/reference mismatch, let ε_t be the maximum coordinate gap and $M_t \ge 1$ bound the magnitudes of both sets of particles. Lemma D.2 gives, with $\Delta_t := |g^{(m)}_{\theta_t} - g^{\mathrm{ref}}_t|$,

$$\Delta_t \le |S^{\mathrm{ref}}_t| + C_tM_t^2\varepsilon_t, \qquad \varepsilon_{t+1} \le C_t(M_t\varepsilon_t + M_t^2\Delta_t), \qquad M_{t+1} \le C_tM_t^5.$$

The first inequality separates empirical fluctuation from the perturbation error; the remaining bounds propagate the particle gap and control nonlinear growth. Starting from $\varepsilon_0 = 0$, the induction keeps

Algorithm 1 μP GD training for domain reweighting

1: **Input:** Width m, learning rate η, KL parameter λ, iteration budget T
2: **Initialize:** $\theta_0^{(m)}$
3: **for** $t = 0, \dots, T-1$ **do**
4: Compute $\ell_t^{(m)} = \ell^{(m)}(\theta_t^{(m)})$ and set $\pi_t^{(m)} = \Phi_\lambda(\ell_t^{(m)})$
5: Update $\theta_{t+1}^{(m)} \leftarrow \theta_t^{(m)} - \eta P_m \sum_{i=1}^K \pi_{i,t}^{(m)} \nabla_\theta \ell_i^{(m)}(\theta_t^{(m)})$
6: **end for**
7: **Output:** $\pi_T^{(m)} = \Phi_\lambda(\ell^{(m)}(\theta_T^{(m)}))$

M_t polylogarithmic in m and $1/\delta$, while ε_t and Δ_t retain the $\widetilde{O}(m^{-1/2})$ scale for all $t \le T$, for each fixed horizon T. The complete induction is given in Theorem D.3.

6 Cross-Width Stability of Domain Reweighting

In this section, we show how to leverage nonasymptotic μP results to establish small-to-large generalization (Khaddaj et al., 2025) for domain reweighting under two-layer neural networks. Consider a multi-domain empirical dataset $\mathcal{D} = \bigcup_{i=1}^K \mathcal{D}_i$, where $\mathcal{D}_i = \{(\xi_{i,j}, y_{i,j})\}_{j=1}^{n_i}$ and $N := \sum_{i=1}^K n_i$. For each domain i, let $\widehat{\mathcal{P}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \delta_{(\xi_{i,j}, y_{i,j})}$ be the empirical distribution. Given a width-m model $g_\theta^{(m)}$, define the domain loss $\ell_i^{(m)}(\theta) := \mathbb{E}_{\widehat{\mathcal{P}}_i}[(g_\theta^{(m)}(\xi) - y)^2]$, and write $\ell^{(m)}(\theta) := (\ell_1^{(m)}(\theta), \dots, \ell_K^{(m)}(\theta))$. Motivated by Xie et al. (2023) and Fan et al. (2024), we aim to solve the following KL-regularized minimax formulation for joint network training and domain reweighting:

$$\min_\theta \max_{\pi^{(m)} \in \Delta^K} \left\{ \sum_{i=1}^K \pi_i^{(m)} \ell_i^{(m)}(\theta) - \frac{1}{\lambda} D_{\mathrm{KL}}(\pi^{(m)} \| \mathbf{u}) \right\}, \qquad \mathbf{u} = \left(\frac{1}{K}, \dots, \frac{1}{K} \right), \tag{2}$$

where $\lambda > 0$. For fixed θ, the inner maximizer is $\pi^{(m)}(\theta) = \Phi_\lambda(\ell^{(m)}(\theta))$, where $\Phi_\lambda(z)_i := \exp(\lambda z_i) / \sum_{j=1}^K \exp(\lambda z_j)$ for $z \in \mathbb{R}^K$. Substituting this optimizer into (2) gives the fixed log-sum-exp objective $\mathcal{R}_m(\theta) := \frac{1}{\lambda} \log \left(\frac{1}{K} \sum_{i=1}^K \exp(\lambda \ell_i^{(m)}(\theta)) \right)$, whose gradient satisfies $\nabla_\theta \mathcal{R}_m(\theta) = \sum_{i=1}^K \pi_i^{(m)}(\theta) \nabla_\theta \ell_i^{(m)}(\theta)$. Thus, although the weights $\pi_i^{(m)}(\theta_t^{(m)})$ evolve along training, the update is simply gradient descent on the fixed objective $\mathcal{R}_m$, with the layerwise μP scaling $P_m = \mathrm{diag}(mI_m, m^{-1}I_m)$ for hidden and actual output weights, respectively (Section 3).

We now study the cross-width stability of the adaptive domain weights. Specifically, we show that the KL-regularized best-response weights generated along a finite-width μP training trajectory remain close to those generated along the corresponding infinite-width trajectory throughout training. Let $\theta_t^{(m)}$ be the width-m trajectory generated by Algorithm 1, and let $\bar{g}_t$ be the corresponding infinite-width μP limit output. Define $\ell_{i,t}^{(\infty)} := \mathbb{E}_{\widehat{\mathcal{P}}_i}[(\bar{g}_t(\xi) - y)^2]$, $\ell_t^{(\infty)} := (\ell_{1,t}^{(\infty)}, \dots, \ell_{K,t}^{(\infty)})$, and $\pi_t^{(\infty)} := \Phi_\lambda(\ell_t^{(\infty)})$, defining the common deterministic domain-weight trajectory across widths.

Theorem 6.1 (Cross-width stability of domain weights)**.** *Run Algorithm 1 for T steps in the two-layer μP setting above, with bounded inputs and labels, square loss, and activation $\sigma(z) = z$ or z^2. Fix K, N, λ independently of m and assume $B_T := \max_{0 \le t \le T} \max_{i \in [K]} \sqrt{\ell_{i,t}^{(\infty)}} < \infty$. For every $\delta \in (0,1)$, there exist $C_T, a_T > 0$, depending only on the fixed problem parameters and T, not on m, δ, such that, writing $\varepsilon_{m,T}(\delta) := C_T m^{-1/2}(1 + \log(2mN(T+1)/\delta))^{a_T}$, with probability at least $1 - \delta$, $\max_{0 \le t \le T} \max_{i \in [K]} \max_{1 \le j \le n_i} |g_{\theta_t}^{(m)}(\xi_{i,j}) - \bar{g}_t(\xi_{i,j})| \le \varepsilon_{m,T}(\delta)$ and $\max_{0 \le t \le T} \|\pi_t^{(m)} - \pi_t^{(\infty)}\|_\infty \le 2\lambda(B_T + \varepsilon_{m,T}(\delta))\varepsilon_{m,T}(\delta) = \widetilde{O}(m^{-1/2})$.*

Remark 6.2 (Connection to small-to-large generalization). Khaddaj et al. (2025) empirically demonstrate that training data often influences model predictions similarly across scales, enabling small proxies to support data attribution and selection. Theorem 6.1 provides an explicit finite-width guarantee for adaptive domain reweighting in the stated two-layer μP setting. For proxy and target widths $m_s \le m_l$ sharing the same data and training setup, apply the theorem to each width with failure probability $\delta/2$. A union bound and the triangle inequality through their common limit $\pi_t^{(\infty)}$

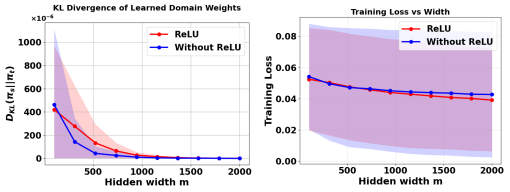


Figure 1: KL divergence of learned domain weights and training loss between small and large networks.

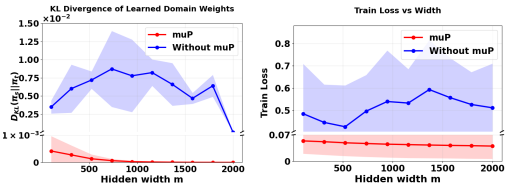


Figure 2: KL divergence and training loss: without μP vs. with μP.

give $\max_{0\le t\le T}\|\pi_t^{(m_s)}-\pi_t^{(m_l)}\|_\infty=\widetilde{O}(m_s^{-1/2}+m_l^{-1/2})$ with probability at least $1-\delta$. This gives a finite-horizon width–accuracy tradeoff for adaptive weights; see Appendix G. It neither asserts global minimax optimality nor analyzes a fresh target trained with frozen proxy weights.

7 Experiments

7.1 Synthetic Experiment

We design a synthetic multi-domain regression experiment to test whether domain weights learned by small models are close to those learned by large models, exploring width stability with multidimensional inputs beyond the scalar-input theory in Section 4. We generate $K=10$ domains, where each input $x\in\mathbb{R}^{500}$ follows $\mathcal{N}(\mu_i,\Sigma_i)$ with domain-specific mean and covariance. Scalar labels are produced by a two-layer generator with Kaiming initialization (He et al., 2015). To simulate data imbalance, domain sample counts follow a long-tailed distribution (Reed, 2001; Liu et al., 2019), with a total of 2000 samples distributed across the ten synthetic domains.

For each student width m, we jointly train the network and domain weights π using Algorithm 1 to solve the minimax problem (2). Hyperparameter searches, selected values, and the full width grid are given in Appendix J.1, along with the common synthetic training settings.

Linear Neural Networks. Both student and teacher models are two-layer linear networks, initialized and parameterized under μP as in Section 3. Their weight matrices are $W_1\in\mathbb{R}^{m\times d}$ and $W_2\in\mathbb{R}^{1\times m}$, with input dimension $d=500$ and scalar output. We vary student widths from 100 to 2000 and use the largest model as the teacher. All models share the selected learning rate and regularization parameter and are trained for $T=500$ steps. After training, we compute the KL divergence between the learned student and teacher domain-weight distributions, π_{s} and π_{tea}. The blue curves in Figure 1 show KL divergence and training loss, averaged over three seeds $(0,1,42)$, with shaded bands indicating standard deviation. For sufficiently wide students ($m\ge 500$), the learned weights closely match those of the teacher, with $D_{\mathrm{KL}}<2\times10^{-8}$.

Nonlinear Neural Networks. Following the same pipeline, we add a ReLU activation after the first linear layer and retain the same depth, width sweep, and training budget. We tune the learning rate over the same range as in the linear experiment and use a common value across widths. The red curves in Figure 1 show nearly identical KL divergence to the linear networks and moderately lower training loss, consistent with the greater representational capacity of the nonlinear model. Thus, domain weights learned by sufficiently wide students closely match those of the teacher in both settings. This cross-width agreement resembles the small-to-large behavior observed by Khaddaj et al. (2025).

Ablation Studies: without μP vs. with μP. We repeat the ReLU experiment under standard parameterization, using Kaiming initialization and an unscaled linear output. Data generation, architectures, widths, regularization, training budget, and random seeds remain unchanged. We tune the learning rate separately for each width; the search range and selected rates appear in Appendix J.1. We compute the student–teacher KL divergence after training, following the same evaluation procedure as in the μP experiments. The blue curves in Figure 2 show less stable domain weights across widths and higher training loss under the reported tuning and training budget. The loss gap also reflects different optimization outcomes. Replacing ReLU with a quadratic activation gives the additional comparison in Figure 5 in Appendix K.

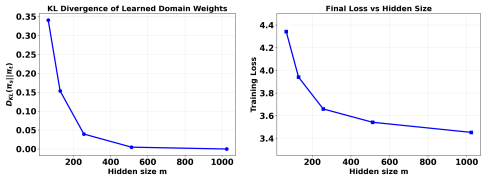


Figure 3: Comparison between small and large GPT models. Left: KL divergence of learned domain weights. Right: Training loss.

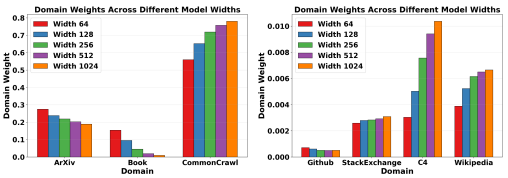


Figure 4: Domain weights learned by GPT model with increasing width.

Domain-weight Discrepancy vs Training Horizon. We study domain-weight stability across training budgets using two-layer ReLU networks under μP. We fix student and teacher widths to 100 and 2000, respectively, and vary T from 100 to 5000, tuning the student learning rate for each budget. Figure 6 in Appendix K shows that the KL divergence, averaged over three seeds, increases with T. This comparison measures the joint effect of the training budget and its selected learning rate on cross-width stability.

7.2 Domain Reweighting with SlimPajama-6B Dataset

We further examine domain-weight stability across model scales using GPT-style models on SlimPajama-6B (Soboleva et al., 2023), beyond the shallow square-loss setting of our theory. The dataset contains seven domains: ArXiv, Books, CommonCrawl, Github, StackExchange, C4, and Wikipedia. Student and teacher models have 4 layers and 8 attention heads and use μP. Student widths are $m \in \{64, 128, 256, 512, 1024\}$, with the width-1024 model serving as the teacher.

We perform exponentiated-gradient ascent on the domain weights, using $\hat{\pi}_t = \pi_{t-1} \odot \exp\left(\eta_\pi \nabla_\pi \mathcal{L}(\theta, \pi_{t-1})\right), \pi_t = \hat{\pi}_t / \sum_{i=1}^{K} \hat{\pi}_{i,t}$. where $\mathcal{L}$ is the KL-regularized weighted cross-entropy loss. The update form is shared with DoGE (Fan et al., 2024), but its bilevel gradient-alignment signal differs from our $\nabla_\pi \mathcal{L}$. This update is a practical variant of Algorithm 1, since computing Φ_λ requires expensive evaluations over all domains' data. The weights evolve iteratively rather than by exact softmax best responses. We train each model with MuAdamW (Yang et al., 2021) for $T = 5000$ steps, using maximum sequence length 1024 and batch size 4 per domain (28 in total). Appendix J.2 provides the dataset source, hyperparameter searches and selected values, and computational resources used to train the GPT-style student and teacher models.

After training, we compare the learned domain-weight distributions and training losses. Figure 3 shows that students with $m \geq 256$ closely match the teacher's domain weights, with $D_{\mathrm{KL}} < 0.05$, and achieve comparable training loss below 3.80.

Figure 4 further visualizes the learned weights across model sizes. Smaller students closely align with the teacher, consistently assigning the largest weight to CommonCrawl, followed by ArXiv and Books. These observations suggest that similar domain-weight stability can persist in GPT-style models across the tested hidden dimensions under the same experimental setup.

8 Conclusion

We develop a non-asymptotic, high-probability theory for gradient descent under μP in width-m two and three-layer linear networks and two-layer nonlinear neural networks with quadratic activation, establishing $\widetilde{O}(m^{-1/2})$ convergence of the finite-width dynamics to their infinite-width limit over finite training horizons. As an application, we establish finite-horizon cross-width stability of adaptive domain weights. An important direction for future work is to extend this framework to deeper architectures and depth-dependent μP parameterizations.

Limitations. The main concentration theorem assumes scalar inputs, bounded data, and square-loss μP dynamics; Appendix I extends the two-layer linear case to vector inputs. Theorem 4.3 shows worst-case exponential fluctuation amplification, while Theorem 4.4 gives a time-uniform bound in a stable label-smoothed cross-entropy example. Extending this stability analysis to practical training remains open. Our experiments provide qualitative evidence beyond the proved settings.

AI USE STATEMENT

The research ideas originated with the authors. GPT assisted with proof development and checking, as well as manuscript preparation. The authors reviewed all AI-assisted content and take full responsibility for the correctness of the proofs and the final manuscript.

ETHICS STATEMENT

Experiments use synthetic data and SlimPajama, with no human participants. Web text may contain biased, harmful, or sensitive content; our results provide no guarantees of downstream fairness, privacy, or safety.

REPRODUCIBILITY STATEMENT

Proofs and auxiliary lemmas are provided in Appendices B–I. Experimental settings, data sources, hyperparameters, and computational resources are documented in Appendix J, with additional results in Appendix K.

REFERENCES

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

Zeyuan Allen-Zhu and Yuanzhi Li. Backward feature correction: How deep learning performs deep (hierarchical) learning. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 4598–4598. PMLR, 2023.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in neural information processing systems*, 32, 2019.

William Beckner. A generalized poincaré inequality for gaussian measures. *Proceedings of the American Mathematical Society*, 105(2):397–400, 1989. doi: 10.1090/S0002-9939-1989-0954373-7.

Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *Foundations of Computational Mathematics*, 25(5):1627–1710, August 2025. ISSN 1615-3383. doi: 10.1007/s10208-024-09664-9. URL https://doi.org/10.1007/s10208-024-09664-9. Published online 22 August 2024.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Advances in Neural Information Processing Systems*, 35:32240–32256, 2022.

Blake Bordelon and Cengiz Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. *Advances in Neural Information Processing Systems*, 36:9707–9750, 2023.

Blake Bordelon and Cengiz Pehlevan. Deep linear network training dynamics from random initialization: Data, width, depth, and hyperparameter transfer. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4968–4997. PMLR, 2025. URL https://proceedings.mlr.press/v267/bordelon25a.html.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4345–4382. PMLR, 2024a. URL https://proceedings.mlr.press/v235/bordelon24a.html.

Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=KZJehvRKGD.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8):084002, 2025. doi: 10.1088/1742-5468/adefb1. URL https://doi.org/10.1088/1742-5468/adefb1.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Mayee Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. Skill-it! a data-driven skills framework for understanding and training language models. *Advances in Neural Information Processing Systems*, 36:36000–36040, 2023.

Zixiang Chen, Greg Yang, Qingyue Zhao, and Quanquan Gu. Global convergence and rich feature learning in l-layer infinite-width neural networks under μ parametrization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9614–9641. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25cd.html.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Lénaïc Chizat, Maria Colombo, Xavier Fernández-Real, and Alessio Figalli. Infinite-width limit of deep linear neural networks. *Communications on Pure and Applied Mathematics*, 77(10): 3958–4007, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Simon S. Du and Jason D. Lee. On the power of over-parametrization in neural networks with quadratic activation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1329–1338. PMLR, 2018. URL https://proceedings.mlr.press/v80/du18a.html.

Logan Engstrom, Axel Feldmann, and Aleksander Madry. Dsdm: Model-aware dataset selection with datamodels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12491–12526. PMLR, 2024. URL https://proceedings.mlr.press/v235/engstrom24a.html.

Simin Fan, Matteo Pagliardini, and Martin Jaggi. DoGE: Domain reweighting with generalization estimation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12895–12915. PMLR, 2024. URL https://proceedings.mlr.press/v235/fan24e.html.

Hengyu Fu, Zihao Wang, Eshaan Nichani, and Jason D Lee. Learning hierarchical polynomials of multiple nonlinear features with three-layer networks. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UZ893n8FXr.

Margalit Glasgow. Sgd finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the xor problem. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=HgOJlxzB16.

Margalit Glasgow, Denny Wu, and Joan Bruna. Propagation of chaos in one-hidden-layer neural networks beyond logarithmic time. *arXiv preprint arXiv:2504.13110*, 2025.

Eugene Golikov and Greg Yang. Non-gaussian tensor programs. *Advances in Neural Information Processing Systems*, 35:21521–21533, 2022.

Soufiane Hayou and Greg Yang. Width and depth limits commute in residual networks. In *International Conference on Machine Learning*, pp. 12700–12723. PMLR, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Alaa Khaddaj, Logan Engstrom, and Aleksander Madry. Small-to-large generalization: Training data influences models consistently across scale. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=79ZkWgY2FI.

Daniel Kunin, Allan Raventós, Clémentine Dominé, Feng Chen, David Klindt, Andrew Saxe, and Surya Ganguli. Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning. *Advances in Neural Information Processing Systems*, 37, 2024. URL https://openreview.net/forum?id=eNM94i7R3A.

Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In *Conference on learning theory*, pp. 2613–2682. PMLR, 2020.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2537–2546, 2019.

Arvind Mahankali, Haochen Zhang, Kefan Dong, Margalit Glasgow, and Tengyu Ma. Beyond ntk with vanilla gradient descent: A mean-field analysis of neural networks with polynomial width, samples, and time. *Advances in Neural Information Processing Systems*, 36:57367–57480, 2023.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on learning theory*, pp. 2388–2464. PMLR, 2019.

Andrea Montanari and Pierfrancesco Urbani. Dynamical decoupling of generalization and overfitting in large two-layer networks. *Advances in Neural Information Processing Systems*, 38, 2025. URL https://openreview.net/forum?id=ImpizBSKcu.

Eshaan Nichani, Alex Damian, and Jason D Lee. Provable guarantees for nonlinear feature learning in three-layer neural networks. *Advances in Neural Information Processing Systems*, 36:10828–10875, 2023.

Lorenzo Noci, Chuning Li, Mufan Li, Bobby He, Thomas Hofmann, Chris J Maddison, and Dan Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. *Advances in Neural Information Processing Systems*, 36:54250–54281, 2023.

Lorenzo Noci, Alexandru Meterez, Thomas Hofmann, and Antonio Orvieto. Super consistency of neural network landscapes and learning rate transfer. *Advances in Neural Information Processing Systems*, 37, 2024. URL https://openreview.net/forum?id=rgwhJ7INtZ.

William J Reed. The pareto, zipf and other power laws. *Economics letters*, 74(1):15–19, 2001.

Yunwei Ren, Mo Zhou, and Rong Ge. Depth separation with multilayer mean-field networks. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=uzFQpkEzOo.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama, June 2023. URL https://huggingface.co/datasets/cerebras/SlimPajama-627B.

Taiji Suzuki, Denny Wu, Kazusato Oko, and Atsushi Nitanda. Feature learning via mean-field langevin dynamics: classifying sparse parities and beyond. *Advances in Neural Information Processing Systems*, 36:34536–34556, 2023.

Nikhil Vyas, Alexander Atanasov, Blake Bordelon, Depen Morwani, Sabarish Sainathan, and Cengiz Pehlevan. Feature-learning networks are consistent across widths at realistic scales. *Advances in Neural Information Processing Systems*, 36, 2023. URL https://openreview.net/forum?id=LTdfYIvbHc.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Mitchell Wortsman, Peter J Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, et al. Small-scale proxies for large-scale transformer training instabilities. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=d8w0pmvXbZ.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36:69798–69818, 2023.

Yizhou Xu and Ziyin Liu. Three mechanisms of feature learning in a linear network. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9941833e8327910ef25daeb9005e4748-Abstract-Conference.html.

Greg Yang. Wide feedforward or recurrent neural networks of any architecture are gaussian processes. *Advances in Neural Information Processing Systems*, 32, 2019.

Greg Yang. Tensor programs ii: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*, 2020.

Greg Yang and Edward J Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737. PMLR, 2021. URL https://proceedings.mlr.press/v139/yang21c.html.

Greg Yang and Etai Littwin. Tensor programs ivb: Adaptive optimization in the infinite-width limit. *arXiv preprint arXiv:2308.01814*, 2023.

Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *Advances in Neural Information Processing Systems*, 34: 17084–17097, 2021.

Greg Yang, James B Simon, and Jeremy Bernstein. A spectral condition for feature learning. *arXiv preprint arXiv:2310.17813*, 2023.

Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs vi: Feature learning in infinite-depth neural networks. In *International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=17pVDnpwwl.

Yu Yang, Siddhartha Mishra, Jeffrey Chiang, and Baharan Mirzasoleiman. SmallToLarge (S2L): Scalable data selection for fine-tuning large language models by summarizing training trajectories of small models. *Advances in Neural Information Processing Systems*, 37:83465–83496, 2024b.

A Comparison with Prior Work

Table 1: Key differences from related width-limit analyses.

Work and model	Main guarantee	Distinction of our results
Yang & Hu (2021) Nonlinear MLPs	Almost-sure infinite-width μP limit.	Explicit finite-width error and confidence dependence.
Mei et al. (2019) 2-layer nonlinear; vector inputs	High-probability risk error between SGD and a mean-field PDE; width and step-size terms.	Output control relative to the discrete-time width limit, with unbounded linear or quadratic features.
Bordelon & Pehlevan (2023) Nonlinear MLPs	Perturbative $m^{-1/2}$ fluctuations and m^{-1} mean correction.	Nonasymptotic high-probability bounds for the stated shallow models.
Chizat et al. (2024) 3-layer linear; vector inputs	Discrete GD/SGD; mean-square error $O(m^{-1+\varepsilon})$ for any fixed $\varepsilon > 0$.	$O(m^{-1})$ mean-square error and expectation bias, with logarithmic confidence dependence, under our Gaussian, scalar-input, square-loss assumptions.
Glasgow et al. (2025) 1-hidden-layer; structured targets	Beyond-logarithmic-time mean-field approximation under structural and stability assumptions.	Exact discrete updates in the stated unprojected models; complementary fixed-horizon guarantees.
This work Theorems 4.1 and 6.1 2–3 layer linear; 2-layer quadratic	High-probability $\widetilde{O}(m^{-1/2})$ output error and adaptive domain-weight error.	Finite-horizon trajectory guarantees under discrete GD, with adaptive domain-weight stability across widths.

Scope. The displayed main theorems assume scalar inputs and Gaussian μP initialization; the domain-weight guarantee is for two-layer models, and the $O(m^{-1})$ mean-square and expectation-bias bounds (Corollary F.18) are for the three-layer linear model. T is fixed independently of m; $\widetilde{O}$ hides horizon-dependent constants and logarithmic factors in m and $1/\delta$. Proposition I.1 separately treats vector-input two-layer linear networks, with fixed logarithmic powers and explicit sensitivity factors.

B Technical Lemmas

Notation. For $\delta \in (0, 1/2]$, write

$$L_\delta := 1 + \log(2/\delta), \quad M_T^{\mathrm{data}} := 1 + \max_{0 \le s \le T+2} \big(|\xi_s| + |y(\xi_s)|\big).$$

Unless stated otherwise, the data are treated as deterministic.

For every width-dependent scalar observable $S = S^{(m)}$, write $\mu_S^{(m)} := \mathbb{E}S^{(m)}$ for its finite-width expectation. Whenever the deterministic infinite-width limit exists, write $S^\infty := \lim_{m\to\infty} \mu_S^{(m)}$.

We reserve the superscript ref for the reference recursions in the two-layer subsections.

We count trainable/evolving parameter blocks. The two-layer quadratic subsection uses the two blocks (U_t, V_t). The three-layer linear subsection uses the three evolving blocks $(u_t, W_t, m\mathbf{v}_t)$.

Lemma B.1 (Gaussian Poincaré and polynomial hypercontractivity)**.** *Let $G = (G_1, \dots, G_n)$ be a standard Gaussian vector in $\mathbb{R}^n$.*

1. *If $F = F(G)$ is real-valued, weakly differentiable, and $F, \nabla F \in L^2$, then*

$$\mathrm{Var}(F) \le \mathbb{E}\|\nabla F\|_2^2.$$

2. *Let $\mathcal{H}$ be a Hilbert space. Suppose $P(G) = \sum_{|\alpha|\le d} h_\alpha G^\alpha$, with $h_\alpha \in \mathcal{H}$, where the sum is finite. Then, for every $p \ge 2$,*

$$\|P(G)\|_{L^p(\mathcal{H})} \le C_d (p-1)^{d/2} \|P(G)\|_{L^2(\mathcal{H})}.$$

One may take $C_d = \sqrt{d+1}$.

Proof. Let γ_n be the standard Gaussian measure on $\mathbb{R}^n$. Let $\{H_\alpha : \alpha \in \mathbb{N}_0^n\}$ be the normalized multivariate Hermite basis of $L^2(\gamma_n)$, with $H_0 \equiv 1$. We use $\partial_i H_\alpha = \sqrt{\alpha_i} H_{\alpha - e_i}$ if $\alpha_i \ge 1$, and $\partial_i H_\alpha = 0$ if $\alpha_i = 0$.

First let $F \in W^{1,2}(\gamma_n)$. Write $F = \sum_{\alpha \in \mathbb{N}_0^n} c_\alpha H_\alpha$ in $L^2(\gamma_n)$, where $c_\alpha = \mathbb{E}[F H_\alpha]$, and set $F_N = \sum_{|\alpha| \le N} c_\alpha H_\alpha$. By Parseval, $\|F - F_N\|^2_{L^2(\gamma_n)} = \sum_{|\alpha|>N} c_\alpha^2 \to 0$.

We also have $F_N \to F$ in $W^{1,2}(\gamma_n)$. Indeed, by Gaussian integration by parts, the adjoint of ∂_i is $\partial_i^* = x_i - \partial_i$. Since $x_i H_\beta = \sqrt{\beta_i + 1} H_{\beta + e_i} + \sqrt{\beta_i} H_{\beta - e_i}$, we get $\partial_i^* H_\beta = \sqrt{\beta_i + 1} H_{\beta + e_i}$. Hence

$$\partial_i F = \sum_{\beta \in \mathbb{N}_0^n} \sqrt{\beta_i + 1}\, c_{\beta + e_i} H_\beta = \sum_{\alpha_i \ge 1} \sqrt{\alpha_i}\, c_\alpha H_{\alpha - e_i} \quad \text{in } L^2(\gamma_n).$$

Therefore $\|\partial_i F\|^2_{L^2(\gamma_n)} = \sum_{\alpha_i \ge 1} \alpha_i c_\alpha^2$, and

$$\|\partial_i F_N - \partial_i F\|^2_{L^2(\gamma_n)} = \sum_{\substack{|\alpha|>N \\ \alpha_i \ge 1}} \alpha_i c_\alpha^2 \to 0.$$

Thus $F_N \to F$ in $W^{1,2}(\gamma_n)$.

For each N, orthonormality gives

$$\mathrm{Var}(F_N) = \sum_{0<|\alpha|\le N} c_\alpha^2, \qquad \mathbb{E}\|\nabla F_N\|_2^2 = \sum_{0<|\alpha|\le N} |\alpha| c_\alpha^2.$$

Since $|\alpha| \ge 1$ for $\alpha \ne 0$, $\mathrm{Var}(F_N) \le \mathbb{E}\|\nabla F_N\|_2^2$. Passing to the limit in $W^{1,2}(\gamma_n)$ gives

$$\mathrm{Var}(F) \le \mathbb{E}\|\nabla F\|_2^2.$$

We now prove the hypercontractive estimate. Let $\mathcal{C}_k = \mathrm{span}\{H_\alpha : |\alpha| = k\}$ be the k-th Wiener chaos. Nelson's inequality gives, for every $Q_k \in \mathcal{C}_k$ and $p \ge 2$,

$$\|Q_k\|_{L^p} \le (p-1)^{k/2} \|Q_k\|_{L^2}.$$

If Q is a scalar Gaussian polynomial of degree at most d, write its chaos decomposition as $Q = \sum_{k=0}^d Q_k$, with $Q_k \in \mathcal{C}_k$. Then, by the triangle inequality, Nelson's inequality, Cauchy–Schwarz, and L^2-orthogonality of distinct chaoses,

$$\|Q\|_{L^p} \le \sum_{k=0}^d (p-1)^{k/2} \|Q_k\|_{L^2} \le \left(\sum_{k=0}^d (p-1)^k \right)^{1/2} \|Q\|_{L^2} \le \sqrt{d+1} (p-1)^{d/2} \|Q\|_{L^2}.$$

Finally, let $P(G) = \sum_{|\alpha|\le d} h_\alpha G^\alpha$ be $\mathcal{H}$-valued. The coefficient vectors h_α span a finite-dimensional subspace $V \subseteq \mathcal{H}$. Choose an orthonormal basis $e_1, \dots, e_M$ of V and write $P(G) = \sum_{j=1}^M Q_j(G) e_j$. Each Q_j is scalar and has degree at most d. Since $p/2 \ge 1$,

$$\|P\|^2_{L^p(\mathcal{H})} = \left\| \sum_{j=1}^M |Q_j|^2 \right\|_{L^{p/2}} \le \sum_{j=1}^M \|Q_j\|^2_{L^p} \le (d+1)(p-1)^d \sum_{j=1}^M \|Q_j\|^2_{L^2}.$$

But $\sum_{j=1}^M \|Q_j\|^2_{L^2} = \|P\|^2_{L^2(\mathcal{H})}$. Taking square roots gives the claimed Hilbert-valued estimate. □

Lemma B.2 (Gaussian tools)**.** *Let z_0 and $m\mathbf{v}_0$ be independent standard Gaussian vectors in $\mathbb{R}^m$. Assume that W_0 is independent of both of them and has entries i.i.d. $N(0, m^{-1})$. If $F = F(z_0, W_0, m\mathbf{v}_0)$ is a smooth scalar-valued function with polynomial growth, then*

$$\mathrm{Var}(F) \le \frac{1}{m}\mathbb{E}\|\nabla_{W_0} F\|_F^2 + \mathbb{E}\|\nabla_{z_0} F\|_2^2 + \mathbb{E}\|\nabla_{m\mathbf{v}_0} F\|_2^2.$$

In the one-hidden-layer quadratic section, this lemma may be applied by identifying the two independent Gaussian vector blocks with $z_0 = U_0$ and $m\mathbf{v}_0 = V_0$; the W_0-block is then absent, or equivalently the function under consideration is independent of a dummy W_0-block. In the three-layer linear μP section, the lemma is applied with $z_0 = u_0$, with $m\mathbf{v}_0$ equal to the output base variable, and with W_0 the middle-layer matrix. Moreover, if P is a scalar Gaussian polynomial of degree at most q, then for every $p \ge 2$,

$$\|P\|_{L^p} \le C_q (p-1)^{q/2} \|P\|_{L^2}.$$

The same estimate holds for Hilbert-valued Gaussian polynomials.

Proof. Set $\Gamma := \sqrt{m}\, W_0$ and $b := m\mathbf{v}_0$. Then Γ has i.i.d. standard Gaussian entries, $W_0 = m^{-1/2}\Gamma$, and z_0, Γ, b are independent standard Gaussian blocks, after identifying Γ with a vector in $\mathbb{R}^{m^2}$.

Define $\widetilde{F}(z, \Gamma, b) := F(z, m^{-1/2}\Gamma, b)$. By Lemma B.1,

$$\mathrm{Var}(\widetilde{F}) \le \mathbb{E}\|\nabla_z \widetilde{F}\|_2^2 + \mathbb{E}\|\nabla_\Gamma \widetilde{F}\|_F^2 + \mathbb{E}\|\nabla_b \widetilde{F}\|_2^2.$$

The chain rule gives $\nabla_z \widetilde{F} = \nabla_{z_0} F$, $\nabla_b \widetilde{F} = \nabla_{m\mathbf{v}_0} F$, and $\nabla_\Gamma \widetilde{F} = m^{-1/2}\nabla_{W_0} F$. Since $\mathrm{Var}(\widetilde{F}) = \mathrm{Var}(F)$, we obtain

$$\mathrm{Var}(F) \le \frac{1}{m}\mathbb{E}\|\nabla_{W_0} F\|_F^2 + \mathbb{E}\|\nabla_{z_0} F\|_2^2 + \mathbb{E}\|\nabla_{m\mathbf{v}_0} F\|_2^2.$$

For the hypercontractive statement, view the underlying standard Gaussian vector as $G = (z_0, \mathrm{vec}(\Gamma), b)$. Replacing W_0 by $m^{-1/2}\Gamma$ does not increase polynomial degree. Hence the scalar and Hilbert-valued estimates follow directly from Lemma B.1 with $d = q$. □

Lemma B.3 (Log-tail to moment conversion)**.** *Let $Z \ge 0$ be a random variable. Assume that there exist constants $A \ge 1$ and $\gamma \ge 1$ such that, for every $\delta \in (0, 1/2]$,*

$$\mathbb{P}\big(Z > A L_\delta^\gamma\big) \le \delta.$$

Then, for every $p \ge 1$,

$$\mathbb{E}Z^p \le C_{p,\gamma} A^p,$$

where $C_{p,\gamma}$ depends only on p and γ.

Proof. Fix $u \ge A$ and define $\delta(u) := 2\exp\Big(1 - (u/A)^{1/\gamma}\Big)$.

If $u \ge A(1 + \log 4)^\gamma$, then $\delta(u) \in (0, 1/2]$ and

$$L_{\delta(u)} = 1 + \log\frac{2}{\delta(u)} = (u/A)^{1/\gamma}.$$

Hence, for all sufficiently large u,

$$\mathbb{P}(Z > u) \le \delta(u) \le 2e\, \exp\big(-(u/A)^{1/\gamma}\big).$$

After increasing the constant, we may therefore write

$$\mathbb{P}(Z > u) \le C_\gamma \exp\big(-(u/A)^{1/\gamma}\big), \qquad u \ge 0.$$

Now use the tail-integral formula:

$$\mathbb{E}Z^p = p\int_0^\infty u^{p-1}\mathbb{P}(Z > u)\, du \le A^p + C_{\gamma,p}\int_A^\infty u^{p-1}\exp\big(-(u/A)^{1/\gamma}\big)\, du.$$

With the change of variables $s = (u/A)^{1/\gamma}$, so that $u = As^\gamma$ and $du = \gamma A s^{\gamma-1} ds$, the second term becomes

$$C_{\gamma,p} A^p \int_1^\infty s^{\gamma p - 1} e^{-s}\, ds \le C'_{\gamma,p} A^p.$$

This proves the claim. □

C ONE-HIDDEN-LAYER LINEAR NEURAL NETWORK SETTING

Setup. In the linear setting $\sigma = \mathrm{id}$, the one-hidden-layer output at the training input ξ_t is written in the form

$$g_{\theta_t}^{(m)} := \frac{1}{m}\sum_{\alpha=1}^{m} V_{t,\alpha}X_{t,\alpha}, \quad X_{t,\alpha} := \xi_t U_{t,\alpha}, \quad \chi_t := g_{\theta_t}^{(m)} - y(\xi_t).$$

Equivalently, $g_{\theta_t}^{(m)} = \frac{1}{m}\sum_{\alpha=1}^{m} V_{t,\alpha}U_{t,\alpha}\xi_t$. The data sequence $\{(\xi_s, y(\xi_s))\}_{0\le s\le T}$ is fixed throughout this subsection and satisfies $|\xi_s| \le 1$, $0 \le s \le T$. If the data are random but independent of the initialization, all probabilities and expectations below are understood conditionally on the data.

The initialization is $(V_{0,\alpha}, U_{0,\alpha})_{\alpha=1}^m$ are i.i.d. centered Gaussian pairs with deterministic covariance, and $X_{0,\alpha} := \xi_0 U_{0,\alpha}$. Thus $(V_{0,\alpha}, X_{0,\alpha})$ is also a centered Gaussian pair with deterministic covariance, but $X_{0,\alpha}$ need not have variance one unless $|\xi_0| = 1$ and $\mathrm{Var}(U_{0,\alpha}) = 1$.

Fix a finite horizon T and set $L_{m,\delta} := 1 + \log(2m/\delta)$. All constants below may depend on T and on the fixed data sequence $\{(\xi_s, y(\xi_s))\}_{0\le s\le T}$, but never on m or δ.

The finite-width coordinates evolve by

$$V_{t+1,\alpha} := V_{t,\alpha} - \chi_t X_{t,\alpha} = V_{t,\alpha} - \chi_t\xi_t U_{t,\alpha}, \quad U_{t+1,\alpha} := U_{t,\alpha} - \chi_t\xi_t V_{t,\alpha}.$$

At each time t, the variable $X_{t,\alpha}$ is always defined by $X_{t,\alpha} := \xi_t U_{t,\alpha}$.

Lemma C.1 (Reference Gaussian recursion)**.** *Define the infinite-width Gaussian reference process by the explicit initialization coupling*

$$V_{0,\alpha}^{\mathrm{ref}} := V_{0,\alpha}, \quad U_{0,\alpha}^{\mathrm{ref}} := U_{0,\alpha}, \quad X_{0,\alpha}^{\mathrm{ref}} := \xi_0 U_{0,\alpha}^{\mathrm{ref}}.$$

For $t \ge 0$, define

$$X_{t,\alpha}^{\mathrm{ref}} := \xi_t U_{t,\alpha}^{\mathrm{ref}}, \quad g_t^{\mathrm{ref}} := \mathbb{E}[V_{t,\alpha}^{\mathrm{ref}} X_{t,\alpha}^{\mathrm{ref}}], \quad \chi_t^{\mathrm{ref}} := g_t^{\mathrm{ref}} - y(\xi_t),$$

and evolve

$$V_{t+1,\alpha}^{\mathrm{ref}} := V_{t,\alpha}^{\mathrm{ref}} - \chi_t^{\mathrm{ref}} X_{t,\alpha}^{\mathrm{ref}} = V_{t,\alpha}^{\mathrm{ref}} - \chi_t^{\mathrm{ref}}\xi_t U_{t,\alpha}^{\mathrm{ref}}, \quad U_{t+1,\alpha}^{\mathrm{ref}} := U_{t,\alpha}^{\mathrm{ref}} - \chi_t^{\mathrm{ref}}\xi_t V_{t,\alpha}^{\mathrm{ref}}.$$

Then for each fixed t, the pairs $(V_{t,\alpha}^{\mathrm{ref}}, U_{t,\alpha}^{\mathrm{ref}})_{\alpha=1}^m$ are i.i.d. centered Gaussian. Consequently, $(V_{t,\alpha}^{\mathrm{ref}}, X_{t,\alpha}^{\mathrm{ref}})_{\alpha=1}^m$ are also i.i.d. centered Gaussian pairs. Moreover, there exists a deterministic covariance matrix $\Gamma_t \in \mathbb{R}^{2\times 2}$ such that

$$(V_{t,\alpha}^{\mathrm{ref}}, U_{t,\alpha}^{\mathrm{ref}}) \sim N(0, \Gamma_t), \quad g_t^{\mathrm{ref}} = \mathbb{E}[V_{t,\alpha}^{\mathrm{ref}} X_{t,\alpha}^{\mathrm{ref}}] = \xi_t(\Gamma_t)_{12}.$$

Proof. Set

$$M_t := \begin{pmatrix} 1 & -\chi_t^{\mathrm{ref}}\xi_t \\ -\chi_t^{\mathrm{ref}}\xi_t & 1 \end{pmatrix}.$$

Then there is

$$\begin{pmatrix} V_{t+1,\alpha}^{\mathrm{ref}} \\ U_{t+1,\alpha}^{\mathrm{ref}} \end{pmatrix} = M_t \begin{pmatrix} V_{t,\alpha}^{\mathrm{ref}} \\ U_{t,\alpha}^{\mathrm{ref}} \end{pmatrix}.$$

The claim follows by induction because deterministic linear images of centered Gaussian vectors are again centered Gaussian, and deterministic linear maps preserve independence across α. The covariance recursion is $\Gamma_{t+1} = M_t\Gamma_t M_t^\top$. Since $X_{t,\alpha}^{\mathrm{ref}} = \xi_t U_{t,\alpha}^{\mathrm{ref}}$, the pair $(V_{t,\alpha}^{\mathrm{ref}}, X_{t,\alpha}^{\mathrm{ref}})$ is also centered Gaussian, and

$$g_t^{\mathrm{ref}} = \mathbb{E}[V_{t,\alpha}^{\mathrm{ref}} X_{t,\alpha}^{\mathrm{ref}}] = \xi_t\mathbb{E}[V_{t,\alpha}^{\mathrm{ref}} U_{t,\alpha}^{\mathrm{ref}}] = \xi_t(\Gamma_t)_{12}.$$

□

Lemma C.2 (Gaussian maxima and centered empirical products)**.** *For each fixed $t \le T$, there is a constant $C_t \ge 1$ such that for every $\delta \in (0, 1/2]$, with probability at least $1-\delta$,*

$$\max_{1\le\alpha\le m}\left(|V_{t,\alpha}^{\mathrm{ref}}| + |U_{t,\alpha}^{\mathrm{ref}}| + |X_{t,\alpha}^{\mathrm{ref}}|\right) \le C_t\, L_{m,\delta}^{1/2},$$

and

$$\left|\frac{1}{m}\sum_{\alpha=1}^{m}\left(V_{t,\alpha}^{\mathrm{ref}} X_{t,\alpha}^{\mathrm{ref}} - g_t^{\mathrm{ref}}\right)\right| \le C_t\, m^{-1/2} L_\delta.$$

Remark C.3. The i.i.d. assertion used here concerns only the reference variables $(V^{\mathrm{ref}}_{t,\alpha}, U^{\mathrm{ref}}_{t,\alpha})_{\alpha\le m}$; no independence across α is asserted for the trained finite-width variables after $t \ge 1$.

Proof. By Lemma C.1, $(V^{\mathrm{ref}}_{t,\alpha}, U^{\mathrm{ref}}_{t,\alpha})$ are i.i.d. centered Gaussian with deterministic covariance depending only on t. Since $X^{\mathrm{ref}}_{t,\alpha} = \xi_t U^{\mathrm{ref}}_{t,\alpha}$ and $|\xi_t| \le 1$, each of $V^{\mathrm{ref}}_{t,\alpha}$, $U^{\mathrm{ref}}_{t,\alpha}$, $X^{\mathrm{ref}}_{t,\alpha}$ is sub-Gaussian with deterministic scale depending only on t. Applying a union bound over α with failure probability $\delta/2$ gives

$$\max_{1\le\alpha\le m}\big(|V^{\mathrm{ref}}_{t,\alpha}| + |U^{\mathrm{ref}}_{t,\alpha}| + |X^{\mathrm{ref}}_{t,\alpha}|\big) \le C_t L^{1/2}_{m,\delta/2} \le C_t L^{1/2}_{m,\delta}.$$

Also $V^{\mathrm{ref}}_{t,\alpha} X^{\mathrm{ref}}_{t,\alpha}$ is a product of jointly Gaussian variables, hence sub-exponential with deterministic scale depending only on t. Since the variables are i.i.d. across α, Bernstein's inequality with failure probability $\delta/2$ yields

$$\left|\frac{1}{m}\sum_{\alpha=1}^{m}\big(V^{\mathrm{ref}}_{t,\alpha} X^{\mathrm{ref}}_{t,\alpha} - g^{\mathrm{ref}}_t\big)\right| \le C_t\left(\sqrt{\frac{L_{\delta/2}}{m}} + \frac{L_{\delta/2}}{m}\right).$$

Since $L_{\delta/2} \le C L_\delta$ and $\delta \le 1/2$, the stated empirical-product bound follows after enlarging C_t.

Intersecting the two events gives probability at least $1-\delta$. □

Lemma C.4 (Exact error recursion)**.** *Define the exact remainder variables*

$$e^v_{t,\alpha} := V_{t,\alpha} - V^{\mathrm{ref}}_{t,\alpha}, \quad e^u_{t,\alpha} := U_{t,\alpha} - U^{\mathrm{ref}}_{t,\alpha}, \quad e^x_{t,\alpha} := X_{t,\alpha} - X^{\mathrm{ref}}_{t,\alpha} = \xi_t e^u_{t,\alpha}.$$

Then

$$e^v_{t+1,\alpha} = e^v_{t,\alpha} - \chi^{\mathrm{ref}}_t\, e^x_{t,\alpha} + (\chi^{\mathrm{ref}}_t - \chi_t) X_{t,\alpha}, \quad e^u_{t+1,\alpha} = e^u_{t,\alpha} - \chi^{\mathrm{ref}}_t \xi_t\, e^v_{t,\alpha} + \xi_t(\chi^{\mathrm{ref}}_t - \chi_t) V_{t,\alpha}.$$

Proof. Subtract the reference recursion from the finite-width recursion coordinatewise:

$$V_{t+1,\alpha} = V_{t,\alpha} - \chi_t X_{t,\alpha}, \qquad V^{\mathrm{ref}}_{t+1,\alpha} = V^{\mathrm{ref}}_{t,\alpha} - \chi^{\mathrm{ref}}_t X^{\mathrm{ref}}_{t,\alpha}.$$

This gives

$$e^v_{t+1,\alpha} = e^v_{t,\alpha} - \chi^{\mathrm{ref}}_t(X_{t,\alpha} - X^{\mathrm{ref}}_{t,\alpha}) + (\chi^{\mathrm{ref}}_t - \chi_t) X_{t,\alpha}.$$

Using $e^x_{t,\alpha} = X_{t,\alpha} - X^{\mathrm{ref}}_{t,\alpha}$ yields the first identity.

Similarly,

$$U_{t+1,\alpha} = U_{t,\alpha} - \chi_t \xi_t V_{t,\alpha}, \qquad U^{\mathrm{ref}}_{t+1,\alpha} = U^{\mathrm{ref}}_{t,\alpha} - \chi^{\mathrm{ref}}_t \xi_t V^{\mathrm{ref}}_{t,\alpha}.$$

Subtracting gives

$$e^u_{t+1,\alpha} = e^u_{t,\alpha} - \chi^{\mathrm{ref}}_t \xi_t e^v_{t,\alpha} + \xi_t(\chi^{\mathrm{ref}}_t - \chi_t) V_{t,\alpha}.$$

□

Lemma C.5 (Output decomposition)**.** *For every $t \le T$,*

$$g^{(m)}_{\theta_t} - g^{\mathrm{ref}}_t = \frac{1}{m}\sum_{\alpha=1}^{m}\big(V^{\mathrm{ref}}_{t,\alpha} X^{\mathrm{ref}}_{t,\alpha} - g^{\mathrm{ref}}_t\big) + \frac{1}{m}\sum_{\alpha=1}^{m}\mathrm{Err}_{t,\alpha},$$

where

$$\mathrm{Err}_{t,\alpha} := V^{\mathrm{ref}}_{t,\alpha} e^x_{t,\alpha} + X^{\mathrm{ref}}_{t,\alpha} e^v_{t,\alpha} + e^v_{t,\alpha} e^x_{t,\alpha}.$$

In particular, $\chi_t - \chi^{\mathrm{ref}}_t = g^{(m)}_{\theta_t} - g^{\mathrm{ref}}_t$.

Proof. Since

$$V_{t,\alpha} = V^{\mathrm{ref}}_{t,\alpha} + e^v_{t,\alpha}, \qquad X_{t,\alpha} = X^{\mathrm{ref}}_{t,\alpha} + e^x_{t,\alpha},$$

we have

$$V_{t,\alpha} X_{t,\alpha} = (V^{\mathrm{ref}}_{t,\alpha} + e^v_{t,\alpha})(X^{\mathrm{ref}}_{t,\alpha} + e^x_{t,\alpha}) = V^{\mathrm{ref}}_{t,\alpha} X^{\mathrm{ref}}_{t,\alpha} + \mathrm{Err}_{t,\alpha}\,.$$

Summing over α, dividing by m, and subtracting $g^{\mathrm{ref}}_t = \mathbb{E}[V^{\mathrm{ref}}_{t,\alpha} X^{\mathrm{ref}}_{t,\alpha}]$ gives the first identity. The identity for $\chi_t - \chi^{\mathrm{ref}}_t$ is immediate from the definitions. □

Lemma C.6 (Error Bound)**.** *Let*

$$\varepsilon_t := \max_{1\le\alpha\le m}\big(|e^v_{t,\alpha}| + |e^u_{t,\alpha}|\big), \quad \Delta_t := |\chi_t - \chi_t^{\mathrm{ref}}| = |g^{(m)}_{\theta_t} - g_t^{\mathrm{ref}}|.$$

On the event

$$\max_{\alpha\le m}\big(|V^{\mathrm{ref}}_{t,\alpha}| + |U^{\mathrm{ref}}_{t,\alpha}| + |X^{\mathrm{ref}}_{t,\alpha}|\big) \le C_t L_{m,\delta}^{1/2},$$

one has

$$\varepsilon_{t+1} \le C_t\Big((1+\Delta_t)\varepsilon_t + \Delta_t L_{m,\delta}^{1/2}\Big).$$

Moreover, on the event from Lemma C.2,

$$\Delta_t \le C_t m^{-1/2} L_\delta + C_t L_{m,\delta}^{1/2}\varepsilon_t + \varepsilon_t^2.$$

Proof. By Lemma C.4,

$$e^v_{t+1,\alpha} = e^v_{t,\alpha} - \chi_t^{\mathrm{ref}} e^x_{t,\alpha} + (\chi_t^{\mathrm{ref}} - \chi_t)X_{t,\alpha},$$

and

$$e^u_{t+1,\alpha} = e^u_{t,\alpha} - \chi_t^{\mathrm{ref}}\xi_t e^v_{t,\alpha} + \xi_t(\chi_t^{\mathrm{ref}} - \chi_t)V_{t,\alpha}.$$

Since $|\xi_t| \le 1$ and $e^x_{t,\alpha} = \xi_t e^u_{t,\alpha}$, there is $|e^x_{t,\alpha}| \le |e^u_{t,\alpha}|$.

Also $|\chi_t^{\mathrm{ref}}|$ is bounded by a constant depending only on the fixed time horizon and data sequence. Therefore,

$$|e^v_{t+1,\alpha}| + |e^u_{t+1,\alpha}| \le C_t\big(|e^v_{t,\alpha}| + |e^u_{t,\alpha}|\big) + C_t\Delta_t\big(|V_{t,\alpha}| + |X_{t,\alpha}|\big).$$

Because $X_{t,\alpha} = \xi_t U_{t,\alpha}$ and $|\xi_t| \le 1$, we have $|X_{t,\alpha}| \le |U_{t,\alpha}|$. Moreover,

$$|V_{t,\alpha}| + |U_{t,\alpha}| \le |V^{\mathrm{ref}}_{t,\alpha}| + |U^{\mathrm{ref}}_{t,\alpha}| + \varepsilon_t.$$

Taking maxima over α and using the reference maximum event gives

$$\varepsilon_{t+1} \le C_t\Big((1+\Delta_t)\varepsilon_t + \Delta_t L_{m,\delta}^{1/2}\Big).$$

For the second bound, Lemma C.5 gives

$$\Delta_t \le \left|\frac{1}{m}\sum_{\alpha=1}^m\big(V^{\mathrm{ref}}_{t,\alpha}X^{\mathrm{ref}}_{t,\alpha} - g_t^{\mathrm{ref}}\big)\right| + \frac{1}{m}\sum_{\alpha=1}^m |\,\mathrm{Err}_{t,\alpha}\,|.$$

The first term is bounded by $C_t m^{-1/2} L_\delta$ on the event from Lemma C.2. For the second term, using $|e^x_{t,\alpha}| \le |e^u_{t,\alpha}| \le \varepsilon_t$, we get

$$\frac{1}{m}\sum_{\alpha=1}^m |V^{\mathrm{ref}}_{t,\alpha}e^x_{t,\alpha}| \le \varepsilon_t \max_{\alpha\le m}|V^{\mathrm{ref}}_{t,\alpha}|, \quad \frac{1}{m}\sum_{\alpha=1}^m |X^{\mathrm{ref}}_{t,\alpha}e^v_{t,\alpha}| \le \varepsilon_t \max_{\alpha\le m}|X^{\mathrm{ref}}_{t,\alpha}|, \quad \frac{1}{m}\sum_{\alpha=1}^m |e^v_{t,\alpha}e^x_{t,\alpha}| \le \varepsilon_t^2.$$

Using the Gaussian maximum bound yields

$$\Delta_t \le C_t m^{-1/2} L_\delta + C_t L_{m,\delta}^{1/2}\varepsilon_t + \varepsilon_t^2.$$

□

Theorem C.7 (Fixed-time concentration in the linear setting)**.** *For every fixed $T < \infty$, define*

$$\gamma_t^{\mathrm{lin}} := \begin{cases} 0, & t = 0, \\ \dfrac{7\cdot 3^{t-1} - 2}{2}, & t \ge 1, \end{cases} \qquad \beta_t^{\mathrm{lin}} := \begin{cases} 1, & t = 0, \\ 7\cdot 3^{t-1} - 1, & t \ge 1. \end{cases}$$

Then there exists a constant $C_T \ge 1$ such that, for every $\delta \in (0, 1/2]$, with probability at least $1-\delta$,

$$|g^{(m)}_{\theta_t} - g_t^{\mathrm{ref}}| + |\chi_t - \chi_t^{\mathrm{ref}}| \le C_T\, m^{-1/2} L_{m,\delta}^{\beta_t^{\mathrm{lin}}}, \quad 0 \le t \le T.$$

Moreover, if $\varepsilon_t := \max_{\alpha\le m}\big(|e^v_{t,\alpha}| + |e^u_{t,\alpha}|\big)$, then on the same event, $\varepsilon_t \le C_T\, m^{-1/2} L_{m,\delta}^{\gamma_t^{\mathrm{lin}}}$, $0 \le t \le T$. Consequently,

$$\max_{0\le t\le T}|g^{(m)}_{\theta_t} - g_t^{\mathrm{ref}}| + \max_{0\le t\le T}|\chi_t - \chi_t^{\mathrm{ref}}| + \max_{0\le t\le T}\varepsilon_t \le C_T\, m^{-1/2} L_{m,\delta}^{\beta_T^{\mathrm{lin}}}.$$

Proof. For each $0 \le t \le T$, let $\mathcal{E}_t(\delta)$ be the intersection of the two events from Lemma C.2 with failure probability $\delta/(2T+2)$. By a union bound,

$$\mathbb{P}\Big(\bigcap_{t=0}^{T} \mathcal{E}_t(\delta)\Big) \ge 1 - \delta.$$

Work on this event and write $L := L_{m,\delta}$.

We prove by induction that

$$\Delta_t \le K_t\, m^{-1/2} L^{\beta_t^{\mathrm{lin}}}, \qquad \varepsilon_t \le K_t\, m^{-1/2} L^{\gamma_t^{\mathrm{lin}}}, \qquad 0 \le t \le T.$$

For $t \ge 1$, the explicit exponents satisfy

$$\gamma_t^{\mathrm{lin}} = \frac{\beta_t^{\mathrm{lin}} - 1}{2}, \qquad \gamma_{t+1}^{\mathrm{lin}} = \beta_t^{\mathrm{lin}} + \gamma_t^{\mathrm{lin}} + 1, \qquad \beta_{t+1}^{\mathrm{lin}} = 2\gamma_{t+1}^{\mathrm{lin}} + 1 = 3\beta_t^{\mathrm{lin}} + 2.$$

At time $t = 0$, the finite-width and reference coordinates coincide, so $\varepsilon_0 = 0$. Lemma C.2 gives

$$\Delta_0 = \left| \frac{1}{m} \sum_{\alpha=1}^{m} \big(V_{0,\alpha}^{\mathrm{ref}} X_{0,\alpha}^{\mathrm{ref}} - g_0^{\mathrm{ref}}\big) \right| \le C_0 m^{-1/2} L.$$

Thus the base case holds.

Now assume the claim at time t. By Lemma C.6,

$$\varepsilon_{t+1} \le C_t \Big((1 + \Delta_t)\varepsilon_t + \Delta_t L^{1/2} \Big).$$

Using the induction hypothesis and $m^{-1} \le m^{-1/2}$, there is

$$\varepsilon_{t+1} \le C_t \Big(m^{-1/2} L^{\gamma_t^{\mathrm{lin}}} + m^{-1/2} L^{\beta_t^{\mathrm{lin}} + \gamma_t^{\mathrm{lin}}} + m^{-1/2} L^{\beta_t^{\mathrm{lin}} + 1/2} \Big).$$

If $t = 0$, then

$$\max\Big\{ \gamma_0^{\mathrm{lin}},\ \beta_0^{\mathrm{lin}} + \gamma_0^{\mathrm{lin}},\ \beta_0^{\mathrm{lin}} + \frac{1}{2} \Big\} = \frac{3}{2} \le \gamma_1^{\mathrm{lin}} = \frac{5}{2}.$$

If $t \ge 1$, then

$$\gamma_t^{\mathrm{lin}} \le \beta_t^{\mathrm{lin}} + \gamma_t^{\mathrm{lin}} < \gamma_{t+1}^{\mathrm{lin}}, \qquad \beta_t^{\mathrm{lin}} + \frac{1}{2} \le \beta_t^{\mathrm{lin}} + \gamma_t^{\mathrm{lin}} + 1 = \gamma_{t+1}^{\mathrm{lin}}.$$

Hence

$$\varepsilon_{t+1} \le K_{t+1} m^{-1/2} L^{\gamma_{t+1}^{\mathrm{lin}}}.$$

Then the second inequality of Lemma C.6 gives

$$\Delta_{t+1} \le C_{t+1} m^{-1/2} L + C_{t+1} L^{1/2} \varepsilon_{t+1} + \varepsilon_{t+1}^2.$$

Using the bound just proved for ε_{t+1} and $m^{-1} \le m^{-1/2}$,

$$\Delta_{t+1} \le C_{t+1} m^{-1/2} \Big(L + L^{\gamma_{t+1}^{\mathrm{lin}} + 1/2} + L^{2\gamma_{t+1}^{\mathrm{lin}}} \Big).$$

If $t = 0$, then

$$\max\Big\{ 1,\ \gamma_1^{\mathrm{lin}} + \frac{1}{2},\ 2\gamma_1^{\mathrm{lin}} \Big\} = 5 < \beta_1^{\mathrm{lin}} = 6.$$

If $t \ge 1$, then

$$2\gamma_{t+1}^{\mathrm{lin}} = 7 \cdot 3^t - 2 = \beta_{t+1}^{\mathrm{lin}} - 1,$$

and

$$\gamma_{t+1}^{\mathrm{lin}} + \frac{1}{2} = \frac{7 \cdot 3^t - 1}{2} \le \beta_{t+1}^{\mathrm{lin}} - 1.$$

Hence

$$\Delta_{t+1} \le K_{t+1} m^{-1/2} L^{\beta_{t+1}^{\mathrm{lin}}}.$$

This closes the induction.

Finally, since for $\forall t \in [T]$. $\chi_t - \chi_t^{\mathrm{ref}} = g_{\theta_t}^{(m)} - g_t^{\mathrm{ref}}$, the displayed scalar estimate is exactly the asserted bound for Δ_t.

Also, since for $\forall t \in [T]$, there is $\gamma_t^{\mathrm{lin}} \le \beta_t^{\mathrm{lin}}$, the same event also yields $\varepsilon_t \le C_T m^{-1/2} L_{m,\delta}^{\beta_t^{\mathrm{lin}}}$. □

Remark C.8. For the collapsed one-hidden-layer identity recursion analyzed in this subsection, the deterministic limit used in the main theorem is defined by $\bar{g}_{\theta_t} := g_t^{\mathrm{ref}}$. Accordingly, Theorem C.7 yields the stated $\widetilde{O}(m^{-1/2})$ concentration around $\bar{g}_{\theta_t}$ in this setting.

D TWO-LAYER NEURAL NETWORK WITH QUADRATIC ACTIVATION SETTING

Notation Throughout this subsection we specialize to $\sigma(z) = z^2$, Accordingly, $X_{t,\alpha} = \xi_t^2 U_{t,\alpha}^2$, $D_{t,\alpha} = 2\xi_t U_{t,\alpha}$, and $g_{\theta_t}^{(m)} := \frac{1}{m}\sum_{\alpha=1}^m V_{t,\alpha} X_{t,\alpha}$, $\chi_t := g_{\theta_t}^{(m)} - y(\xi_t)$, where the initialization is specialized to independent standard Gaussians: $U_{0,\alpha} \sim \mathcal{N}(0,1)$, $V_{0,\alpha} \sim \mathcal{N}(0,1)$, with $\{U_{0,\alpha}, V_{0,\alpha} : 1 \le \alpha \le m\}$ independent.

The finite-width coordinate recursions thus become

$$V_{t+1,\alpha} = V_{t,\alpha} - \chi_t \xi_t^2 U_{t,\alpha}^2, \qquad U_{t+1,\alpha} = U_{t,\alpha} - 2\chi_t \xi_t^2 V_{t,\alpha} U_{t,\alpha}.$$

Fix a finite horizon T and set $L_{m,\delta} := 1 + \log(2m/\delta)$, all constants below may depend on T and on the fixed data sequence $\{(\xi_s, y(\xi_s))\}_{0\le s\le T}$, but never on m or δ. Thus the pairs $(U_{0,\alpha}, V_{0,\alpha})$ are i.i.d. standard Gaussian pairs with identity covariance, and the two coordinates in each pair are independent.

Define the deterministic reference recursion by

$$U_{0,\alpha}^{\mathrm{ref}} := U_{0,\alpha}, \quad V_{0,\alpha}^{\mathrm{ref}} := V_{0,\alpha}, \quad X_{t,\alpha}^{\mathrm{ref}} := \xi_t^2 (U_{t,\alpha}^{\mathrm{ref}})^2, \quad D_{t,\alpha}^{\mathrm{ref}} := 2\xi_t U_{t,\alpha}^{\mathrm{ref}},$$

$g_t^{\mathrm{ref}} := \mathbb{E}\big[V_{t,\alpha}^{\mathrm{ref}} X_{t,\alpha}^{\mathrm{ref}}\big]$, $\chi_t^{\mathrm{ref}} := g_t^{\mathrm{ref}} - y(\xi_t)$, $V_{t+1,\alpha}^{\mathrm{ref}} := V_{t,\alpha}^{\mathrm{ref}} - \chi_t^{\mathrm{ref}} X_{t,\alpha}^{\mathrm{ref}}$, $U_{t+1,\alpha}^{\mathrm{ref}} := U_{t,\alpha}^{\mathrm{ref}} - \chi_t^{\mathrm{ref}} \xi_t V_{t,\alpha}^{\mathrm{ref}} D_{t,\alpha}^{\mathrm{ref}}$, and $Y_{t,\alpha}^{\mathrm{ref}} := V_{t,\alpha}^{\mathrm{ref}} X_{t,\alpha}^{\mathrm{ref}} = \xi_t^2 V_{t,\alpha}^{\mathrm{ref}} (U_{t,\alpha}^{\mathrm{ref}})^2$.

Lemma D.1 (Reference polynomial structure and concentration)**.** *For each fixed $t \le T$, the coordinates $(U_{t,\alpha}^{\mathrm{ref}}, V_{t,\alpha}^{\mathrm{ref}})_{\alpha=1}^m$ are i.i.d. Moreover there exist deterministic polynomials P_t^u, $P_t^v : \mathbb{R}^2 \to \mathbb{R}$ of degree at most 2^t such that $U_{t,\alpha}^{\mathrm{ref}} = P_t^u(U_{0,\alpha}, V_{0,\alpha})$, $V_{t,\alpha}^{\mathrm{ref}} = P_t^v(U_{0,\alpha}, V_{0,\alpha})$.*

In particular, for every $p \ge 1$,

$$\sup_{\alpha\le m} \mathbb{E}\Big[|U_{t,\alpha}^{\mathrm{ref}}|^p + |V_{t,\alpha}^{\mathrm{ref}}|^p + |X_{t,\alpha}^{\mathrm{ref}}|^p + |D_{t,\alpha}^{\mathrm{ref}}|^p + |Y_{t,\alpha}^{\mathrm{ref}}|^p\Big] \le C_{p,t}.$$

Moreover, there exists $C_t \ge 1$ such that for every $\delta \in (0, 1/2]$, with probability at least $1-\delta$,

$$\max_{\alpha\le m}\big(|U_{t,\alpha}^{\mathrm{ref}}| + |V_{t,\alpha}^{\mathrm{ref}}| + |X_{t,\alpha}^{\mathrm{ref}}| + |D_{t,\alpha}^{\mathrm{ref}}|\big) \le C_t L_{m,\delta}^{2^t},$$

and

$$\left|\frac{1}{m}\sum_{\alpha=1}^m \big(Y_{t,\alpha}^{\mathrm{ref}} - g_t^{\mathrm{ref}}\big)\right| \le C_t m^{-1/2} L_{m,\delta}^{3\cdot 2^t}.$$

Proof. The reference dynamics are coordinatewise once the deterministic scalar sequence $(\chi_s^{\mathrm{ref}})_{s\le t-1}$ is fixed. Hence independence and identical distribution across α are preserved.

We prove the polynomial representation by induction on t. At time 0, $P_0^u(u,v) := u$, $P_0^v(u,v) := v$ have degree $1 = 2^0$.

Assume $U_{t,\alpha}^{\mathrm{ref}} = P_t^u(U_{0,\alpha}, V_{0,\alpha})$, $V_{t,\alpha}^{\mathrm{ref}} = P_t^v(U_{0,\alpha}, V_{0,\alpha})$ with $\deg P_t^u, \deg P_t^v \le 2^t$. Then $X_{t,\alpha}^{\mathrm{ref}} = \xi_t^2 (U_{t,\alpha}^{\mathrm{ref}})^2$ is a polynomial of degree at most 2^{t+1}, while $D_{t,\alpha}^{\mathrm{ref}} = 2\xi_t U_{t,\alpha}^{\mathrm{ref}}$ is a polynomial of degree at most 2^t.

Because χ_t^{ref} is deterministic, $V_{t+1,\alpha}^{\mathrm{ref}} = V_{t,\alpha}^{\mathrm{ref}} - \chi_t^{\mathrm{ref}} X_{t,\alpha}^{\mathrm{ref}}$ and

$$U_{t+1,\alpha}^{\mathrm{ref}} = U_{t,\alpha}^{\mathrm{ref}} - \chi_t^{\mathrm{ref}} \xi_t V_{t,\alpha}^{\mathrm{ref}} D_{t,\alpha}^{\mathrm{ref}} = U_{t,\alpha}^{\mathrm{ref}} - 2\chi_t^{\mathrm{ref}} \xi_t^2 V_{t,\alpha}^{\mathrm{ref}} U_{t,\alpha}^{\mathrm{ref}}$$

are polynomials of degree at most 2^{t+1}. This closes the induction.

The fixed-moment bound now follows because every displayed quantity is a polynomial of bounded degree in the Gaussian pair $(U_{0,\alpha}, V_{0,\alpha})$.

For the maximum bound, fix one of the random variables $Z_{t,\alpha} \in \{U_{t,\alpha}^{\mathrm{ref}}, V_{t,\alpha}^{\mathrm{ref}}, X_{t,\alpha}^{\mathrm{ref}}, D_{t,\alpha}^{\mathrm{ref}}\}$. Each $Z_{t,\alpha}$ is a Gaussian polynomial of degree at most 2^{t+1}, so Lemma B.2 gives $\|Z_{t,\alpha}\|_{L^p} \le C_t p^{2^t}$, $p \ge 2$. Choose $p := \max\{2, \lceil L_{m,\delta/2}\rceil\}$. Then Markov's inequality implies

$$\mathbb{P}\Big(|Z_{t,\alpha}| > A_t L_{m,\delta/2}^{2^t}\Big) \le \left(\frac{C_t p^{2^t}}{A_t L_{m,\delta/2}^{2^t}}\right)^p.$$

After choosing A_t large enough, RHS is at most $\delta/(8m)$. A union bound over the four choices of $Z_{t,\alpha}$ and over $\alpha \le m$ gives, with probability at least $1-\delta/2$,

$$\max_{\alpha\le m}\big(|U^{\mathrm{ref}}_{t,\alpha}| + |V^{\mathrm{ref}}_{t,\alpha}| + |X^{\mathrm{ref}}_{t,\alpha}| + |D^{\mathrm{ref}}_{t,\alpha}|\big) \le C_t L_{m,\delta}^{2^t},$$

where we used $L_{m,\delta/2} \le C L_{m,\delta}$.

For the empirical mean, the random variable $S^{\mathrm{ref}}_t := \frac{1}{m}\sum_{\alpha=1}^m \big(Y^{\mathrm{ref}}_{t,\alpha} - g^{\mathrm{ref}}_t\big)$ is a centered Gaussian polynomial of degree at most $3\cdot 2^t$ in the independent Gaussian family $\big(U_{0,\alpha}, V_{0,\alpha}\big)_{\alpha=1}^m$. Also the coordinates are i.i.d., so

$$\|S^{\mathrm{ref}}_t\|_{L^2}^2 = \mathrm{Var}(S^{\mathrm{ref}}_t) = \frac{1}{m}\,\mathrm{Var}(Y^{\mathrm{ref}}_{t,1}) \le C_t m^{-1}.$$

Another application of Lemma B.2 gives $\|S^{\mathrm{ref}}_t\|_{L^p} \le C_t m^{-1/2} p^{3\cdot 2^{t-1}}$, Choose $p := \max\{2, \lceil L_{m,\delta/2}\rceil\}$, Markov's inequality gives

$$\mathbb{P}\Big(|S^{\mathrm{ref}}_t| > A_t m^{-1/2} L_{m,\delta/2}^{3\cdot 2^t}\Big) \le \left(\frac{C_t p^{3\cdot 2^{t-1}}}{A_t L_{m,\delta/2}^{3\cdot 2^t}}\right)^p.$$

For A_t sufficiently large, this is at most $\delta/2$. Since $L_{m,\delta/2} \le C L_{m,\delta}$, the stated empirical-mean bound follows after enlarging C_t. Intersecting the maximum event and the empirical-mean event gives probability at least $1-\delta$. □

Lemma D.2 (Deterministic growth and exact error bounds for $\sigma(z) = z^2$)**.** *Define*

$$e^u_{t,\alpha} := U_{t,\alpha} - U^{\mathrm{ref}}_{t,\alpha}, \quad e^v_{t,\alpha} := V_{t,\alpha} - V^{\mathrm{ref}}_{t,\alpha}, \quad \varepsilon_t := \max_{\alpha\le m}\big(|e^u_{t,\alpha}| + |e^v_{t,\alpha}|\big), \quad \Delta_t := |g^{(m)}_{\theta_t} - g^{\mathrm{ref}}_t|,$$

and

$$M_t := 1 + \max_{\alpha\le m}\big(|U_{t,\alpha}| + |V_{t,\alpha}| + |U^{\mathrm{ref}}_{t,\alpha}| + |V^{\mathrm{ref}}_{t,\alpha}|\big).$$

Then, for every $t \le T$,

$$|X_{t,\alpha}| + |X^{\mathrm{ref}}_{t,\alpha}| \le C_t M_t^2, \qquad |D_{t,\alpha}| + |D^{\mathrm{ref}}_{t,\alpha}| \le C_t M_t,$$

$$|X_{t,\alpha} - X^{\mathrm{ref}}_{t,\alpha}| \le C_t M_t |e^u_{t,\alpha}|, \qquad |D_{t,\alpha} - D^{\mathrm{ref}}_{t,\alpha}| \le C_t |e^u_{t,\alpha}|.$$

Moreover,

$$M_{t+1} \le C_t M_t^5, \quad \varepsilon_{t+1} \le C_t\big(M_t\varepsilon_t + M_t^2\Delta_t\big), \quad \Delta_t \le \left|\frac{1}{m}\sum_{\alpha=1}^m \big(Y^{\mathrm{ref}}_{t,\alpha} - g^{\mathrm{ref}}_t\big)\right| + C_t M_t^2 \varepsilon_t.$$

Proof. Since $\sigma(z) = z^2$ and $\sigma'(z) = 2z$, there is

$$X_{t,\alpha} = \xi_t^2 U_{t,\alpha}^2, \quad X^{\mathrm{ref}}_{t,\alpha} = \xi_t^2 (U^{\mathrm{ref}}_{t,\alpha})^2, \quad D_{t,\alpha} = 2\xi_t U_{t,\alpha}, \quad D^{\mathrm{ref}}_{t,\alpha} = 2\xi_t U^{\mathrm{ref}}_{t,\alpha}.$$

Therefore

$$|X_{t,\alpha}| + |X^{\mathrm{ref}}_{t,\alpha}| \le C_t M_t^2, \qquad |D_{t,\alpha}| + |D^{\mathrm{ref}}_{t,\alpha}| \le C_t M_t.$$

Also, because

$$X_{t,\alpha} - X^{\mathrm{ref}}_{t,\alpha} = \xi_t^2\big(U^2_{t,\alpha} - (U^{\mathrm{ref}}_{t,\alpha})^2\big) = \xi_t^2\big(U_{t,\alpha} + U^{\mathrm{ref}}_{t,\alpha}\big)e^u_{t,\alpha},$$

there is

$$|X_{t,\alpha} - X^{\mathrm{ref}}_{t,\alpha}| \le C_t M_t |e^u_{t,\alpha}|.$$

Likewise, $D_{t,\alpha} - D^{\mathrm{ref}}_{t,\alpha} = 2\xi_t e^u_{t,\alpha}$ gives $|D_{t,\alpha} - D^{\mathrm{ref}}_{t,\alpha}| \le C_t |e^u_{t,\alpha}|$.

Next, due to

$$|g^{(m)}_{\theta_t}| \le \frac{1}{m}\sum_{\alpha=1}^m |V_{t,\alpha}|\,|X_{t,\alpha}| \le C_t M_t^3.$$

and g^{ref}_t is deterministic and finite, so after increasing C_t and using $M_t \ge 1$,

$$|g^{\mathrm{ref}}_t| + |\chi_t| + |\chi^{\mathrm{ref}}_t| \le C_t M_t^3.$$

Hence the finite-width update gives

$$|V_{t+1,\alpha}| \le |V_{t,\alpha}| + |\chi_t|\,|X_{t,\alpha}| \le C_t(M_t + M_t^3 M_t^2) \le C_t M_t^5,$$

and

$$|U_{t+1,\alpha}| \le |U_{t,\alpha}| + 2|\chi_t|\,|\xi_t|^2\,|V_{t,\alpha}|\,|U_{t,\alpha}| \le C_t(M_t + M_t^3 M_t M_t) \le C_t M_t^5.$$

The same bound holds for the reference coordinates, so $M_{t+1} \le C_t M_t^5$.

Subtracting the finite-width and reference recursions gives

$$e^v_{t+1,\alpha} = e^v_{t,\alpha} - \chi_t^{\mathrm{ref}}\big(X_{t,\alpha} - X^{\mathrm{ref}}_{t,\alpha}\big) + (\chi_t^{\mathrm{ref}} - \chi_t)X_{t,\alpha},$$

and

$$e^u_{t+1,\alpha} = e^u_{t,\alpha} - 2\chi_t^{\mathrm{ref}}\xi_t^2\Big(V_{t,\alpha}U_{t,\alpha} - V^{\mathrm{ref}}_{t,\alpha}U^{\mathrm{ref}}_{t,\alpha}\Big) + 2\xi_t^2(\chi_t^{\mathrm{ref}} - \chi_t)V_{t,\alpha}U_{t,\alpha}.$$

Now

$$\big|V_{t,\alpha}U_{t,\alpha} - V^{\mathrm{ref}}_{t,\alpha}U^{\mathrm{ref}}_{t,\alpha}\big| \le |e^v_{t,\alpha}|\,|U_{t,\alpha}| + |V^{\mathrm{ref}}_{t,\alpha}|\,|e^u_{t,\alpha}| \le C_t M_t\big(|e^u_{t,\alpha}| + |e^v_{t,\alpha}|\big).$$

Using the pointwise bounds above together with $|\chi_t - \chi_t^{\mathrm{ref}}| = \Delta_t$, we obtain

$$|e^v_{t+1,\alpha}| \le |e^v_{t,\alpha}| + C_t M_t|e^u_{t,\alpha}| + C_t M_t^2\Delta_t,$$

and

$$|e^u_{t+1,\alpha}| \le |e^u_{t,\alpha}| + C_t M_t\big(|e^u_{t,\alpha}| + |e^v_{t,\alpha}|\big) + C_t M_t^2\Delta_t.$$

Taking maxima over α gives $\varepsilon_{t+1} \le C_t\big(M_t\varepsilon_t + M_t^2\Delta_t\big)$.

For the output difference, use the exact factorization

$$\begin{aligned} V_{t,\alpha}X_{t,\alpha} - Y^{\mathrm{ref}}_{t,\alpha} &= \xi_t^2\Big(V_{t,\alpha}U_{t,\alpha}^2 - V^{\mathrm{ref}}_{t,\alpha}(U^{\mathrm{ref}}_{t,\alpha})^2\Big) \\ &= \xi_t^2\Big((V_{t,\alpha} - V^{\mathrm{ref}}_{t,\alpha})U_{t,\alpha}^2 + V^{\mathrm{ref}}_{t,\alpha}(U_{t,\alpha} - U^{\mathrm{ref}}_{t,\alpha})(U_{t,\alpha} + U^{\mathrm{ref}}_{t,\alpha})\Big), \end{aligned}$$

there is

$$|V_{t,\alpha}X_{t,\alpha} - Y^{\mathrm{ref}}_{t,\alpha}| \le C_t M_t^2|e^v_{t,\alpha}| + C_t M_t^2|e^u_{t,\alpha}| \le C_t M_t^2\varepsilon_t.$$

Therefore

$$\Delta_t \le \left|\frac{1}{m}\sum_{\alpha=1}^m\big(Y^{\mathrm{ref}}_{t,\alpha} - g_t^{\mathrm{ref}}\big)\right| + \frac{1}{m}\sum_{\alpha=1}^m|V_{t,\alpha}X_{t,\alpha} - Y^{\mathrm{ref}}_{t,\alpha}| \le \left|\frac{1}{m}\sum_{\alpha=1}^m\big(Y^{\mathrm{ref}}_{t,\alpha} - g_t^{\mathrm{ref}}\big)\right| + C_t M_t^2\varepsilon_t.$$

$\chi_t - \chi_t^{\mathrm{ref}} = g^{(m)}_{\theta_t} - g_t^{\mathrm{ref}}$ is immediate from the definition. □

Theorem D.3 (Fixed-time concentration in the two-layer pure quadratic setting)**.** *Fix $T < \infty$ and define $a_t := 5^t$, $\gamma_t := 5^t$, $\beta_t := 3\cdot 5^t$. Then there exists a constant $C_T \ge 1$ such that, for every $\delta \in (0, 1/2]$, with probability at least $1-\delta$, $|g^{(m)}_{\theta_t} - g_t^{\mathrm{ref}}| + |\chi_t - \chi_t^{\mathrm{ref}}| \le C_T m^{-1/2}L^{\beta_t}_{m,\delta}$, $\varepsilon_t \le C_T m^{-1/2}L^{\gamma_t}_{m,\delta}$, $M_t \le C_T L^{a_t}_{m,\delta}$, $0 \le t \le T$, Consequently, $\max_{0\le t\le T}\big(|g^{(m)}_{\theta_t} - g_t^{\mathrm{ref}}| + |\chi_t - \chi_t^{\mathrm{ref}}|\big) \le C_T m^{-1/2}L^{\beta_T}_{m,\delta}$, $\max_{0\le t\le T}\varepsilon_t \le C_T m^{-1/2}L^{\beta_T}_{m,\delta}$, $\max_{0\le t\le T}M_t \le C_T L^{a_T}_{m,\delta}$.*

Proof. For each $0 \le t \le T$, apply Lemma D.1 with failure probability $\eta := \frac{\delta}{T+1}$, and let $\mathcal{E}_t$ be the intersection of the corresponding reference-maximum and empirical-mean events. A union bound gives

$$\mathbb{P}\Big(\bigcap_{t=0}^T\mathcal{E}_t\Big) \ge 1-\delta, \quad L_{m,\eta} = 1 + \log\frac{2m(T+1)}{\delta} \le C_T L_{m,\delta} := C_T L.$$

We prove by induction on t that on $\bigcap_{s=0}^T\mathcal{E}_s$,

$$M_t \le K_t L^{a_t}, \quad \varepsilon_t \le K_t m^{-1/2}L^{\gamma_t}, \quad \Delta_t \le K_t m^{-1/2}L^{\beta_t}, \quad 0 \le t \le T,$$

where $\Delta_t = |g^{(m)}_{\theta_t} - g_t^{\mathrm{ref}}| = |\chi_t - \chi_t^{\mathrm{ref}}|$.

At time $t = 0$, the initial finite-width and reference coordinates coincide, so $\varepsilon_0 = 0$ and $M_0 = 1 + 2\max_{\alpha \le m}\big(|U_{0,\alpha}| + |V_{0,\alpha}|\big) \le K_0 L$, after enlarging K_0. This is because Gaussian maxima scale like $L^{1/2}$. Since $a_0 = 1$, this is exactly $M_0 \le K_0 L^{a_0}$. Finally, Lemma D.2 and $\varepsilon_0 = 0$ give

$$\Delta_0 = \left| \frac{1}{m} \sum_{\alpha=1}^{m} \big(Y_{0,\alpha}^{\mathrm{ref}} - g_0^{\mathrm{ref}}\big) \right| \le K_0 m^{-1/2} L^3 = K_0 m^{-1/2} L^{\beta_0}.$$

Thus the base case holds.

Assume now that the bounds hold at time $t < T$. By Lemma D.2, $M_{t+1} \le C_t M_t^5$. Using the induction hypothesis, $M_{t+1} \le C_t L^{5a_t} = C_t L^{a_{t+1}}$.

Again by Lemma D.2 we get $\varepsilon_{t+1} \le C_t\big(M_t \varepsilon_t + M_t^2 \Delta_t\big)$. Substituting the induction bounds gives

$$\varepsilon_{t+1} \le C_t m^{-1/2}\big(L^{a_t+\gamma_t} + L^{2a_t+\beta_t}\big).$$

Now $a_t + \gamma_t = 2 \cdot 5^t \le 5^{t+1} = \gamma_{t+1}$, $2a_t + \beta_t = 2 \cdot 5^t + 3 \cdot 5^t = 5^{t+1} = \gamma_{t+1}$. Therefore $\varepsilon_{t+1} \le K_{t+1} m^{-1/2} L^{\gamma_{t+1}}$.

Finally, apply the output bound from Lemma D.2 at time $t + 1$:

$$\Delta_{t+1} \le \left| \frac{1}{m} \sum_{\alpha=1}^{m} \big(Y_{t+1,\alpha}^{\mathrm{ref}} - g_{t+1}^{\mathrm{ref}}\big) \right| + C_{t+1} M_{t+1}^2 \varepsilon_{t+1}.$$

On $\mathcal{E}_{t+1}$ the empirical term is bounded by $C_{t+1} m^{-1/2} L^{3 \cdot 2^{t+1}}$. Using the already proved bounds on M_{t+1} and ε_{t+1}, there is

$$M_{t+1}^2 \varepsilon_{t+1} \le K_{t+1} m^{-1/2} L^{2a_{t+1}+\gamma_{t+1}} = K_{t+1} m^{-1/2} L^{3 \cdot 5^{t+1}}.$$

Since $3 \cdot 2^{t+1} \le \beta_{t+1}$, $2a_{t+1} + \gamma_{t+1} = \beta_{t+1}$, we conclude that $\Delta_{t+1} \le K_{t+1} m^{-1/2} L^{\beta_{t+1}}$. This closes the induction.

Since $\gamma_t \le \beta_t$ for every t, the same event also yields $\varepsilon_t \le C_T m^{-1/2} L_{m,\delta}^{\beta_t}$. Taking the maximum over $0 \le t \le T$ proves the final statement. □

Lemma D.4 (Agreement of the two-layer reference limits with the Tensor Program limits)**.** *Fix a finite horizon $T < \infty$. Consider either the one-hidden-layer linear model $\sigma(z) = z$ from Section C or the one-hidden-layer pure quadratic model $\sigma(z) = z^2$ from Section D, under the scalar-input μP recursions described in the main text. Let $g_{\theta_t}^{\mathrm{TP}}(\xi)$ denote the standard Tensor Program infinite-width output limit for the corresponding compiled finite-time computation. Let $g_t^{\mathrm{ref}}(\xi)$ denote the deterministic reference output constructed in the corresponding subsection:*

$$g_t^{\mathrm{ref}}(\xi) = \xi\, \mathbb{E}\big[V_{t,\alpha}^{\mathrm{ref}} U_{t,\alpha}^{\mathrm{ref}}\big] \quad \text{if } \sigma(z) = z,$$

while

$$g_t^{\mathrm{ref}}(\xi) = \xi^2\, \mathbb{E}\big[V_{t,\alpha}^{\mathrm{ref}} (U_{t,\alpha}^{\mathrm{ref}})^2\big] \quad \text{if } \sigma(z) = z^2.$$

Then, for every fixed $0 \le t \le T$ and every fixed bounded input ξ,

$$g_{\theta_t}^{\mathrm{TP}}(\xi) = g_t^{\mathrm{ref}}(\xi).$$

Consequently, in the one-hidden-layer linear and pure-quadratic settings one may define

$$\bar{g}_{\theta_t}(\xi) := g_t^{\mathrm{ref}}(\xi)$$

and this deterministic reference output is the Tensor Program infinite-width output limit.

Proof. For the linear model,

$$g_{\theta_t}^{(m)}(\xi) = \xi\, \frac{1}{m} \sum_{\alpha=1}^{m} V_{t,\alpha} U_{t,\alpha}.$$

The proof of Theorem C.7 applies verbatim to any fixed bounded evaluation input ξ, after replacing $X_{t,\alpha} = \xi_t U_{t,\alpha}$ by $\xi U_{t,\alpha}$ only in the displayed output evaluation. Hence

$$g_{\theta_t}^{(m)}(\xi) - g_t^{\mathrm{ref}}(\xi) \xrightarrow[m\to\infty]{\mathbb{P}} 0.$$

For the quadratic model,

$$g_{\theta_t}^{(m)}(\xi) = \xi^2 \frac{1}{m} \sum_{\alpha=1}^{m} V_{t,\alpha} U_{t,\alpha}^2,$$

and the proof of Theorem D.3, again applied to a fixed bounded evaluation input ξ, gives

$$g_{\theta_t}^{(m)}(\xi) - g_t^{\mathrm{ref}}(\xi) \xrightarrow[m\to\infty]{\mathbb{P}} 0.$$

On the other hand, the Tensor Program theorem applied to the same compiled finite-time μP computation gives

$$g_{\theta_t}^{(m)}(\xi) \xrightarrow[m\to\infty]{\mathbb{P}} g_{\theta_t}^{\mathrm{TP}}(\xi).$$

Limits in probability are unique, so $g_{\theta_t}^{\mathrm{TP}}(\xi) = g_t^{\mathrm{ref}}(\xi)$. □

E EXPECTATION-INDUCTION TOOLS

The only expectation input used below is the following elementary fact: a fixed polynomial of scalar observables whose fluctuations are of order $m^{-1/2}$ has expectation equal to the same polynomial applied to the expectations, up to an $O(m^{-1})$ error. The proof is a finite algebraic induction after writing each observable as its expectation plus a centered fluctuation.

Lemma E.1 (Polynomial expectations from centered fluctuations)**.** *Fix integers $n, D \geq 1$. Let $Z^{(m)} = (Z_1^{(m)}, \ldots, Z_n^{(m)})$ be a random vector and write $\mu_i^{(m)} := \mathbb{E} Z_i^{(m)}$, $\mu^{(m)} := (\mu_1^{(m)}, \ldots, \mu_n^{(m)})$. Assume that, for every fixed $q \geq 2$,*

$$\max_{1\leq i\leq n} \|Z_i^{(m)} - \mu_i^{(m)}\|_{L^q} \leq C_q m^{-1/2}, \quad \max_{1\leq i\leq n} \left(\|Z_i^{(m)}\|_{L^q} + |\mu_i^{(m)}|\right) \leq C_q.$$

Then, for every polynomial $P : \mathbb{R}^n \to \mathbb{R}$ of degree at most D,

$$\mathbb{E} P(Z^{(m)}) = P(\mu^{(m)}) + O(m^{-1}).$$

The constant depends on P, n, D and the fixed moment constants, but not on m.

Proof. By linearity it is enough to treat a monomial $P(z) = z^\alpha = \prod_{i=1}^n z_i^{\alpha_i}$, $|\alpha| \leq D$.

Write $Z_i^{(m)} = \mu_i^{(m)} + \eta_i^{(m)}$, $\mathbb{E}\eta_i^{(m)} = 0$, expanding the product gives

$$\prod_{i=1}^{n} (\mu_i^{(m)} + \eta_i^{(m)})^{\alpha_i} = \prod_{i=1}^{n} (\mu_i^{(m)})^{\alpha_i} + \mathcal{L} + \mathcal{R},$$

where $\mathcal{L}$ is the sum of all terms containing exactly one centered factor and $\mathcal{R}$ is the sum of all terms containing at least two centered factors. Every term in $\mathcal{L}$ has zero expectation because all remaining factors are deterministic.

Consider a term in $\mathcal{R}$:

$$C \prod_i (\mu_i^{(m)})^{a_i} \prod_i (\eta_i^{(m)})^{b_i}, \quad \sum_i b_i \geq 2, \quad \sum_i (a_i + b_i) \leq D.$$

The deterministic factors are uniformly bounded. For the centered factors, Hölder's inequality and the assumed L^q bounds give

$$\mathbb{E}\left|\prod_i (\eta_i^{(m)})^{b_i}\right| \leq \prod_{i:b_i>0} \|\eta_i^{(m)}\|_{L^{Db_i}}^{b_i} \leq C_D m^{-\frac{1}{2}\sum_i b_i} \leq C_D m^{-1}.$$

There are only finitely many terms, so $\mathbb{E} Z^\alpha = (\mu^{(m)})^\alpha + O(m^{-1})$. Summing over the finitely many monomials of P proves the lemma. □

Proposition E.2 (Closed scalar-state expectation induction)**.** *Fix $T < \infty$. For each $0 \leq t \leq T$, let $Y_t^{(m)} = (Y_{t,1}^{(m)}, \ldots, Y_{t,N_t}^{(m)})$ be a finite scalar state. Assume the following.*

1. *There is a deterministic vector $Y_0^\infty \in \mathbb{R}^{N_0}$ such that $\mathbb{E}Y_{0,i}^{(m)} = Y_{0,i}^\infty + O(m^{-1})$, $\forall i \in [N_0]$.*
2. *For each $0 \le t < T$, there are fixed polynomials $P_{t,i} : \mathbb{R}^{N_t} \to \mathbb{R}$, $1 \le i \le N_{t+1}$, independent of m, such that*
$$Y_{t+1,i}^{(m)} = P_{t,i}(Y_t^{(m)}) + R_{t,i}^{(m)},\ 1 \le i \le N_{t+1}, \quad \mathbb{E}R_{t,i}^{(m)} = O(m^{-1}).$$
3. *The deterministic scalar state is defined by the matching recursion $Y_{t+1,i}^\infty := P_{t,i}(Y_t^\infty)$.*
4. *For every fixed $q \ge 2$ and every $t \le T$,*
$$\max_{i\in[N_t]} \|Y_{t,i}^{(m)} - \mathbb{E}Y_{t,i}^{(m)}\|_{L^q} \le C_{t,q} m^{-1/2}, \quad \max_{i\in[N_t]} \big(\|Y_{t,i}^{(m)}\|_{L^q} + |\mathbb{E}Y_{t,i}^{(m)}|\big) \le C_{t,q}.$$

Then, for every $0 \le t \le T$ and every coordinate i, $\mathbb{E}Y_{t,i}^{(m)} = Y_{t,i}^\infty + O(m^{-1})$.

Proof. The proof is by induction on t. The claim at $t = 0$ is satisfied by assumption 1. Assume it holds at time t. For coordinate i at time $t+1$, there is $\mathbb{E}Y_{t+1,i}^{(m)} = \mathbb{E}P_{t,i}(Y_t^{(m)}) + \mathbb{E}R_{t,i}^{(m)}$. By Lemma E.1 we have
$$\mathbb{E}P_{t,i}(Y_t^{(m)}) = P_{t,i}(\mathbb{E}Y_t^{(m)}) + O(m^{-1}).$$
By $\mathbb{E}R_{t,i}^{(m)} = O(m^{-1})$ there is $\mathbb{E}Y_{t+1,i}^{(m)} = P_{t,i}(\mathbb{E}Y_t^{(m)}) + O(m^{-1})$. By the induction hypothesis, $\mathbb{E}Y_t^{(m)} = Y_t^\infty + O(m^{-1})$ coordinate-wisely. Since $P_{t,i}$ is a fixed polynomial and the deterministic trajectory is bounded for fixed T,
$$P_{t,i}(\mathbb{E}Y_t^{(m)}) = P_{t,i}(Y_t^\infty) + O(m^{-1}) = Y_{t+1,i}^\infty + O(m^{-1}).$$
This proves the induction step. □

Lemma E.3 (Initialization output bias)**.** *In both settings of this appendix, the initialized output has zero expectation. In the three-layer linear notation, $\mathbb{E}g_{\theta_0}^{(m)} = \mathbb{E}\frac{\xi_0}{m}\langle m\mathbf{v}_0, W_0u_0\rangle = 0$. In the pure quadratic notation, $\mathbb{E}g_{\theta_0}^{(m)} = 0$. Consequently, if the deterministic initialized output is defined by $\bar{g}_{\theta_0} := 0$, then $|\mathbb{E}g_{\theta_0}^{(m)} - \bar{g}_{\theta_0}| = 0$.*

Proof. For the three layer linear model, condition on u_0 and W_0. The vector $m\mathbf{v}_0$ is centered and independent of (u_0, W_0), so
$$\mathbb{E}\left[g_{\theta_0}^{(m)} \mid u_0, W_0\right] = \frac{\xi_0}{m}\langle \mathbb{E}[m\mathbf{v}_0], W_0u_0\rangle = 0.$$
Taking expectation gives $\mathbb{E}g_{\theta_0}^{(m)} = 0$.

For the pure quadratic model, the output weights $V_{0,\alpha}$ are again centered and independent of the hidden features at initialization. Conditioning on all non-output initialization variables gives zero conditional expectation for the output. Thus $\mathbb{E}g_{\theta_0}^{(m)} = 0$ in that case as well. □

F TWO HIDDEN-LAYER LINEAR NEURAL NETWORK SETTING

Setup and notation. We analyze the scalar-input three-layer linear μP model by tracking the first-layer parameter vector rather than an input-dependent first-layer activation. Let $u_t \in \mathbb{R}^m$ be the first-layer parameter vector, let $W_t \in \mathbb{R}^{m\times m}$ be the middle-layer matrix, and let $m\mathbf{v}_t \in \mathbb{R}^m$ be the μP base output variable. For an evaluation input ξ, define $g_{\theta_t}^{(m)}(\xi) := \frac{\xi}{m}\langle m\mathbf{v}_t, W_tu_t\rangle$. At training step t, write $z_t := W_tu_t$, $d_t := W_t^\top(m\mathbf{v}_t)$, $G_t^{(m)} := m^{-1}\langle m\mathbf{v}_t, z_t\rangle$, $g_{\theta_t}^{(m)} := g_{\theta_t}^{(m)}(\xi_t) = \xi_tG_t^{(m)}$, $\chi_t := g_{\theta_t}^{(m)} - y(\xi_t)$, and $\rho_t := \xi_t\chi_t$. Under the normal initialization, the matrix W_0 has i.i.d. centered Gaussian entries with variance m^{-1}, while $u_0, m\mathbf{v}_0 \in \mathbb{R}^m$ are independent standard Gaussian vectors. Let $\mathcal{G}_0 := (W_0, u_0, m\mathbf{v}_0)$ denote the initial Gaussian input. All polynomial and Gaussian-chaos statements below are taken with respect to the Gaussian coordinates of $\mathcal{G}_0$, conditional on the

deterministic data sequence $(\xi_t, y(\xi_t))_{t\geq 0}$. All constants C_T below may depend on T and M_T^{data}, but never on m or δ.

Using the shorthand $\rho_t = \xi_t \chi_t$, the exact μP gradient-descent recursions are

$$m\mathbf{v}_{t+1} = m\mathbf{v}_t - \rho_t z_t, \quad u_{t+1} = u_t - \rho_t d_t, \quad W_{t+1} = W_t - \frac{\rho_t}{m}(m\mathbf{v}_t)u_t^\top.$$

For $0 \leq r \leq t$, define $K_{r,t}^{xx} := \frac{1}{m} u_r^\top u_t$, $K_{t,r}^{vv} := \frac{1}{m}\langle m\mathbf{v}_t, m\mathbf{v}_r\rangle$, $G_{t,r}^{xv} := \frac{1}{m}\langle z_t, m\mathbf{v}_r\rangle$, $J_t^x := \frac{1}{m} u_t^\top W_0^\top (m\mathbf{v}_t)$, $D_{r,t}^x := K_{r,t+1}^{xx} - K_{r,t}^{xx}$. The superscripts x, xx, and xv are retained only for notational compatibility with the rest of the appendix.

Also define

$$A_t^{vd} := \frac{1}{m}\langle m\mathbf{v}_t, W_0 d_t\rangle, \quad Q_t^z := \frac{1}{m}\|z_t\|_2^2, \quad H_t^{z,Wd} := \frac{1}{m}\langle z_t, W_0 d_t\rangle.$$

$$B_0^v := \frac{1}{m}\|W_0 d_0\|_2^2, \quad S_0^z := \frac{1}{m}\langle z_0, W_0 W_0^\top z_0\rangle, \quad T_0^z := \frac{1}{m}\langle W_0 d_0, W_0 W_0^\top z_0\rangle.$$

For scalar-valued F, $\nabla_{W_0} F$ is the usual matrix gradient and $\|\nabla_{W_0} F\|_F$ is its Frobenius norm. For vector-valued U, set $\|D_{W_0}U\|_{\mathrm{F}\to 2} := \sup_{\|H\|_F=1} \|D_{W_0}U[H]\|_2$, $\|D_{u_0}U\|_{2\to 2} := \sup_{\|h\|_2=1}\|D_{u_0}U[h]\|_2$, and $\|D_{m\mathbf{v}_0}U\|_{2\to 2} := \sup_{\|h\|_2=1}\|D_{m\mathbf{v}_0}U[h]\|_2$. We abbreviate $\|\nabla_{W_0}U\|_F := \|D_{W_0}U\|_{\mathrm{F}\to 2}$, $\|\nabla_{u_0}U\|_2 := \|D_{u_0}U\|_{2\to 2}$, and $\|\nabla_{m\mathbf{v}_0}U\|_2 := \|D_{m\mathbf{v}_0}U\|_{2\to 2}$. For matrix-valued M, define $\|D_{W_0}M\|_{\mathrm{F}\to\mathrm{F}} := \sup_{\|H\|_F=1}\|D_{W_0}M[H]\|_F$, $\|D_{u_0}M\|_{2\to\mathrm{F}} := \sup_{\|h\|_2=1}\|D_{u_0}M[h]\|_F$, and $\|D_{m\mathbf{v}_0}M\|_{2\to\mathrm{F}} := \sup_{\|h\|_2=1}\|D_{m\mathbf{v}_0}M[h]\|_F$. When no confusion is possible, we also write $\|\nabla_{W_0}M\|_{F\to F} := \|D_{W_0}M\|_{\mathrm{F}\to\mathrm{F}}$, $\|\nabla_{u_0}M\|_{2\to F} := \|D_{u_0}M\|_{2\to\mathrm{F}}$, and $\|\nabla_{m\mathbf{v}_0}M\|_{2\to F} := \|D_{m\mathbf{v}_0}M\|_{2\to\mathrm{F}}$.

We repeatedly use Lemma B.2 in the form of the weighted Gaussian Poincaré inequality associated with the Gaussian law of $\mathcal{G}_0$.

Lemma F.1. *For every $t \geq 0$, $W_{t+1} = W_0 - \sum_{r=0}^t \frac{\rho_r}{m}(m\mathbf{v}_r)u_r^\top$. Consequently,*

$$z_{t+1} = W_0 u_{t+1} - \sum_{r=0}^t \rho_r\, m\mathbf{v}_r\, K_{r,t+1}^{xx}, \quad d_{t+1} = W_0^\top(m\mathbf{v}_{t+1}) - \sum_{r=0}^t \rho_r\, u_r\, K_{t+1,r}^{vv}.$$

Proof. The weight recursion $W_{t+1} = W_t - \frac{\rho_t}{m}(m\mathbf{v}_t)u_t^\top$ is equal to

$$W_{t+1} = W_0 - \sum_{r=0}^t \frac{\rho_r}{m}(m\mathbf{v}_r)u_r^\top.$$

Multiplying u_{t+1} on both sides gives

$$z_{t+1} = W_0 u_{t+1} - \sum_{r=0}^t \frac{\rho_r}{m}(m\mathbf{v}_r)u_r^\top u_{t+1} = W_0 u_{t+1} - \sum_{r=0}^t \rho_r\, m\mathbf{v}_r\, K_{r,t+1}^{xx}.$$

Similarly, multiplying the transpose identity on $m\mathbf{v}_{t+1}$ yields

$$d_{t+1} = W_0^\top(m\mathbf{v}_{t+1}) - \sum_{r=0}^t \frac{\rho_r}{m} u_r \langle m\mathbf{v}_r, m\mathbf{v}_{t+1}\rangle = W_0^\top(m\mathbf{v}_{t+1}) - \sum_{r=0}^t \rho_r\, u_r\, K_{t+1,r}^{vv}.$$

□

Lemma F.2 (Step-0 scalar)**.** *There exists a constant C, depending only on M_T^{data} and in this lemma only on $|\xi_0| + |y(\xi_0)|$, such that for every $\delta \in (0, 1/2]$, with probability at least $1-\delta$,*

$$\begin{aligned}
&|K_{0,0}^{xx} - \mathbb{E}K_{0,0}^{xx}| + |K_{0,0}^{vv} - \mathbb{E}K_{0,0}^{vv}| + |G_{0,0}^{xv} - \mathbb{E}G_{0,0}^{xv}| + |g_{\theta_0}^{(m)} - \mathbb{E}g_{\theta_0}^{(m)}| \\
&+ |\chi_0 - \mathbb{E}\chi_0| + |\rho_0 - \mathbb{E}\rho_0| + |Q_0^z - \mathbb{E}Q_0^z| + |A_0^{vd} - \mathbb{E}A_0^{vd}| + |H_0^{z,Wd} - \mathbb{E}H_0^{z,Wd}| \\
&+ |B_0^v - \mathbb{E}B_0^v| + |S_0^z - \mathbb{E}S_0^z| + |T_0^z - \mathbb{E}T_0^z| + |K_{0,1}^{xx} - \mathbb{E}K_{0,1}^{xx}| \\
&+ |D_{0,0}^x - \mathbb{E}D_{0,0}^x| + |K_{1,0}^{vv} - \mathbb{E}K_{1,0}^{vv}| \leq C m^{-1/2} L_\delta^8.
\end{aligned}$$

Moreover, on the same event,

$$|K^{xx}_{0,0}| + |K^{vv}_{0,0}| + |G^{xv}_{0,0}| + |g^{(m)}_{\theta_0}| + |\chi_0| + |\rho_0| + |Q^z_0| + |A^{vd}_0| + |H^{z,Wd}_0| + |B^v_0| + |S^z_0| + |T^z_0| + |K^{xx}_{0,1}| + |D^x_{0,0}| + |K^{vv}_{1,0}| \le CL^8_\delta.$$

Finally,

$$\|\nabla_{W_0}\chi_0\|_F + \|\nabla_{W_0}\rho_0\|_F + \|\nabla_{W_0}K^{xx}_{0,1}\|_F + \|\nabla_{W_0}K^{vv}_{1,0}\|_F \le CL^4_\delta,$$

and

$$\|\nabla_{u_0}\chi_0\|_2 + \|\nabla_{m\mathbf{v}_0}\chi_0\|_2 + \|\nabla_{u_0}\rho_0\|_2 + \|\nabla_{m\mathbf{v}_0}\rho_0\|_2 + \|\nabla_{u_0}K^{xx}_{0,1}\|_2 + \|\nabla_{m\mathbf{v}_0}K^{xx}_{0,1}\|_2 + \|\nabla_{u_0}K^{vv}_{1,0}\|_2 + \|\nabla_{m\mathbf{v}_0}K^{vv}_{1,0}\|_2 \le Cm^{-1/2}L^4_\delta.$$

Proof. Mark $a := m\mathbf{v}_0$, $b := u_0$, $M := W_0$, note a and b are standard Gaussian vectors in $\mathbb{R}^m$, M has independent $\mathcal{N}(0, 1/m)$ entries, and a, b, M are independent. Introduce the time-0 building blocks

$$q_0 := b,\ q_1 := a,\ q_2 := Mb,\ q_3 := M^\top a,\ q_4 := MM^\top a,\ q_5 := M^\top Mb,\ q_6 := MM^\top Mb,$$

which gives $z_0 = q_2$, $d_0 = q_3$, $W_0 d_0 = q_4$, $W_0 W_0^\top z_0 = q_6$.

The basic scalars are

$$K^{xx}_{0,0} = \frac{1}{m}\|q_0\|_2^2, \qquad K^{vv}_{0,0} = \frac{1}{m}\|q_1\|_2^2, \qquad G^{xv}_{0,0} = \frac{1}{m}\langle q_2, q_1\rangle,$$

$$g^{(m)}_{\theta_0} = \xi_0 G^{xv}_{0,0}, \qquad \chi_0 = \xi_0 G^{xv}_{0,0} - y(\xi_0), \qquad \rho_0 = \xi_0\chi_0,$$

$$Q^z_0 = \frac{1}{m}\|q_2\|_2^2, \qquad A^{vd}_0 = \frac{1}{m}\langle q_1, q_4\rangle = \frac{1}{m}\|q_3\|_2^2, \qquad H^{z,Wd}_0 = \frac{1}{m}\langle q_2, q_4\rangle,$$

$$B^v_0 = \frac{1}{m}\|q_4\|_2^2, \qquad S^z_0 = \frac{1}{m}\langle q_2, q_6\rangle, \qquad T^z_0 = \frac{1}{m}\langle q_4, q_6\rangle.$$

Since

$$m\mathbf{v}_1 = q_1 - \rho_0 q_2, \qquad u_1 = q_0 - \rho_0 q_3,$$

we also have

$$K^{xx}_{0,1} = K^{xx}_{0,0} - \rho_0 G^{xv}_{0,0} = K^{xx}_{0,0} - \chi_0 g^{(m)}_{\theta_0},$$

$$D^x_{0,0} = -\rho_0 G^{xv}_{0,0} = -\chi_0 g^{(m)}_{\theta_0}, \qquad K^{vv}_{1,0} = K^{vv}_{0,0} - \rho_0 G^{xv}_{0,0} = K^{vv}_{0,0} - \chi_0 g^{(m)}_{\theta_0}.$$

Degree bookkeeping. The degrees of the building blocks in the Gaussian coordinates are

$$\deg q_0 = \deg q_1 = 1, \qquad \deg q_2 = \deg q_3 = 2, \qquad \deg q_4 = \deg q_5 = 3, \qquad \deg q_6 = 4.$$

Consequently every scalar appearing in the lemma is a Gaussian polynomial of degree at most 7.

High-probability size bounds. Let

$$\mathcal{E}_0 := \left\{ \|a\|_2 + \|b\|_2 \le C\sqrt{m}\, L_\delta^{1/2}, \quad \|M\|_{\mathrm{op}} \le CL_\delta^{1/2} \right\}.$$

By standard Gaussian concentration and the standard operator-norm bound for Gaussian matrices with entries of variance $1/m$, C can be chosen so that $\mathbb{P}(\mathcal{E}_0) \ge 1 - \delta/2$. On $\mathcal{E}_0$,

$$\|q_0\|_2 + \|q_1\|_2 \le C\sqrt{m}\, L_\delta^{1/2}, \quad \|q_2\|_2 + \|q_3\|_2 \le C\sqrt{m}\, L_\delta,$$

$$\|q_4\|_2 + \|q_5\|_2 \le C\sqrt{m}\, L_\delta^{3/2}, \quad \|q_6\|_2 \le C\sqrt{m}\, L_\delta^2.$$

Cauchy–Schwarz and boundedness of $|\xi_0| + |y(\xi_0)|$ then give the displayed simultaneous size bound after increasing the logarithmic power to 8.

Gradient bounds for χ_0, ρ_0, $K^{xx}_{0,1}$, and $K^{vv}_{1,0}$. For a perturbation (H, h_u, h_v),

$$DG^{xv}_{0,0}[H, h_u, h_v] = \frac{1}{m}\left(\langle h_v, Mb\rangle + \langle a, Hb\rangle + \langle a, Mh_u\rangle\right).$$

Hence

$$\nabla_{W_0} G^{xv}_{0,0} = \frac{1}{m}ab^\top, \qquad \nabla_{u_0} G^{xv}_{0,0} = \frac{1}{m}M^\top a, \qquad \nabla_{m\mathbf{v}_0} G^{xv}_{0,0} = \frac{1}{m}Mb.$$

Since $g_{\theta_0}^{(m)} = \xi_0 G_{0,0}^{xv}$, $\chi_0 = g_{\theta_0}^{(m)} - y(\xi_0)$, and $\rho_0 = \xi_0 \chi_0$, the same estimates hold for $g_{\theta_0}^{(m)}$, χ_0, and ρ_0 up to deterministic factors bounded by M_T^{data}. In particular, on $\mathcal{E}_0$,

$$\|\nabla_{W_0}\chi_0\|_F + \|\nabla_{W_0}\rho_0\|_F \le CL_\delta,$$

and

$$\|\nabla_{u_0}\chi_0\|_2 + \|\nabla_{m\mathbf{v}_0}\chi_0\|_2 + \|\nabla_{u_0}\rho_0\|_2 + \|\nabla_{m\mathbf{v}_0}\rho_0\|_2 \le Cm^{-1/2}L_\delta.$$

Furthermore,

$$K_{0,1}^{xx} = K_{0,0}^{xx} - \chi_0 g_{\theta_0}^{(m)}, \qquad K_{1,0}^{vv} = K_{0,0}^{vv} - \chi_0 g_{\theta_0}^{(m)}.$$

Differentiating these identities and using $|g_{\theta_0}^{(m)}| + |\chi_0| \le CL_\delta^{3/2}$ on $\mathcal{E}_0$ gives the stated W_0-, u_0-, and $m\mathbf{v}_0$-block gradient bounds.

Variance bounds by Poincaré. The deterministic estimates

$$\mathbb{E}\|q_i\|_2^4 \le Cm^2, \qquad \mathbb{E}\|D_{W_0}q_i\|_{\mathrm{F}\to 2}^4 \le Cm^2, \qquad \mathbb{E}\|D_{u_0}q_i\|_{2\to 2}^4 + \mathbb{E}\|D_{m\mathbf{v}_0}q_i\|_{2\to 2}^4 \le C$$

for $0 \le i \le 6$ follow from the displayed formulas for the q_i's and standard fixed moments of $\|M\|_{\mathrm{op}}$. Therefore, for every scalar of the form $m^{-1}\langle q_i, q_j\rangle$ or $m^{-1}\|q_i\|_2^2$, the weighted Gaussian Poincaré inequality gives variance at most Cm^{-1}. The same bound applies to deterministic affine combinations such as $g_{\theta_0}^{(m)}$, χ_0, and ρ_0.

For $K_{0,1}^{xx}$, $D_{0,0}^{x}$, and $K_{1,0}^{vv}$, use their representation through $\chi_0 g_{\theta_0}^{(m)}$. The product rule, the fixed fourth moments above, and the gradient bounds for $G_{0,0}^{xv}$ imply

$$\mathbb{E}\|\nabla_{W_0}(\chi_0 g_{\theta_0}^{(m)})\|_F^2 \le C, \qquad \mathbb{E}\|\nabla_{u_0}(\chi_0 g_{\theta_0}^{(m)})\|_2^2 + \mathbb{E}\|\nabla_{m\mathbf{v}_0}(\chi_0 g_{\theta_0}^{(m)})\|_2^2 \le Cm^{-1}.$$

Weighted Poincaré again yields variance $O(m^{-1})$ for these endpoint scalars.

Hypercontractive upgrade to high probability. Each centered scalar in the lemma is a Gaussian polynomial of degree at most 7 and has L^2 norm at most $Cm^{-1/2}$. Gaussian hypercontractivity and Markov's inequality with $p = \max\{2, \lceil L_\delta \rceil\}$ give

$$|S - \mathbb{E}S| \le Cm^{-1/2}L_\delta^8$$

with failure probability at most $\delta/(2N_*)$ for each scalar S, where N_* is the fixed number of scalars in the statement. A union bound and intersection with $\mathcal{E}_0$ prove the simultaneous deviation, size, and gradient bounds. □

Lemma F.3. *[Polynomial stability under small perturbations] Fix integers $n, D \ge 1$, and let $P : \mathbb{R}^n \to \mathbb{R}$ be a polynomial of total degree at most D. Let $Z = (Z_1, \dots, Z_n)$ be a random vector, and write $\mu_i := \mathbb{E}Z_i$, $\mu = (\mu_1, \dots, \mu_n)$. Assume that there exist $M_\delta \ge 1$ and $\varepsilon > 0$ such that $|\mu_i| \le M_\delta$, $i \in [n]$, that on an event $\mathcal{E}$,*

$$|Z_i| \le M_\delta, \quad |Z_i - \mu_i| \le \varepsilon, \quad \|Z_i\|_{L^{2D}} \le M_\delta, \quad \|Z_i - \mu_i\|_{L^2} \le \varepsilon, \quad i \in [n].$$

Then there exists a constant $C = C(P, n, D)$, independent of M_δ and ε, such that on $\mathcal{E}$,

$$|P(Z) - \mathbb{E}P(Z)| \le C(1 + M_\delta)^{D-1}\varepsilon.$$

Proof. Throughout the proof, C denotes a positive constant depending only on P, n, D, whose value may change from line to line.

Write

$$P(z) = \sum_{|\alpha|\le D} c_\alpha z^\alpha, \quad z^\alpha = \prod_{j=1}^n z_j^{\alpha_j}.$$

For each i, the partial derivative $\partial_i P$ has degree at most $D-1$. In fact, $\partial_i P(z) = \sum_{\substack{|\alpha|\le D \\ \alpha_i \ge 1}} c_\alpha \alpha_i z^{\alpha - e_i}$.

If $z \in [-M_\delta, M_\delta]^n$, then

$$|z^{\alpha - e_i}| = \prod_{j=1}^n |z_j|^{\alpha_j - \mathbf{1}_{\{j=i\}}} \le M_\delta^{|\alpha|-1} \le (1 + M_\delta)^{D-1}.$$

Therefore,

$$|\partial_i P(z)| \le \sum_{\substack{|\alpha|\le D\\ \alpha_i\ge 1}} |c_\alpha|\alpha_i| z^{\alpha-e_i}| \le C(P,n,D)(1+M_\delta)^{D-1}.$$

On the event $\mathcal{E}$, both Z and μ belong to the box $[-M_\delta, M_\delta]^n$. Since this box is convex, the whole segment $\mu + t(Z-\mu)$, $t \in [0,1]$, also belongs to $[-M_\delta, M_\delta]^n$. By the mean-value formula,

$$P(Z) - P(\mu) = \int_0^1 \nabla P(\mu + t(Z-\mu)) \cdot (Z-\mu)\, dt.$$

Hence, on $\mathcal{E}$,

$$\begin{aligned} |P(Z) - P(\mu)| &\le \int_0^1 \sum_{i=1}^n |\partial_i P(\mu + t(Z-\mu))|\, |Z_i - \mu_i|\, dt \\ &\le C(P,n,D)(1+M_\delta)^{D-1} \sum_{i=1}^n |Z_i - \mu_i|. \end{aligned}$$

Since $|Z_i - \mu_i| \le \varepsilon$ on $\mathcal{E}$, we obtain

$$|P(Z) - P(\mu)| \le C(P,n,D)(1+M_\delta)^{D-1}\varepsilon.$$

It remains to compare $P(\mu)$ with $\mathbb{E}P(Z)$. By linearity, it suffices to treat one monomial z^α, where $|\alpha| \le D$. Define intermediate vectors

$$x^{(0)} = \mu, \quad x^{(i)} = (Z_1, \dots, Z_i, \mu_{i+1}, \dots, \mu_n), \quad 1 \le i \le n,$$

so that $x^{(n)} = Z$. Then

$$Z^\alpha - \mu^\alpha = \sum_{i=1}^n \left[(x^{(i)})^\alpha - (x^{(i-1)})^\alpha \right].$$

Each difference changes only the i-th coordinate, hence

$$(x^{(i)})^\alpha - (x^{(i-1)})^\alpha = (Z_i^{\alpha_i} - \mu_i^{\alpha_i}) \prod_{j<i} Z_j^{\alpha_j} \prod_{j>i} \mu_j^{\alpha_j}.$$

Using $u^k - v^k = (u-v)\sum_{\ell=0}^{k-1} u^{k-1-\ell} v^\ell$, we obtain

$$Z^\alpha - \mu^\alpha = \sum_{i=1}^n (Z_i - \mu_i) Q_{\alpha,i}(Z,\mu),$$

where each $Q_{\alpha,i}$ is a polynomial of total degree at most $D-1$.

Taking expectations and using the triangle inequality gives

$$|\mathbb{E}Z^\alpha - \mu^\alpha| \le \sum_{i=1}^n \mathbb{E}\left[|Z_i - \mu_i|\, |Q_{\alpha,i}(Z,\mu)| \right].$$

By Cauchy–Schwarz,

$$\mathbb{E}\left[|Z_i - \mu_i|\, |Q_{\alpha,i}(Z,\mu)| \right] \le \|Z_i - \mu_i\|_{L^2} \|Q_{\alpha,i}(Z,\mu)\|_{L^2}.$$

Using $\|Z_i - \mu_i\|_{L^2} \le \varepsilon$, it remains to bound $\|Q_{\alpha,i}(Z,\mu)\|_{L^2}$.

Each monomial in $Q_{\alpha,i}$ has the form

$$\prod_{j=1}^n Z_j^{\beta_j} \prod_{j=1}^n \mu_j^{\gamma_j}, \qquad |\beta| + |\gamma| \le D-1.$$

Since $|\mu_j| \le M_\delta$,

$$\left\| \prod_{j=1}^n Z_j^{\beta_j} \prod_{j=1}^n \mu_j^{\gamma_j} \right\|_{L^2} \le M_\delta^{|\gamma|} \left\| \prod_{j=1}^n Z_j^{\beta_j} \right\|_{L^2}.$$

Let $B = |\beta|$. If $B > 0$, Hölder's inequality yields

$$\mathbb{E}\prod_{j=1}^{n}|Z_j|^{2\beta_j} \le \prod_{\beta_j>0}\left(\mathbb{E}|Z_j|^{2B}\right)^{\beta_j/B}.$$

Since $B \le D$ and $\|Z_j\|_{L^{2D}} \le M_\delta$, monotonicity of L^p-norms gives $\|Z_j\|_{L^{2B}} \le M_\delta$. Thus

$$\left\|\prod_{j=1}^{n} Z_j^{\beta_j}\right\|_{L^2} \le M_\delta^B.$$

Therefore,

$$\|Q_{\alpha,i}(Z,\mu)\|_{L^2} \le C(\alpha,n,D)(1+M_\delta)^{D-1}.$$

It follows that

$$|\mathbb{E}Z^\alpha - \mu^\alpha| \le C(\alpha,n,D)(1+M_\delta)^{D-1}\varepsilon.$$

Summing over all monomials yields

$$|\mathbb{E}P(Z) - P(\mu)| \le C(P,n,D)(1+M_\delta)^{D-1}\varepsilon.$$

Finally, on $\mathcal{E}$,

$$|P(Z) - \mathbb{E}P(Z)| \le |P(Z)-P(\mu)| + |P(\mu)-\mathbb{E}P(Z)| \le C(P,n,D)(1+M_\delta)^{D-1}\varepsilon,$$

which completes the proof. □

Lemma F.4 (Base vector and blockwise gradient bounds)**.** *Let $\mathcal{U}$ denote the collection of vectors*

$$\mathcal{U} := \{u_s,\, m\mathbf{v}_s,\, z_s,\, d_s,\, W_0^\top(m\mathbf{v}_s),\, W_0u_s,\, W_0d_s,\, s=0,1\}.$$

There exists a constant C, depending only on M_T^{data} and, in this lemma, on the distinguished datum only through $|\xi_0| + |y(\xi_0)|$, such that for every $\delta \in (0,1/2]$, with probability at least $1-\delta$ and $s = 0, 1$,

$$\max\Big\{\|u_s\|_2, \|m\mathbf{v}_s\|_2, \|z_s\|_2, \|d_s\|_2, \|W_0^\top(m\mathbf{v}_s)\|_2, \|W_0u_s\|_2, \|W_0d_s\|_2\Big\} \le C\sqrt{m}\,L_\delta^{20},$$

Moreover, for every $U \in \mathcal{U}$,

$$\sup_{\|H\|_{\mathrm{F}}=1}\|D_{W_0}U[H]\|_2 \le C\sqrt{m}\,L_\delta^{20}, \qquad \sup_{\|h_u\|_2=1}\|D_{u_0}U[h_u]\|_2 + \sup_{\|h_v\|_2=1}\|D_{m\mathbf{v}_0}U[h_v]\|_2 \le CL_\delta^{20}.$$

Finally, for every $U \in \mathcal{U}$, $\mathbb{E}\|U\|_2^4 \le Cm^2$, $\mathbb{E}\left(\sup_{\|H\|_{\mathrm{F}}=1}\|D_{W_0}U[H]\|_2\right)^4 \le Cm^2$, $\mathbb{E}\left(\sup_{\|h_u\|_2=1}\|D_{u_0}U[h_u]\|_2\right)^4 + \mathbb{E}\left(\sup_{\|h_v\|_2=1}\|D_{m\mathbf{v}_0}U[h_v]\|_2\right)^4 \le C$.

Proof. We treat $(\xi_0, y(\xi_0))$ as deterministic. Write $a := m\mathbf{v}_0$, $b := u_0$, $M := W_0$, and use the time-0 blocks $q_0 = b$, $q_1 = a$, $q_2 = Mb$, $q_3 = M^\top a$, $q_4 = MM^\top a$, $q_5 = M^\top Mb$, and $q_6 = MM^\top Mb$. Apply Lemma F.2 with failure probability $\delta/2$, and intersect with the Gaussian norm event

$$\|a\|_2 + \|b\|_2 \le C\sqrt{m}\,L_\delta^{1/2}, \qquad \|M\|_{\mathrm{op}} \le CL_\delta^{1/2}.$$

On this intersection, which has probability at least $1-\delta$, the blocks satisfy $\|q_i\|_2 \le C\sqrt{m}\,L_\delta^2$, $0 \le i \le 6$. And Lemma F.2 gives $|\rho_0| + |K_{0,1}^{xx}| + |K_{1,0}^{vv}| \le CL_\delta^8$ together with the corresponding scalar gradient bounds for ρ_0, $K_{0,1}^{xx}$, and $K_{1,0}^{vv}$.

At time 0, there is $u_0 = q_0$, $m\mathbf{v}_0 = q_1$, $z_0 = q_2$, $d_0 = q_3$, $W_0^\top(m\mathbf{v}_0) = q_3$, $W_0u_0 = q_2$, $W_0d_0 = q_4$. For time 1, the recursions and Lemma F.1 give

$$m\mathbf{v}_1 = q_1 - \rho_0 q_2, \quad u_1 = q_0 - \rho_0 q_3, \quad z_1 = q_2 - \rho_0 q_4 - \rho_0 q_1 K_{0,1}^{xx}, \quad d_1 = q_3 - \rho_0 q_5 - \rho_0 q_0 K_{1,0}^{vv},$$

and

$$W_0^\top(m\mathbf{v}_1) = q_3 - \rho_0 q_5, \quad W_0u_1 = q_2 - \rho_0 q_4, \quad W_0d_1 = q_4 - \rho_0 q_6 - \rho_0 q_2 K_{1,0}^{vv}.$$

These finite sums immediately yield the displayed norm bounds.

For derivatives, differentiate the displayed formulas. The deterministic derivatives of the q_i's satisfy, on the Gaussian norm event,

$$\sup_{\|H\|_{\mathrm{F}}=1} \|D_{W_0} q_i[H]\|_2 \le C\sqrt{m}\, L_\delta^2, \qquad \sup_{\|h_u\|_2=1} \|D_{u_0} q_i[h_u]\|_2 + \sup_{\|h_v\|_2=1} \|D_{m\mathbf{v}_0} q_i[h_v]\|_2 \le C L_\delta^2.$$

Each product-rule term contains either a derivative of a q_i, or a scalar derivative from Lemma F.2 multiplied by a vector of size $O(\sqrt{m}L_\delta^2)$. The u_0- and $m\mathbf{v}_0$-block scalar derivatives have the extra $m^{-1/2}$ factor, while the W_0-block scalar derivatives are $O(L_\delta^4)$. Therefore every W_0-block derivative is at most $C\sqrt{m}L_\delta^{20}$ and every u_0- or $m\mathbf{v}_0$-block derivative is at most CL_δ^{20}.

The high-probability estimates above are uniform in the tail parameter. Applying Lemma B.3 to the normalized quantities $m^{-1/2}\|U\|_2$, $m^{-1/2}\|D_{W_0}U\|_{\mathrm{F}\to 2}$, $\|D_{u_0}U\|_{2\to 2}$, and $\|D_{m\mathbf{v}_0}U\|_{2\to 2}$ gives the displayed fourth-moment bounds. □

Lemma F.5 (Higher-order scalar observables at time 1)**.** *There exists a constant C, depending only on M_T^{data} and in this lemma only on $|\xi_0| + |y(\xi_0)|$, such that for every $\delta \in (0, 1/2]$, with probability at least $1-\delta$,*

$$|K_{1,1}^{vv} - \mathbb{E}K_{1,1}^{vv}| + |Q_1^z - \mathbb{E}Q_1^z| + |G_{1,0}^{xv} - \mathbb{E}G_{1,0}^{xv}| + |A_1^{vd} - \mathbb{E}A_1^{vd}| + |H_1^{z,Wd} - \mathbb{E}H_1^{z,Wd}| \le Cm^{-1/2}L_\delta^{40}.$$

Proof. Define

$$Z_0 := \Big(K_{0,0}^{xx}, K_{0,0}^{vv}, G_{0,0}^{xv}, g_{\theta_0}^{(m)}, \chi_0, \rho_0, Q_0^z, A_0^{vd}, H_0^{z,Wd}, B_0^v, S_0^z, T_0^z, K_{0,1}^{xx}, K_{1,0}^{vv}\Big).$$

By Lemma F.2, with probability at least $1-\delta$,

$$\max_i |(Z_0)_i - \mathbb{E}(Z_0)_i| \le C_0 m^{-1/2} L_\delta^8, \qquad \max_i \big(|(Z_0)_i| + |\mathbb{E}(Z_0)_i|\big) \le C_0 L_\delta^8.$$

Moreover, fixed moments and centered L^2 bounds for the coordinates of Z_0 follow from the same Poincaré/hypercontractive argument used in Lemma F.2. Thus the hypotheses of Lemma F.3 hold with $M_\delta = C_0 L_\delta^8$, $\varepsilon = C_0 m^{-1/2} L_\delta^8$.

Each time-1 scalar is an explicit polynomial in Z_0. With $G_0 := G_{0,0}^{xv}$ and $\rho_0 = \xi_0\chi_0$, the recursions give

$$K_{1,1}^{vv} = K_{0,0}^{vv} - 2\rho_0 G_0 + \rho_0^2 Q_0^z, \quad G_{1,0}^{xv} = G_0 - \rho_0 A_0^{vd} - \rho_0 K_{0,1}^{xx} K_{0,0}^{vv},$$

$$Q_1^z = Q_0^z - 2\rho_0 H_0^{z,Wd} - 2\rho_0 K_{0,1}^{xx} G_0 + \rho_0^2 B_0^v + 2\rho_0^2 K_{0,1}^{xx} A_0^{vd} + \rho_0^2 (K_{0,1}^{xx})^2 K_{0,0}^{vv},$$

$$A_1^{vd} = A_0^{vd} - 2\rho_0 H_0^{z,Wd} - \rho_0 K_{1,0}^{vv} G_0 + \rho_0^2 S_0^z + \rho_0^2 K_{1,0}^{vv} Q_0^z,$$

and

$$H_1^{z,Wd} = H_0^{z,Wd} - \rho_0 S_0^z - \rho_0 K_{1,0}^{vv} Q_0^z - \rho_0 B_0^v + \rho_0^2 T_0^z + \rho_0^2 K_{1,0}^{vv} H_0^{z,Wd} - \rho_0 K_{0,1}^{xx} A_0^{vd} + \rho_0^2 K_{0,1}^{xx} H_0^{z,Wd} + \rho_0^2 K_{0,1}^{xx} K_{1,0}^{vv} G_0.$$

Each expression is a fixed polynomial of total degree at most 5 in the coordinates of Z_0. Applying Lemma F.3 and union bounding over the five observables proves the claim. □

Lemma F.6 (Degree growth up to time T)**.** *Fix $T < \infty$, and define $D_T := 2 \cdot 7^{T+2}$, $q_T := 2D_T = 4 \cdot 7^{T+2}$. Then every vector among $\{u_t,\, m\mathbf{v}_t,\, z_t,\, d_t,\, W_t\}$ where $0 \le t \le T+2$, is a polynomial in the Gaussian coordinates of $\mathcal{G}_0$ of degree at most D_T.*

Moreover, every vector among $\{W_0^\top(m\mathbf{v}_t),\, W_0 u_t,\, W_0 d_t\}$ $0 \le t \le T+1$, is a polynomial in $\mathcal{G}_0$ of degree at most D_T. Consequently, every scalar observable appearing in Theorem F.8 up to time T is a polynomial in $\mathcal{G}_0$ of degree at most q_T.

The restriction of the W_0-derived vectors $W_0^\top(m\mathbf{v}_t)$, $W_0 u_t$, $W_0 d_t$ to $0 \le t \le T+1$ is intentional. These are the only such degree bounds needed for scalar concentration up to time T. The final high-probability vector bounds at time $T+2$ in Theorem F.8 are obtained separately by the terminal bootstrap in the theorem proof, not by applying this degree lemma at level $T+2$.

Proof. Let D_t be the maximum polynomial degree among $\{u_t,\, m\mathbf{v}_t,\, z_t,\, d_t,\, W_t\}$. At time 0,

$$\deg(u_0) = \deg(m\mathbf{v}_0) = \deg(W_0) = 1, \quad \deg(z_0) \le 2, \quad \deg(d_0) \le 2,$$

so $D_0 \le 2$.

Since $g_{\theta_t}^{(m)} = \xi_t \frac{1}{m}\langle m\mathbf{v}_t, z_t\rangle$, $\chi_t = g_{\theta_t}^{(m)} - y(\xi_t)$, $\rho_t = \xi_t\chi_t$, we have $\deg(\chi_t) \le 2D_t$, $\deg(\rho_t) \le 2D_t$.

The recursions $m\mathbf{v}_{t+1} = m\mathbf{v}_t - \rho_t z_t$, $u_{t+1} = u_t - \rho_t d_t$ therefore imply $\deg(m\mathbf{v}_{t+1}) \le 3D_t$, $\deg(u_{t+1}) \le 3D_t$.

Likewise, $W_{t+1} = W_t - \frac{\rho_t}{m}(m\mathbf{v}_t)u_t^\top$ gives $\deg(W_{t+1}) \le 4D_t$. Hence

$$\deg(z_{t+1}) \le \deg(W_{t+1}) + \deg(u_{t+1}) \le 7D_t,$$

and similarly $\deg(d_{t+1}) \le 7D_t$. Thus $D_{t+1} \le 7D_t$. Since $D_0 \le 2$, it follows that $D_t \le 2 \cdot 7^t \le 2 \cdot 7^{T+2} = D_T$, $0 \le t \le T+2$.

For the W_0-derived vectors, if $0 \le s \le T+1$, then

$$\deg(W_0 u_s) \le 1 + \deg(u_s) \le 1 + 2 \cdot 7^{T+1} \le 2 \cdot 7^{T+2} = D_T,$$

due to $1 + 2 \cdot 7^{T+1} \le 14 \cdot 7^{T+1} = 2 \cdot 7^{T+2}$. The same estimation gives $\deg(W_0^\top(m\mathbf{v}_s)) \le D_T$, $\deg(W_0 d_s) \le D_T$.

Finally, every scalar observable in the theorem is either a normalized inner product of two vectors from these families or a fixed finite linear combination of such terms. Hence its degree is at most $2D_T = q_T$. □

Lemma F.7 (Scalar concentration from blockwise vector moments)**.** *Fix $T < \infty$, and define $q_T := 4 \cdot 7^{T+2}$, $\alpha_T := \frac{q_T}{2} + 1 = 2 \cdot 7^{T+2} + 1$.*

Let $\mathcal{V}_{T+1}$ denote the family of vectors

$$\mathcal{V}_{T+1} := \{u_s,\, m\mathbf{v}_s,\, z_s,\, d_s,\, W_0^\top(m\mathbf{v}_s),\, W_0 u_s,\, W_0 d_s,\, 0 \le s \le T+1\}.$$

For vector-valued maps U, write $\|D_{W_0}U\|_{\mathrm{F}\to 2} := \sup_{\|H\|_{\mathrm{F}}=1} \|D_{W_0}U[H]\|_2$ and

$$\|D_{u_0}U\|_{2\to 2} := \sup_{\|h_u\|_2=1} \|D_{u_0}U[h_u]\|_2, \qquad \|D_{m\mathbf{v}_0}U\|_{2\to 2} := \sup_{\|h_v\|_2=1} \|D_{m\mathbf{v}_0}U[h_v]\|_2.$$

Assume that there exists $B_T \ge 1$ such that every $U \in \mathcal{V}_{T+1}$ satisfies

$$\mathbb{E}\|U\|_2^4 \le B_T m^2, \quad \mathbb{E}\|D_{W_0}U\|_{\mathrm{F}\to 2}^4 \le B_T m^2, \quad \mathbb{E}\|D_{u_0}U\|_{2\to 2}^4 + \mathbb{E}\|D_{m\mathbf{v}_0}U\|_{2\to 2}^4 \le B_T.$$

Then there exists a constant C_T, depending only on T, B_T, and M_T^{data}, such that for every $\delta \in (0, 1/2]$, with probability at least $1-\delta$, every scalar observable S appearing in Theorem F.8 up to time T satisfies $|S - \mathbb{E}S| \le C_T m^{-1/2} L_\delta^{\alpha_T}$. Moreover $|\mathbb{E}S| \le C_T$, the same conclusion holds in one-observable form for any single normalized inner product $S = \frac{1}{m}\langle U, V\rangle$ provided U, V satisfy the displayed fourth-moment assumptions and $S - \mathbb{E}S$ is a Gaussian polynomial of degree at most q_T.

Proof. We treat the data sequence $(\xi_t, y(\xi_t))_{0\le t\le T+1}$ as deterministic. Equivalently, if the data are random but independent of the Gaussian initialization, all probabilities and expectations below are understood conditionally on the data. The constants are uniform whenever the data are bounded by M_T^{data}.

We first prove the one-observable estimate. Let $S = \frac{1}{m}\langle U, V\rangle$, where U, V satisfy the displayed fourth-moment assumptions and where $S - \mathbb{E}S$ is a Gaussian polynomial of degree at most q_T.

For scalar-valued maps F, write $\|D_{W_0}F\|_{\mathrm{F}\to\mathbb{R}} := \sup_{\|H\|_{\mathrm{F}}=1} |D_{W_0}F[H]|$ and

$$\|D_{u_0}F\|_{2\to\mathbb{R}} := \sup_{\|h_u\|_2=1} |D_{u_0}F[h_u]|, \qquad \|D_{m\mathbf{v}_0}F\|_{2\to\mathbb{R}} := \sup_{\|h_v\|_2=1} |D_{m\mathbf{v}_0}F[h_v]|.$$

Differentiating S blockwisely gives

$$D_{W_0}S[H] = \frac{1}{m}\left(\langle D_{W_0}U[H], V\rangle + \langle U, D_{W_0}V[H]\rangle\right).$$

Therefore

$$\|D_{W_0}S\|_{\mathrm{F}\to\mathbb{R}} \le \frac{1}{m}\left(\|D_{W_0}U\|_{\mathrm{F}\to 2}\|V\|_2 + \|U\|_2\|D_{W_0}V\|_{\mathrm{F}\to 2}\right).$$

Similarly,

$$\|D_{u_0}S\|_{2\to\mathbb{R}} \le \frac{1}{m}\left(\|D_{u_0}U\|_{2\to 2}\|V\|_2 + \|U\|_2\|D_{u_0}V\|_{2\to 2}\right),$$

and

$$\|D_{m\mathbf{v}_0}S\|_{2\to\mathbb{R}} \le \frac{1}{m}\left(\|D_{m\mathbf{v}_0}U\|_{2\to 2}\|V\|_2 + \|U\|_2\|D_{m\mathbf{v}_0}V\|_{2\to 2}\right).$$

By Cauchy–Schwarz and the fourth-moment assumptions,

$$\mathbb{E}\|D_{W_0}S\|^2_{\mathrm{F}\to\mathbb{R}} \le \frac{2}{m^2}\Big[\left(\mathbb{E}\|D_{W_0}U\|^4_{\mathrm{F}\to 2}\right)^{1/2}\left(\mathbb{E}\|V\|_2^4\right)^{1/2} + \left(\mathbb{E}\|U\|_2^4\right)^{1/2}\left(\mathbb{E}\|D_{W_0}V\|^4_{\mathrm{F}\to 2}\right)^{1/2}\Big] \le CB_T.$$

Likewise,

$$\mathbb{E}\|D_{u_0}S\|^2_{2\to\mathbb{R}} \le \frac{2}{m^2}\Big[\left(\mathbb{E}\|D_{u_0}U\|^4_{2\to 2}\right)^{1/2}\left(\mathbb{E}\|V\|_2^4\right)^{1/2} + \left(\mathbb{E}\|U\|_2^4\right)^{1/2}\left(\mathbb{E}\|D_{u_0}V\|^4_{2\to 2}\right)^{1/2}\Big] \le CB_Tm^{-1},$$

and the same argument gives $\mathbb{E}\|D_{m\mathbf{v}_0}S\|^2_{2\to\mathbb{R}} \le CB_Tm^{-1}$.

Since the entries of W_0 have variance $1/m$, while the coordinates of u_0 and $m\mathbf{v}_0$ have variance 1, the weighted Gaussian Poincaré inequality gives

$$\mathrm{Var}(S) \le \frac{1}{m}\mathbb{E}\|D_{W_0}S\|^2_{\mathrm{F}\to\mathbb{R}} + \mathbb{E}\|D_{u_0}S\|^2_{2\to\mathbb{R}} + \mathbb{E}\|D_{m\mathbf{v}_0}S\|^2_{2\to\mathbb{R}}.$$

Hence $\mathrm{Var}(S) \le CB_Tm^{-1}$, and therefore $\|S-\mathbb{E}S\|_{L^2} \le CB_T^{1/2}m^{-1/2}$.

By assumption, $S-\mathbb{E}S$ is a centered Gaussian polynomial of degree at most q_T. Scalar Gaussian hypercontractivity yields, for every $p \ge 2$,

$$\|S-\mathbb{E}S\|_{L^p} \le C_Tp^{q_T/2}\|S-\mathbb{E}S\|_{L^2} \le C_TB_T^{1/2}p^{q_T/2}m^{-1/2}.$$

Let $p := \max\{2, \lceil L_\eta\rceil\}$, where $\eta \in (0, 1/2]$. By Markov's inequality, for a sufficiently large constant A_T,

$$\mathbb{P}\left(|S-\mathbb{E}S| > A_TB_T^{1/2}m^{-1/2}L_\eta^{q_T/2+1}\right) \le \left(\frac{C_TB_T^{1/2}p^{q_T/2}m^{-1/2}}{A_TB_T^{1/2}m^{-1/2}L_\eta^{q_T/2+1}}\right)^p \le \eta.$$

Absorbing $B_T^{1/2}$ into the constant, and recalling that $\alpha_T = \frac{q_T}{2}+1$, we obtain the one-observable bound $|S-\mathbb{E}S| \le C_Tm^{-1/2}L_\eta^{\alpha_T}$ with probability at least $1-\eta$.

Next,

$$|\mathbb{E}S| \le \frac{1}{m}\mathbb{E}\big[\|U\|_2\|V\|_2\big] \le \frac{1}{m}\left(\mathbb{E}\|U\|_2^4\right)^{1/4}\left(\mathbb{E}\|V\|_2^4\right)^{1/4} \le B_T^{1/2}.$$

Thus, after increasing C_T, $|\mathbb{E}S| \le C_T$.

We now pass from one normalized inner product to the full finite scalar family in Theorem F.8. By Lemma F.6, every scalar observable up to time T is a Gaussian polynomial of degree at most q_T. Moreover, by the definition of the scalar family in the theorem, each such observable is either a normalized inner product $\frac{1}{m}\langle U, V\rangle$, $U, V \in \mathcal{V}_{T+1}$, or a fixed finite linear combination of such normalized inner products, with the number of terms and coefficients bounded in terms of T and M_T^{data}.

Let N_T be the total number of normalized-inner-product summands needed to represent all scalar observables appearing up to time T. Apply the one-observable estimate to each summand with failure probability $\eta := \frac{\delta}{N_T}$. Since N_T depends only on T, $L_{\delta/N_T} \le C_TL_\delta$. A union bound gives, with probability at least $1-\delta$, the bound

$$\left|\frac{1}{m}\langle U, V\rangle - \mathbb{E}\frac{1}{m}\langle U, V\rangle\right| \le C_Tm^{-1/2}L_\delta^{\alpha_T}$$

for every normalized-inner-product summand in the finite family. Taking the corresponding finite linear combinations and absorbing the fixed number of terms and coefficient bounds into C_T yields, simultaneously for every scalar observable S appearing in the theorem up to time T, $|S - \mathbb{E}S| \le C_T m^{-1/2} L_\delta^{\alpha_T}$. The same finite-combination argument also gives $|\mathbb{E}S| \le C_T$. This completes the proof. □

Theorem F.8 (Simultaneous concentration around finite-width expectations for GD)**.** *Fix $T < \infty$. Define $\alpha_t := 2 \cdot 7^{t+2} + 1$, $\beta_t := 200 \cdot 8^{t+2}$, Then there exists a constant $C_T \ge 1$, depending only on T and M_T^{data}, such that for every $\delta \in (0, 1/2]$, with probability at least $1 - \delta$, the following holds.*

For each $t \ge 0$, define

$$\varepsilon_t(\delta) := C_T m^{-1/2} L_\delta^{\alpha_t}, \quad R_t(\delta) := C_T \sqrt{m}\, L_\delta^{\beta_t}, \quad r_t(\delta) := C_T L_\delta^{\beta_t}.$$

Then, for every $0 \le s \le T + 2$,

$$\|u_s\|_2 + \|m\mathbf{v}_s\|_2 + \|z_s\|_2 + \|d_s\|_2 + \|W_0^\top (m\mathbf{v}_s)\|_2 + \|W_0 u_s\|_2 + \|W_0 d_s\|_2 \le R_s(\delta),$$

and, for every such vector U,

$$\|\nabla_{W_0} U\|_F \le R_s(\delta), \qquad \|\nabla_{u_0} U\|_2 + \|\nabla_{m\mathbf{v}_0} U\|_2 \le r_s(\delta).$$

Moreover, for every $0 \le t \le T$, every $0 \le r \le t$, and every observable scalar S

$$S \in \Big\{ g_{\theta_t}^{(m)},\ \chi_t,\ \rho_t,\ J_t^x,\ D_{r,t}^x,\ A_t^{vd},\ H_t^{z,Wd},\ K_{r,t+1}^{xx},\ K_{t,r}^{vv},\ G_{t,r}^{xv},\ Q_t^z \Big\},$$

we have

$$|S - \mu_S^{(m)}| \le \varepsilon_t(\delta), \quad |S| + |\mu_S^{(m)}| \le C_T L_\delta^{\alpha_t}.$$

Proof. Fix $T < \infty$. All constants below may change from line to line and depend only on T and M_T^{data}. The induction below is used in a uniform tail sense. More precisely, for each level t we prove that, for every $\eta \in (0, 1/2]$, there is an event $\mathcal{E}_t(\eta)$ of probability at least $1 - \eta$ on which the estimates defining $\mathcal{P}_t(\eta)$ below hold with L_η in place of L_δ. The constants C_t are independent of m and η. This uniformity is what allows Lemma B.3 to be applied inside the induction. In the displayed argument we write δ instead of the arbitrary tail parameter η.

Every auxiliary high-probability statement is invoked with failure probability $\delta_* := \delta / N_T$, where N_T is a sufficiently large deterministic constant depending only on T. Since $L_{\delta_*} = 1 + \log(2N_T/\delta) \le C_T L_\delta$, all logarithmic losses caused by this replacement may be absorbed into L_δ after enlarging C_T. We also include in the auxiliary event the standard Gaussian operator-norm event $\|W_0\|_{\mathrm{op}} \le C_T L_\delta^{1/2}$. We work on the intersection of all such auxiliary events and suppress this bookkeeping below.

We record the deterministic exponent relations used repeatedly. Since $\alpha_t = 2 \cdot 7^{t+2} + 1, \beta_t = 200 \cdot 8^{t+2}$, we have $\alpha_t \le \beta_t$, $3\beta_t + 1 \le \beta_{t+1}$, $4\beta_t + 1 \le \beta_{t+1}$ and

$$\alpha_{t+1} + \beta_t + 1 \le \beta_{t+1},\ \alpha_{t+1} + 2\beta_t + 1 \le \beta_{t+1},\ 2\alpha_{t+1} + 3\beta_t + 1 \le \beta_{t+1},\ \alpha_{t+1} + 5\beta_t + 1 \le \beta_{t+1}.$$

The additive $+1$ is included to absorb all factors $L_\delta^{1/2}$ coming from $\|W_0\|_{\mathrm{op}}$.

Inside the induction we repeatedly apply Lemma B.3 to the normalized quantities $m^{-1/2}\|U\|_2$, $m^{-1/2}\|\nabla_{W_0} U\|_F$, $\|\nabla_{u_0} U\|_2$, $\|\nabla_{m\mathbf{v}_0} U\|_2$, for vector nodes U, and to the scalar observables S themselves. This makes explicit which random variables naturally live on the $\sqrt{m}$ scale and which live on the 1 scale in the weighted Gaussian Poincaré argument.

For each $t \ge 0$, let $\mathcal{S}_t$ be the scalar family

$$\mathcal{S}_t := \Big\{ g_{\theta_s}^{(m)},\ \chi_s,\ \rho_s,\ J_s^x,\ A_s^{vd},\ Q_s^z,\ H_s^{z,Wd},\ K_{r,s+1}^{xx},\ K_{s,r}^{vv},\ G_{s,r}^{xv},\ D_{r,s}^x : 0 \le r \le s \le t \Big\}.$$

We prove by induction on t the following uniform statement $\mathcal{P}_t(\delta)$.

Note there is an event of probability at least $1 - \delta$ on which the following four assertions hold.

1. Every scalar $S \in \mathcal{S}_t$ satisfies

$$|S - \mathbb{E}S| \le C_t m^{-1/2} L_\delta^{\alpha_t}, \quad |S| + |\mathbb{E}S| \le C_t L_\delta^{\alpha_t}.$$

2. Every vector $\{u_s,\, m\mathbf{v}_s,\, z_s,\, d_s,\, W_0^\top(m\mathbf{v}_s),\, W_0u_s,\, W_0d_s\}$, $0 \le s \le t+1$, satisfies $\|U\|_2 \le C_t\sqrt{m}\,L_\delta^{\beta_t}$.
3. The same vectors satisfy
$$\|\nabla_{W_0}U\|_F \le C_t\sqrt{m}\,L_\delta^{\beta_t}, \qquad \|\nabla_{u_0}U\|_2 + \|\nabla_{m\mathbf{v}_0}U\|_2 \le C_tL_\delta^{\beta_t}.$$
4. For every $0 \le s \le t$, $\|\nabla_{W_0}\chi_s\|_F + \|\nabla_{W_0}\rho_s\|_F \le C_tL_\delta^{2\beta_t}$ and
$$\|\nabla_{u_0}\chi_s\|_2 + \|\nabla_{m\mathbf{v}_0}\chi_s\|_2 + \|\nabla_{u_0}\rho_s\|_2 + \|\nabla_{m\mathbf{v}_0}\rho_s\|_2 \le C_tm^{-1/2}L_\delta^{2\beta_t}.$$

Base case. Lemma F.2 gives item 1 for all time-0 scalars, including
$$\{J_0^x := G_{0,0}^{xv},\; g_{\theta_0}^{(m)},\; \chi_0,\; \rho_0,\; A_0^{vd},\; Q_0^z,\; H_0^{z,Wd},\; K_{0,1}^{xx},\; K_{0,0}^{vv},\; G_{0,0}^{xv},\; D_{0,0}^x\}.$$
Lemma F.4 gives items 2 and 3 at times 0 and 1. Item 4 for $s=0$ is again part of Lemma F.2. Since $\alpha_0 = 99,\ \beta_0 = 200\cdot 8^2$, the powers 8, 20, and 40 appearing in the base lemmas are absorbed after enlarging C_0. Thus $\mathcal{P}_0(\delta)$ holds. Since δ was arbitrary, the base case holds uniformly in δ.

Inductive. Assume that $\mathcal{P}_t(\eta)$ has been proved for every $\eta \in (0, 1/2]$, for some $0 \le t < T$. We prove $\mathcal{P}_{t+1}(\delta)$ for an arbitrary $\delta \in (0, 1/2]$.

Step 1: Concentrate $g_{\theta_{t+1}}^{(m)}$ and χ_{t+1}. The uniform induction hypothesis implies that every vector already covered by $\mathcal{P}_t$ has a tail bound of the form $Z \le C_tL_\eta^{\beta_t}$ with probability at least $1-\eta, 0 < \eta \le 1/2$ after normalizing by $\sqrt{m}$ for vector norms and W_0-block gradients, and by 1 for the u_0- and $m\mathbf{v}_0$-block gradients. Lemma B.3 therefore yields
$$\mathbb{E}\|U\|_2^4 \le C_tm^2, \quad \mathbb{E}\|\nabla_{W_0}U\|_F^4 \le C_tm^2, \quad \mathbb{E}\|\nabla_{u_0}U\|_2^4 + \mathbb{E}\|\nabla_{m\mathbf{v}_0}U\|_2^4 \le C_t$$
for every vector U already covered by $\mathcal{P}_t$. In particular, these estimates hold for $U = m\mathbf{v}_{t+1}$, $U = z_{t+1}$. By Lemma F.6, the centered scalar
$$g_{\theta_{t+1}}^{(m)} - \mathbb{E}g_{\theta_{t+1}}^{(m)} = \xi_{t+1}\left(\frac{1}{m}\langle m\mathbf{v}_{t+1}, z_{t+1}\rangle - \mathbb{E}\left[\frac{1}{m}\langle m\mathbf{v}_{t+1}, z_{t+1}\rangle\right]\right)$$
is a Gaussian polynomial of degree bounded only in terms of t. The one-observable form of Lemma F.7 gives
$$|g_{\theta_{t+1}}^{(m)} - \mathbb{E}g_{\theta_{t+1}}^{(m)}| + |\chi_{t+1} - \mathbb{E}\chi_{t+1}| + |\rho_{t+1} - \mathbb{E}\rho_{t+1}| \le C_tm^{-1/2}L_\delta^{\alpha_{t+1}}.$$
Moreover by
$$|\mathbb{E}g_{\theta_{t+1}}^{(m)}| \le \frac{1}{m}(\mathbb{E}\|m\mathbf{v}_{t+1}\|_2^4)^{1/4}(\mathbb{E}\|z_{t+1}\|_2^4)^{1/4} \le C_t,$$
there is $|g_{\theta_{t+1}}^{(m)}| + |\chi_{t+1}| + |\rho_{t+1}| \le C_tL_\delta^{\alpha_{t+1}}$.

For every block $B \in \{W_0, u_0, m\mathbf{v}_0\}$, there is
$$D_Bg_{\theta_{t+1}}^{(m)} = \frac{\xi_{t+1}}{m}\Big(\langle D_B(m\mathbf{v}_{t+1}), z_{t+1}\rangle + \langle m\mathbf{v}_{t+1}, D_Bz_{t+1}\rangle\Big).$$
Items 2 and 3 of $\mathcal{P}_t(\delta)$ gives $\|\nabla_{W_0}\chi_{t+1}\|_F + \|\nabla_{W_0}\rho_{t+1}\|_F \le C_tL_\delta^{2\beta_t}$ and
$$\|\nabla_{u_0}\chi_{t+1}\|_2 + \|\nabla_{m\mathbf{v}_0}\chi_{t+1}\|_2 + \|\nabla_{u_0}\rho_{t+1}\|_2 + \|\nabla_{m\mathbf{v}_0}\rho_{t+1}\|_2 \le C_tm^{-1/2}L_\delta^{2\beta_t}.$$
Since $2\beta_t \le 2\beta_{t+1}$, this gives the new χ- and ρ-gradient estimates required for item 4 of $\mathcal{P}_{t+1}(\delta)$.

Step 2: Vector and gradient bounds to time $t+2$. Set $\lambda_t := \alpha_{t+1} + \beta_t$, $\mu_t := \alpha_{t+1} + 2\beta_t$, $\nu_t := \max\{4\beta_t, \alpha_{t+1} + 2\beta_t\}$. We keep these sharper intermediate exponents until the last line of each estimate. From $m\mathbf{v}_{t+2} = m\mathbf{v}_{t+1} - \rho_{t+1}z_{t+1}$, $u_{t+2} = u_{t+1} - \rho_{t+1}d_{t+1}$, items 2 and 3 of $\mathcal{P}_t(\delta)$, and Step 1, we get
$$\|m\mathbf{v}_{t+2}\|_2 + \|u_{t+2}\|_2 \le C_t\sqrt{m}\Big(L_\delta^{\beta_t} + L_\delta^{\alpha_{t+1}+\beta_t}\Big) \le C_t\sqrt{m}\,L_\delta^{\lambda_t}.$$

Differentiating the same recursions gives, for every block B,

$$D_B(m\mathbf{v}_{t+2}) = D_B(m\mathbf{v}_{t+1}) - (D_B\rho_{t+1})z_{t+1} - \rho_{t+1}D_B z_{t+1},$$
$$D_B u_{t+2} = D_B u_{t+1} - (D_B\rho_{t+1})d_{t+1} - \rho_{t+1}D_B d_{t+1}.$$

Consequently,

$$\|\nabla_{W_0} m\mathbf{v}_{t+2}\|_F + \|\nabla_{W_0} u_{t+2}\|_F \le C_t\sqrt{m}\Big(L_\delta^{\beta_t} + L_\delta^{3\beta_t} + L_\delta^{\lambda_t}\Big) \le C_t\sqrt{m}\, L_\delta^{\max\{3\beta_t,\lambda_t\}},$$

and similarly

$$\|\nabla_{u_0} m\mathbf{v}_{t+2}\|_2 + \|\nabla_{m\mathbf{v}_0} m\mathbf{v}_{t+2}\|_2 + \|\nabla_{u_0} u_{t+2}\|_2 + \|\nabla_{m\mathbf{v}_0} u_{t+2}\|_2 \le C_t L_\delta^{\max\{3\beta_t,\lambda_t\}}.$$

Lemma F.1 gives $W_{t+2} = W_0 - \sum_{r=0}^{t+1} \frac{\rho_r}{m}(m\mathbf{v}_r)u_r^\top$.

Using Step 1 for $r = t+1$ and item 1 of $\mathcal{P}_t(\delta)$ for $r \le t$, together with item 2 of $\mathcal{P}_t(\delta)$, yields

$$\|W_{t+2}\|_{\mathrm{op}} \le \|W_0\|_{\mathrm{op}} + \sum_{r=0}^{t+1} \frac{|\rho_r|}{m}\|m\mathbf{v}_r\|_2\|u_r\|_2 \le C_t L_\delta^{\mu_t}.$$

Differentiating the unrolled formula gives

$$D_B W_{t+2} = \mathbf{1}_{\{B=W_0\}}H - \sum_{r=0}^{t+1}\frac{1}{m}\Big((D_B\rho_r)(m\mathbf{v}_r)u_r^\top + \rho_r D_B(m\mathbf{v}_r)u_r^\top + \rho_r(m\mathbf{v}_r)(D_B u_r)^\top\Big).$$

For the W_0-block, item 4 of $\mathcal{P}_t(\delta)$ and Step 1 give

$$\|\nabla_{W_0} W_{t+2}\|_{F\to F} \le C_t\Big(1 + L_\delta^{4\beta_t} + L_\delta^{\alpha_{t+1}+2\beta_t}\Big) \le C_t L_\delta^{\nu_t}.$$

For the u_0- and $m\mathbf{v}_0$-blocks,

$$\|\nabla_{u_0} W_{t+2}\|_{2\to F} + \|\nabla_{m\mathbf{v}_0} W_{t+2}\|_{2\to F} \le C_t m^{-1/2}\Big(L_\delta^{4\beta_t} + L_\delta^{\alpha_{t+1}+2\beta_t}\Big) \le C_t m^{-1/2} L_\delta^{\nu_t}.$$

Since $z_{t+2} = W_{t+2}u_{t+2}$, $d_{t+2} = W_{t+2}^\top(m\mathbf{v}_{t+2})$, the sharper bounds above give

$$\|z_{t+2}\|_2 + \|d_{t+2}\|_2 \le \|W_{t+2}\|_{\mathrm{op}}\big(\|u_{t+2}\|_2 + \|m\mathbf{v}_{t+2}\|_2\big) \le C_t\sqrt{m}\, L_\delta^{\mu_t+\lambda_t}.$$

Here $\mu_t + \lambda_t = 2\alpha_{t+1} + 3\beta_t \le \beta_{t+1}$, so $\|z_{t+2}\|_2 + \|d_{t+2}\|_2 \le C_{t+1}\sqrt{m}\, L_\delta^{\beta_{t+1}}$.

Differentiating gives $D_B z_{t+2} = (D_B W_{t+2})u_{t+2} + W_{t+2}D_B u_{t+2}$, $D_B d_{t+2} = (D_B W_{t+2})^\top(m\mathbf{v}_{t+2}) + W_{t+2}^\top D_B(m\mathbf{v}_{t+2})$. Therefore $\|\nabla_{W_0} z_{t+2}\|_F + \|\nabla_{W_0} d_{t+2}\|_F \le C_t\sqrt{m}\, L_\delta^{\zeta_t}$ and

$$\|\nabla_{u_0} z_{t+2}\|_2 + \|\nabla_{m\mathbf{v}_0} z_{t+2}\|_2 + \|\nabla_{u_0} d_{t+2}\|_2 + \|\nabla_{m\mathbf{v}_0} d_{t+2}\|_2 \le C_t L_\delta^{\zeta_t},$$

where

$$\zeta_t := \max\{\nu_t + \lambda_t,\, \mu_t + \max\{3\beta_t, \lambda_t\}\} \le \max\{\alpha_{t+1} + 5\beta_t,\, 2\alpha_{t+1} + 3\beta_t\}.$$

By the exponent inequalities recorded above and $\zeta_t \le \beta_{t+1}$, we have

$$\|\nabla_{W_0} z_{t+2}\|_F + \|\nabla_{W_0} d_{t+2}\|_F \le C_{t+1}\sqrt{m}\, L_\delta^{\beta_{t+1}},$$

while

$$\|\nabla_{u_0} z_{t+2}\|_2 + \|\nabla_{m\mathbf{v}_0} z_{t+2}\|_2 + \|\nabla_{u_0} d_{t+2}\|_2 + \|\nabla_{m\mathbf{v}_0} d_{t+2}\|_2 \le C_{t+1} L_\delta^{\beta_{t+1}}.$$

Using $\|W_0\|_{\mathrm{op}} \le C_T L_\delta^{1/2}$ and the sharper estimates from Steps 2a–2c,

$$\|W_0^\top(m\mathbf{v}_{t+2})\|_2 + \|W_0 u_{t+2}\|_2 \le C_t\sqrt{m}\, L_\delta^{\lambda_t+1/2}, \quad \|W_0 d_{t+2}\|_2 \le C_t\sqrt{m}\, L_\delta^{2\alpha_{t+1}+3\beta_t+1/2}.$$

The exponent inequalities with the additive $+1$ imply that both displays are bounded by $C_{t+1}\sqrt{m}\, L_\delta^{\beta_{t+1}}$.

For the W_0-block, there are

$$D_{W_0}(W_0 u_{t+2})[H] = H u_{t+2} + W_0 D_{W_0} u_{t+2}[H],$$
$$D_{W_0}\big(W_0^\top (m\mathbf{v}_{t+2})\big)[H] = H^\top (m\mathbf{v}_{t+2}) + W_0^\top D_{W_0}(m\mathbf{v}_{t+2})[H],$$
$$D_{W_0}(W_0 d_{t+2})[H] = H d_{t+2} + W_0 D_{W_0} d_{t+2}[H].$$

Thus the first two derivatives are bounded by $C_t\sqrt{m}\, L_\delta^{\max\{\lambda_t,\, \max\{3\beta_t, \lambda_t\}+1/2\}}$, and the last one is bounded by $C_t\sqrt{m}\, L_\delta^{\max\{2\alpha_{t+1}+3\beta_t,\, \zeta_t+1/2\}}$.

Because

$$\max\{3\beta_t, \lambda_t\} + 1 \le \beta_{t+1}, \quad 2\alpha_{t+1} + 3\beta_t + 1 \le \beta_{t+1}, \quad \zeta_t + 1 \le \beta_{t+1},$$

all three W_0-block derivatives are at most $C_{t+1}\sqrt{m}\, L_\delta^{\beta_{t+1}}$. For the u_0- and $m\mathbf{v}_0$-blocks there is no derivative of the leading W_0, so the same operator-norm factor $L_\delta^{1/2}$ multiplies the bounds from above. The same exponent relations yield

$$\|\nabla_{u_0}(W_0^\top (m\mathbf{v}_{t+2}))\|_2 + \|\nabla_{m\mathbf{v}_0}(W_0^\top (m\mathbf{v}_{t+2}))\|_2 + \|\nabla_{u_0}(W_0 u_{t+2})\|_2 +$$
$$\|\nabla_{m\mathbf{v}_0}(W_0 u_{t+2})\|_2 + \|\nabla_{u_0}(W_0 d_{t+2})\|_2 + \|\nabla_{m\mathbf{v}_0}(W_0 d_{t+2})\|_2 \le C_{t+1} L_\delta^{\beta_{t+1}}.$$

We have proved items 2 and 3 of $\mathcal{P}_{t+1}(\delta)$.

Step 3: Fourth moments from the new tails. The estimates in items 2 and 3 of $\mathcal{P}_{t+1}(\eta)$ hold uniformly for every $\eta \in (0, 1/2]$. Applying Lemma B.3 gives, for every vector U covered by $\mathcal{P}_{t+1}$,

$$\mathbb{E}\|U\|_2^4 \le C_{t+1} m^2, \quad \mathbb{E}\|\nabla_{W_0} U\|_F^4 \le C_{t+1} m^2, \quad \mathbb{E}\|\nabla_{u_0} U\|_2^4 + \mathbb{E}\|\nabla_{m\mathbf{v}_0} U\|_2^4 \le C_{t+1}.$$

These are the moment hypotheses needed for the next scalar concentration step.

Step 4: Concentrate new scalar at time $t+1$. The vector family up to time $t+2$ now satisfies the hypotheses of Lemma F.7 with horizon $t+1$. Hence every newly appearing scalar at time $t+1$ that is a normalized inner product satisfies $|S - \mathbb{E}S| \le C_{t+1} m^{-1/2} L_\delta^{\alpha_{t+1}}$, $|\mathbb{E}S| \le C_{t+1}$. This includes $\{J^x_{t+1},\, A^{vd}_{t+1},\, Q^z_{t+1},\, H^{z,Wd}_{t+1},\, K^{xx}_{r,t+2},\, K^{vv}_{t+1,r},\, G^{xv}_{t+1,r}\}$, $0 \le r \le t+1$. For the increments, we use the one-observable form of Lemma F.7 directly on the two normalized inner products in

$$D^x_{r,t+1} := \frac{1}{m}\langle u_r, u_{t+2}\rangle - \frac{1}{m}\langle u_r, u_{t+1}\rangle, \qquad 0 \le r \le t+1.$$

Thus both summands are concentrated at scale $C_{t+1} m^{-1/2} L_\delta^{\alpha_{t+1}}$, and their expectations are $O(1)$. This also covers the endpoint $r = t+1$, where the second summand is the diagonal $m^{-1}\|u_{t+1}\|_2^2 = K^{xx}_{t+1,t+1}$, which is not one of the previously listed induction scalars. Therefore

$$|D^x_{r,t+1} - \mathbb{E}D^x_{r,t+1}| \le C_{t+1} m^{-1/2} L_\delta^{\alpha_{t+1}}, \quad |\mathbb{E}D^x_{r,t+1}| \le C_{t+1}.$$

Combining the concentration estimates with the expectation estimates gives $|S| + |\mathbb{E}S| \le C_{t+1} L_\delta^{\alpha_{t+1}}$ for every scalar newly appearing at time $t+1$. Together with item 1 of $\mathcal{P}_t(\delta)$ and $\alpha_t \le \alpha_{t+1}$, this proves item 1 of $\mathcal{P}_{t+1}(\delta)$. The induction closes.

Consequently $\mathcal{P}_t(\delta)$ holds for every $0 \le t \le T$, uniformly in δ. In particular, all scalar observables are controlled up to time T, and all vectors in the theorem family are controlled up to time $T+1$.

To reach the final vector level, repeat Steps 1 and 2 with $t = T$. Step 3 at level T supplies the fourth-moment hypotheses for $m\mathbf{v}_{T+1}$ and z_{T+1}, so the one-observable form of Lemma F.7 and step 1 gives the same bounds for $g^{(m)}_{\theta_{T+1}}$, χ_{T+1}, ρ_{T+1}, and the blockwise gradients of χ_{T+1} and ρ_{T+1} as in Step 1. Substituting those estimates into the recursions for $\{m\mathbf{v}_{T+2},\, u_{T+2},\, W_{T+2},\, z_{T+2},\, d_{T+2}\}$ and then into $\{W_0^\top (m\mathbf{v}_{T+2}),\, W_0 u_{T+2},\, W_0 d_{T+2}\}$ gives vector and block-gradient bounds at time $T+2$ with exponent β_{T+1}, hence also with the theorem's larger exponent β_{T+2}.

This terminal step uses only the already-established high-probability and fourth-moment bounds. It does not require the degree bound for the W_0-derived vectors at time $T+2$ from Lemma F.6.

The event used for the theorem is the finite intersection of the events produced above, with the failure probabilities allocated by δ_*. It therefore has probability at least $1-\delta$. For a scalar observable at time t,

use item 1 of $\mathcal{P}_t(\delta_*)$ and absorb $L_{\delta_*} \le C_T L_\delta$; this gives the stated exponent α_t. For vectors at time s, use $\mathcal{P}_0$ if $s = 0$, use $\mathcal{P}_{s-1}$ if $1 \le s \le T+1$, and use the terminal bootstrap if $s = T+2$. Since $(\beta_s)_{s\ge 0}$ is increasing, the resulting bounds are all absorbed by $R_s(\delta) = C_T\sqrt{m}\, L_\delta^{\beta_s}$, $r_s(\delta) = C_T L_\delta^{\beta_s}$. Enlarging C_T once more to dominate the finitely many constants completes the proof. □

Expectation asymptotics We now upgrade Theorem F.8 from concentration around finite-width expectations to concentration around a deterministic scalar trajectory. The proof uses only the elementary polynomial expectation-induction principle from Proposition E.2.

Throughout this paragraph write $a_t := m\mathbf{v}_t$. Thus the three-layer linear recursions are

$$a_{t+1} = a_t - \rho_t z_t, \quad u_{t+1} = u_t - \rho_t d_t, \quad W_{t+1} = W_t - \frac{\rho_t}{m} a_t u_t^\top,$$

with $z_t = W_t u_t$, $d_t = W_t^\top a_t$, $g_{\theta_t}^{(m)}(\xi) = \xi\, \frac{1}{m}\langle a_t, z_t\rangle$.

For derivatives with respect to the initialized output base vector, we write D_{a_0} for $D_{m\mathbf{v}_0}$.

Let $\mathfrak{W}_\ell$ be the set of all words of length ℓ in the two letters W_0 and $W_0^\top$. The empty word belongs to $\mathfrak{W}_0$ and acts as the identity. For $L \ge 0$, define $\mathfrak{W}_{\le L} := \bigcup_{\ell=0}^{L} \mathfrak{W}_\ell$. If $R = B_1 B_2 \cdots B_\ell$, $B_j \in \{W_0, W_0^\top\}$, then $|R| = \ell$, and Rx denotes the corresponding matrix product applied to $x \in \mathbb{R}^m$.

Fix the training horizon T. Choose $L_* := 2T + 4$, $L(t) := L_* - t$, $0 \le t \le T + 1$. The only property of this choice used below is $L(t + 1) + 1 = L(t)$. Hence, whenever $R \in \mathfrak{W}_{\le L(t+1)}$, both concatenated words RW_0, $RW_0^\top$ belong to $\mathfrak{W}_{\le L(t)}$.

For $0 \le t \le T + 1$, define

$$\mathcal{V}_t^\star := \Big\{ Ru_s,\ Ra_s,\ Rz_s,\ Rd_s : 0 \le s \le t + 1,\ R \in \mathfrak{W}_{\le L(t)} \Big\}.$$

This is a finite dictionary whose size depends on T, but not on m.

Let $\mathcal{I}_t^{(m)}$ be the collection of all normalized inner products generated by $\mathcal{V}_t^\star$: $\mathcal{I}_t^{(m)} := \big\{ \frac{1}{m}\langle U, V\rangle : U, V \in \mathcal{V}_t^\star \big\}$. Define $Y_t^{\mathrm{lin},(m)}$ to be an arbitrary fixed ordering of the finite list

$$Y_t^{\mathrm{lin},(m)} := \Big(\mathcal{I}_t^{(m)},\ \{ g_{\theta_s}^{(m)}, \chi_s, \rho_s : 0 \le s \le t + 1 \} \Big).$$

The last scalars are included explicitly only for readability. They are already polynomial functions of the inner products in $\mathcal{I}_t^{(m)}$, since

$$G_{s,s}^{xv} := \frac{1}{m}\langle z_s, a_s\rangle \in \mathcal{I}_t^{(m)}, \quad g_{\theta_s}^{(m)} = \xi_s G_{s,s}^{xv}, \quad \chi_s = g_{\theta_s}^{(m)} - y(\xi_s), \quad \rho_s = \xi_s \chi_s.$$

The scalar family appearing in Theorem F.8 is

$$\mathcal{S}_t^{\mathrm{lin}} := \Big\{ g_{\theta_t}^{(m)}, \chi_t, \rho_t, J_t^x, A_t^{vd}, Q_t^z, H_t^{z,Wd} \Big\} \cup \Big\{ K_{r,t+1}^{xx}, K_{t,r}^{vv}, G_{t,r}^{xv}, D_{r,t}^x : 0 \le r \le t \Big\}.$$

Every $S \in \mathcal{S}_t^{\mathrm{lin}}$ is a coordinate of $Y_t^{\mathrm{lin},(m)}$ or a fixed polynomial function of its coordinates. Indeed,

$$K_{r,s}^{xx} = \frac{1}{m}\langle u_r, u_s\rangle, \quad K_{s,r}^{vv} = \frac{1}{m}\langle a_s, a_r\rangle, \quad G_{s,r}^{xv} = \frac{1}{m}\langle z_s, a_r\rangle, \quad J_s^x = \frac{1}{m}\langle u_s, W_0^\top a_s\rangle,$$

$$A_s^{vd} = \frac{1}{m}\langle a_s, W_0 d_s\rangle, \quad Q_s^z = \frac{1}{m}\langle z_s, z_s\rangle, \quad H_s^{z,Wd} = \frac{1}{m}\langle z_s, W_0 d_s\rangle.$$

Finally, $D_{r,s}^x = K_{r,s+1}^{xx} - K_{r,s}^{xx}$ is an exact difference of two included kernel coordinates.

The base estimate below proves that all coordinates of $\mathbb{E}Y_0^{\mathrm{lin},(m)}$ have coordinatewise limits. We define $Y_0^{\mathrm{lin},\infty} := \lim_{m\to\infty} \mathbb{E}Y_0^{\mathrm{lin},(m)}$. Below we prove that the word-closed state satisfies an exact polynomial recursion $Y_{t+1}^{\mathrm{lin},(m)} = P_t^{\mathrm{lin}}\big(Y_t^{\mathrm{lin},(m)}\big)$, where P_t^{lin} is independent of m. We then define the deterministic trajectory by $Y_{t+1}^{\mathrm{lin},\infty} := P_t^{\mathrm{lin}}\big(Y_t^{\mathrm{lin},\infty}\big)$. For any observable $S \in \mathcal{S}_t^{\mathrm{lin}}$, let S^∞ denote the corresponding coordinate, or the corresponding fixed polynomial function, evaluated at $Y_t^{\mathrm{lin},\infty}$. Define $\bar{g}_{\theta_t}(\xi) := \xi\,(G_{t,t}^{xv})^\infty$, $\bar{g}_{\theta_t} := \bar{g}_{\theta_t}(\xi_t)$.

Lemma F.9 (Exact word-closed polynomial closure)**.** *Fix $T < \infty$ and use the word-length schedule $L(t) = L_* - t$ above. For each $0 \le t \le T$, there is a finite index set $\mathcal{J}_t$, compatible with the fixed ordering of $Y_t^{\mathrm{lin},(m)}$, and a polynomial map $P_t^{\mathrm{lin}} : \mathbb{R}^{\mathcal{J}_t} \to \mathbb{R}^{\mathcal{J}_{t+1}}$, independent of m, such that $Y_{t+1}^{\mathrm{lin},(m)} = P_t^{\mathrm{lin}}\big(Y_t^{\mathrm{lin},(m)}\big)$ exactly.*

Moreover, every observable in $\mathcal{S}_s^{\mathrm{lin}}$, $0 \le s \le t$, is either a coordinate of $Y_t^{\mathrm{lin},(m)}$ or a fixed polynomial function of coordinates of $Y_t^{\mathrm{lin},(m)}$.

Proof. Define the vector-coordinate index set

$$\mathcal{J}_t^{\mathrm{vec}} := \big\{(\tau, s, R) : \tau \in \{u, a, z, d\},\ 0 \le s \le t+1,\ R \in \mathfrak{W}_{\le L(t)}\big\},$$

where

$$V_{u,s,R} := Ru_s, \quad V_{a,s,R} := Ra_s, \quad V_{z,s,R} := Rz_s, \quad V_{d,s,R} := Rd_s.$$

The scalar coordinates of $Y_t^{\mathrm{lin},(m)}$ are indexed by

$$\mathcal{J}_t := \big(\mathcal{J}_t^{\mathrm{vec}} \times \mathcal{J}_t^{\mathrm{vec}}\big) \sqcup \big(\{g, \chi, \rho\} \times \{0, \dots, t+1\}\big), \quad Y_{t,(\iota,\iota')}^{\mathrm{lin},(m)} = \frac{1}{m}\langle V_\iota, V_{\iota'}\rangle, \quad \iota, \iota' \in \mathcal{J}_t^{\mathrm{vec}},$$

and with the remaining coordinates equal to $g_{\theta_s}^{(m)}, \chi_s, \rho_s$, $0 \le s \le t+1$. This only formalizes the fixed ordering of $Y_t^{\mathrm{lin},(m)}$ used above.

Fix $R \in \mathfrak{W}_{\le L(t+1)}$. Since $L(t+1)+1 = L(t)$, both RW_0 and $RW_0^\top$ belong to $\mathfrak{W}_{\le L(t)}$.

First, the direct recursions give

$$Ra_{t+2} = Ra_{t+1} - \rho_{t+1} Rz_{t+1}, \quad Ru_{t+2} = Ru_{t+1} - \rho_{t+1} Rd_{t+1}.$$

All vectors on the right belong to $\mathcal{V}_t^\star$, and ρ_{t+1} is a coordinate of $Y_t^{\mathrm{lin},(m)}$.

Next, for $0 \le r \le t+1$,

$$K_{r,t+2}^{xx} = K_{r,t+1}^{xx} - \rho_{t+1}\frac{1}{m}\langle u_r, d_{t+1}\rangle, \quad K_{t+2,r}^{vv} = K_{t+1,r}^{vv} - \rho_{t+1}\frac{1}{m}\langle z_{t+1}, a_r\rangle.$$

Each normalized inner product on the right-hand side is a coordinate of $Y_t^{\mathrm{lin},(m)}$. Hence $K_{r,t+2}^{xx}$ and $K_{t+2,r}^{vv}$ are fixed polynomial functions of $Y_t^{\mathrm{lin},(m)}$.

Using Lemma F.1,

$$z_{t+2} = W_0 u_{t+2} - \sum_{r=0}^{t+1} \rho_r a_r K_{r,t+2}^{xx}, \qquad d_{t+2} = W_0^\top a_{t+2} - \sum_{r=0}^{t+1} \rho_r u_r K_{t+2,r}^{vv}.$$

Multiplying by R, and then using the two direct recursions above, gives the explicit new-vector formulas

$$Rz_{t+2} = RW_0 u_{t+1} - \rho_{t+1} RW_0 d_{t+1} - \sum_{r=0}^{t+1} \rho_r\, Ra_r\, K_{r,t+2}^{xx},$$

$$Rd_{t+2} = RW_0^\top a_{t+1} - \rho_{t+1} RW_0^\top z_{t+1} - \sum_{r=0}^{t+1} \rho_r\, Ru_r\, K_{t+2,r}^{vv}.$$

Because $RW_0, RW_0^\top \in \mathfrak{W}_{\le L(t)}$, every vector appearing on the right-hand side belongs to $\mathcal{V}_t^\star$, and every scalar coefficient is a fixed polynomial function of $Y_t^{\mathrm{lin},(m)}$. Thus every new vector in $\mathcal{V}_{t+1}^\star$ is a finite linear combination of vectors in $\mathcal{V}_t^\star$ with polynomial coefficients in $Y_t^{\mathrm{lin},(m)}$.

Consequently, if

$$U' = \sum_i c_i(Y_t^{\mathrm{lin},(m)}) U_i, \qquad V' = \sum_j d_j(Y_t^{\mathrm{lin},(m)}) V_j,$$

where $U_i, V_j \in \mathcal{V}_t^\star$, then

$$\frac{1}{m}\langle U', V'\rangle = \sum_{i,j} c_i(Y_t^{\mathrm{lin},(m)}) d_j(Y_t^{\mathrm{lin},(m)}) \frac{1}{m}\langle U_i, V_j\rangle.$$

The last normalized inner products are coordinates of $Y_t^{\mathrm{lin},(m)}$. Therefore every coordinate of $\mathcal{I}_{t+1}^{(m)}$ is a fixed polynomial in coordinates of $Y_t^{\mathrm{lin},(m)}$.Finally,

$$g_{\theta_{t+2}}^{(m)} = \xi_{t+2}\frac{1}{m}\langle a_{t+2}, z_{t+2}\rangle, \qquad \chi_{t+2} = g_{\theta_{t+2}}^{(m)} - y(\xi_{t+2}), \qquad \rho_{t+2} = \xi_{t+2}\chi_{t+2},$$

are also fixed polynomial functions of $Y_t^{\mathrm{lin},(m)}$. The older explicit scalar coordinates are copied from $Y_t^{\mathrm{lin},(m)}$. This proves the exact recursion $Y_{t+1}^{\mathrm{lin},(m)} = P_t^{\mathrm{lin}}\big(Y_t^{\mathrm{lin},(m)}\big)$ with no stochastic remainder.

It remains only to identify the scalar observables used in the concentration theorem. For $0 \le s \le t$,

$$K_{r,s}^{xx} = \frac{1}{m}\langle u_r, u_s\rangle, \quad K_{s,r}^{vv} = \frac{1}{m}\langle a_s, a_r\rangle, \quad G_{s,r}^{xv} = \frac{1}{m}\langle z_s, a_r\rangle,$$

are coordinates of $Y_t^{\mathrm{lin},(m)}$. Also

$$J_s^x = \frac{1}{m}\langle u_s, W_0^\top a_s\rangle, \quad A_s^{vd} = \frac{1}{m}\langle a_s, W_0 d_s\rangle, \quad Q_s^z = \frac{1}{m}\langle z_s, z_s\rangle, \quad H_s^{z,Wd} = \frac{1}{m}\langle z_s, W_0 d_s\rangle$$

are normalized inner products of vectors in $\mathcal{V}_t^\star$. Finally, $D_{r,s}^x = K_{r,s+1}^{xx} - K_{r,s}^{xx}$. Thus every $S \in \mathcal{S}_s^{\mathrm{lin}}$, $0 \le s \le t$, is either a coordinate of $Y_t^{\mathrm{lin},(m)}$ or a fixed polynomial function of its coordinates. □

Lemma F.10 (Fixed high moments for the basic vector nodes)**.** *Fix $T < \infty$, and let $\mathcal{B}_T := \{u_s, a_s, z_s, d_s : 0 \le s \le T+2\}$. For every fixed $p \ge 1$ and every $U \in \mathcal{B}_T$,*

$$\mathbb{E}\|U\|_2^p \le C_{T,p} m^{p/2}, \quad \mathbb{E}\|D_{W_0}U\|_{F\to 2}^p \le C_{T,p} m^{p/2}, \quad \mathbb{E}\|D_{u_0}U\|_{2\to 2}^p + \mathbb{E}\|D_{a_0}U\|_{2\to 2}^p \le C_{T,p}.$$

Proof. The proof of Theorem F.8 is uniform in the tail parameter. In particular, for every $\eta \in (0, 1/2]$, every $U \in \mathcal{B}_T$, and constants depending only on T,

$$m^{-1/2}\|U\|_2 + m^{-1/2}\|D_{W_0}U\|_{F\to 2} + \|D_{u_0}U\|_{2\to 2} + \|D_{a_0}U\|_{2\to 2} \le C_T L_\eta^{\beta_{T+2}}$$

with probability at least $1-\eta$. Applying Lemma B.3 to each normalized quantity gives, for every fixed $p \ge 1$, $\mathbb{E}\left(m^{-1/2}\|U\|_2\right)^p \le C_{T,p}$, $\mathbb{E}\left(m^{-1/2}\|D_{W_0}U\|_{F\to 2}\right)^p \le C_{T,p}$, $\mathbb{E}\|D_{u_0}U\|_{2\to 2}^p + \mathbb{E}\|D_{a_0}U\|_{2\to 2}^p \le C_{T,p}$. Multiplying back the powers of $m^{1/2}$ proves the claim. □

Lemma F.11 (Moment and concentration bounds for the word-closed dictionary)**.** *Fix $T < \infty$. For every fixed $q \ge 1$, every $0 \le t \le T+1$, and every $U \in \mathcal{V}_t^\star$,*

$$\mathbb{E}\|U\|_2^{2q} \le C_{T,q} m^q, \quad \mathbb{E}\|D_{W_0}U\|_{F\to 2}^{2q} \le C_{T,q} m^q, \quad \mathbb{E}\|D_{u_0}U\|_{2\to 2}^{2q} + \mathbb{E}\|D_{a_0}U\|_{2\to 2}^{2q} \le C_{T,q}.$$

Consequently, for every scalar coordinate S of $Y_t^{\mathrm{lin},(m)}$ and every fixed $q \ge 2$,

$$\|S - \mathbb{E}S\|_{L^q} \le C_{T,q} m^{-1/2}, \qquad \|S\|_{L^q} + |\mathbb{E}S| \le C_{T,q}.$$

Proof. Every vector $U \in \mathcal{V}_t^\star$ has the form $U = RU_0$, where $U_0 \in \mathcal{B}_T$, $R \in \mathfrak{W}_{\le L_*}$, and $L_* = 2T+4$. Write $\ell := |R|$. The matrix W_0 has i.i.d. $N(0, m^{-1})$ entries, so for every fixed $p \ge 1$, $\mathbb{E}\|W_0\|_{\mathrm{op}}^p \le C_p$.

First, $\|RU_0\|_2 \le \|W_0\|_{\mathrm{op}}^\ell \|U_0\|_2$. Since U_0 depends on W_0, we do not separate expectations. Instead, by Hölder,

$$\mathbb{E}\|RU_0\|_2^{2q} \le \left(\mathbb{E}\|W_0\|_{\mathrm{op}}^{4q\ell}\right)^{1/2} \left(\mathbb{E}\|U_0\|_2^{4q}\right)^{1/2}.$$

By Lemma F.10, $\mathbb{E}\|U_0\|_2^{4q} \le C_{T,q} m^{2q}$, and the operator-norm moment is bounded by a constant depending only on T, q. Therefore $\mathbb{E}\|RU_0\|_2^{2q} \le C_{T,q} m^q$.

Next consider the u_0- and a_0-block derivatives. Since R depends only on W_0, $D_{u_0}(RU_0) = R\,D_{u_0}U_0$, $D_{a_0}(RU_0) = R\,D_{a_0}U_0$. Hence $\|D_{u_0}(RU_0)\|_{2\to 2} \le \|W_0\|_{\mathrm{op}}^{\ell}\|D_{u_0}U_0\|_{2\to 2}$, and similarly for D_{a_0}. Hölder gives

$$\mathbb{E}\|D_{u_0}(RU_0)\|_{2\to 2}^{2q} \le \left(\mathbb{E}\|W_0\|_{\mathrm{op}}^{4q\ell}\right)^{1/2}\left(\mathbb{E}\|D_{u_0}U_0\|_{2\to 2}^{4q}\right)^{1/2} \le C_{T,q},$$

and the same bound holds for $D_{a_0}(RU_0)$.

It remains to treat the W_0-block derivative. Write $R = B_1B_2\cdots B_\ell$, $B_j \in \{W_0, W_0^\top\}$. For a perturbation H, the product rule gives

$$D_{W_0}(RU_0)[H] = R\,D_{W_0}U_0[H] + \sum_{j=1}^{\ell} B_1\cdots B_{j-1}\,\widetilde{H}_j\,B_{j+1}\cdots B_\ell\,U_0,$$

where $\widetilde{H}_j = H$ if $B_j = W_0$, and $\widetilde{H}_j = H^\top$ if $B_j = W_0^\top$. Since $\|H\|_{\mathrm{op}} \le \|H\|_F$, taking the supremum over $\|H\|_F = 1$ gives

$$\|D_{W_0}(RU_0)\|_{F\to 2} \le \|W_0\|_{\mathrm{op}}^{\ell}\|D_{W_0}U_0\|_{F\to 2} + \ell\,\|W_0\|_{\mathrm{op}}^{\ell-1}\|U_0\|_2,$$

with the second term omitted when $\ell = 0$. Therefore, using $(x+y)^{2q} \le C_q(x^{2q}+y^{2q})$,

$$\mathbb{E}\|D_{W_0}(RU_0)\|_{F\to 2}^{2q} \le C_q\mathbb{E}\Big[\|W_0\|_{\mathrm{op}}^{2q\ell}\|D_{W_0}U_0\|_{F\to 2}^{2q}\Big] + C_{q,\ell}\mathbb{E}\Big[\|W_0\|_{\mathrm{op}}^{2q(\ell-1)}\|U_0\|_2^{2q}\Big].$$

By Hölder and Lemma F.10,

$$\mathbb{E}\Big[\|W_0\|_{\mathrm{op}}^{2q\ell}\|D_{W_0}U_0\|_{F\to 2}^{2q}\Big] \le \left(\mathbb{E}\|W_0\|_{\mathrm{op}}^{4q\ell}\right)^{1/2}\left(\mathbb{E}\|D_{W_0}U_0\|_{F\to 2}^{4q}\right)^{1/2} \le C_{T,q}m^q.$$

Similarly,

$$\mathbb{E}\Big[\|W_0\|_{\mathrm{op}}^{2q(\ell-1)}\|U_0\|_2^{2q}\Big] \le C_{T,q}m^q.$$

Thus $\mathbb{E}\|D_{W_0}(RU_0)\|_{F\to 2}^{2q} \le C_{T,q}m^q$.

Now let $S = \frac{1}{m}\langle U, V\rangle$, $U, V \in \mathcal{V}_t^\star$. The moment bounds above imply

$$\|S\|_{L^q} \le \frac{1}{m}\left(\mathbb{E}\|U\|_2^{2q}\right)^{1/(2q)}\left(\mathbb{E}\|V\|_2^{2q}\right)^{1/(2q)} \le C_{T,q}.$$

In particular, $|\mathbb{E}S| \le C_{T,q}$.

For concentration, differentiate S. For the W_0-block,

$$D_{W_0}S[H] = \frac{1}{m}\left(\langle D_{W_0}U[H], V\rangle + \langle U, D_{W_0}V[H]\rangle\right),$$

so

$$\|D_{W_0}S\|_{F\to\mathbb{R}} \le \frac{1}{m}\left(\|D_{W_0}U\|_{F\to 2}\|V\|_2 + \|U\|_2\|D_{W_0}V\|_{F\to 2}\right).$$

Using the fourth-moment case of the bounds just proved, there is $\mathbb{E}\|D_{W_0}S\|_{F\to\mathbb{R}}^2 \le C_T$. Similarly,

$$\|D_{u_0}S\|_{2\to\mathbb{R}} \le \frac{1}{m}\left(\|D_{u_0}U\|_{2\to 2}\|V\|_2 + \|U\|_2\|D_{u_0}V\|_{2\to 2}\right),$$

and hence $\mathbb{E}\|D_{u_0}S\|_{2\to\mathbb{R}}^2 \le C_T m^{-1}$. The same bound holds for the a_0-block: $\mathbb{E}\|D_{a_0}S\|_{2\to\mathbb{R}}^2 \le C_T m^{-1}$. The weighted Gaussian Poincaré inequality gives

$$\mathrm{Var}(S) \le \frac{1}{m}\mathbb{E}\|D_{W_0}S\|_{F\to\mathbb{R}}^2 + \mathbb{E}\|D_{u_0}S\|_{2\to\mathbb{R}}^2 + \mathbb{E}\|D_{a_0}S\|_{2\to\mathbb{R}}^2 \le C_T m^{-1}.$$

Thus $\|S - \mathbb{E}S\|_{L^2} \le C_T m^{-1/2}$.

Finally, $S - \mathbb{E}S$ is a centered Gaussian polynomial of degree bounded only in terms of T. Indeed, each basic vector $U_0 \in \mathcal{B}_T$ has Gaussian polynomial degree bounded by Lemma F.6, and multiplying by a word $R \in \mathfrak{W}_{\le L_*}$ increases the degree by at most L_*. Therefore the scalar $m^{-1}\langle U, V\rangle$ has

degree bounded by a constant depending only on T. Gaussian hypercontractivity gives, for every fixed $q \geq 2$,

$$\|S - \mathbb{E}S\|_{L^q} \leq C_{T,q}\|S - \mathbb{E}S\|_{L^2} \leq C_{T,q} m^{-1/2}.$$

The explicitly included scalars $g_{\theta_s}^{(m)}$, χ_s, ρ_s are affine functions of the normalized inner product

$$G_{s,s}^{xv} := \frac{1}{m}\langle z_s, a_s\rangle, \qquad g_{\theta_s}^{(m)} = \xi_s G_{s,s}^{xv}, \quad \chi_s = g_{\theta_s}^{(m)} - y(\xi_s), \quad \rho_s = \xi_s \chi_s.$$

Hence the same L^q bounds hold for every scalar coordinate of $Y_t^{\mathrm{lin},(m)}$. □

Lemma F.12 (Wick–Isserlis formula for centered Gaussian vectors)**.** *Let $X = (X_1, \ldots, X_N)$ be a centered Gaussian vector with covariance matrix $\Sigma_{ij} := \mathbb{E}[X_i X_j]$. Then, for any indices $i_1, \ldots, i_r \in [N]$, $\mathbb{E}[X_{i_1} \cdots X_{i_r}] = 0$ if r is odd,*

and, if $r = 2k$ is even, then

$$\mathbb{E}[X_{i_1} \cdots X_{i_{2k}}] = \sum_{\pi \in \mathcal{P}_2(2k)} \prod_{\{p,q\} \in \pi} \Sigma_{i_p i_q},$$

where $\mathcal{P}_2(2k)$ denotes the set of all pair partitions of $\{1, \ldots, 2k\}$.

Proof. Let $t = (t_1, \ldots, t_N) \in \mathbb{R}^N$. The moment generating function of X is

$$M_X(t) := \mathbb{E}e^{t^\top X} = \exp\left(\frac{1}{2} t^\top \Sigma t\right) = \exp\left(\frac{1}{2} \sum_{a,b=1}^{N} \Sigma_{ab} t_a t_b\right).$$

For any indices $i_1, \ldots, i_r$,

$$\mathbb{E}[X_{i_1} \cdots X_{i_r}] = \partial_{i_1} \cdots \partial_{i_r} M_X(t)|_{t=0}.$$

The exponential $\exp\left(\frac{1}{2}\sum_{a,b} \Sigma_{ab} t_a t_b\right)$ is a power series containing only monomials of even total degree, because its exponent is homogeneous of degree 2. Hence, if r is odd, the derivative of total order r evaluated at $t = 0$ is zero.

Now let $r = 2k$. Only the k-th term in the exponential series can contribute to the coefficient of a monomial of total degree $2k$. Thus

$$M_X(t) = \frac{1}{k!}\left(\frac{1}{2} \sum_{a,b=1}^{N} \Sigma_{ab} t_a t_b\right)^k + \{\text{terms of degree different from } 2k\}.$$

Expanding the k-th power gives a sum of products of k quadratic factors

$$\Sigma_{a_1 b_1} \cdots \Sigma_{a_k b_k} t_{a_1} t_{b_1} \cdots t_{a_k} t_{b_k}.$$

After applying $\partial_{i_1} \cdots \partial_{i_{2k}}$ and setting $t = 0$, a term contributes exactly when the derivative indices $i_1, \ldots, i_{2k}$ can be matched to the $2k$ variables appearing in the k quadratic factors. Equivalently, the $2k$ positions are partitioned into k unordered pairs. Each pair $\{p, q\}$ contributes the covariance $\Sigma_{i_p i_q}$. The combinatorial factor $1/(2^k k!)$ from the exponential expansion is exactly canceled by the $2^k k!$ different orderings of the same unordered pairing. Therefore

$$\mathbb{E}[X_{i_1} \cdots X_{i_{2k}}] = \sum_{\pi \in \mathcal{P}_2(2k)} \prod_{\{p,q\} \in \pi} \Sigma_{i_p i_q}.$$

This proves the formula. □

Lemma F.13 (Fixed-word Gaussian expectation expansion)**.** *Let $a, b \in \mathbb{R}^m$ be independent standard Gaussian vectors, and let $M \in \mathbb{R}^{m \times m}$ have independent $N(0, m^{-1})$ entries, independent of a, b. Fix finitely many words $R_1', R_1'', \ldots, R_r', R_r''$ in the two letters M and $M^\top$, with all word lengths bounded independently of m. For each ℓ, let $x_\ell, y_\ell \in \{a, b\}$, and define $Z_\ell^{(m)} := \frac{1}{m}\langle R_\ell' x_\ell, R_\ell'' y_\ell\rangle$. Then $\mathbb{E}\prod_{\ell=1}^r Z_\ell^{(m)} = c + O(m^{-1})$, where c depends only on the words and on the choices x_ℓ, y_ℓ, not on m. Consequently, if P is any fixed polynomial in finitely many such normalized fixed-word inner products, then $\mathbb{E}P(Z^{(m)}) = P^\infty + O(m^{-1})$ for some deterministic constant P^∞.*

Proof. Write $M_{ij} = m^{-1/2} G_{ij}$, where $(G_{ij})_{1 \le i,j \le m}$ are independent standard Gaussian variables, independent of a, b. Thus the entire family

$$\{a_i : 1 \le i \le m\} \cup \{b_i : 1 \le i \le m\} \cup \{G_{ij} : 1 \le i, j \le m\}$$

is a finite independent centered Gaussian family.

We first prove the claim for the product $\prod_{\ell=1}^{r} Z_\ell^{(m)}$. Expanding every word R'_ℓ and R''_ℓ into matrix entries gives a finite index sum. More explicitly, since all word lengths are fixed, there exists a fixed integer D, independent of m, such that each summand has the form $m^{-r-D/2} \prod_{\nu=1}^{D} G_{\alpha_\nu \beta_\nu} \prod_{\mu=1}^{A} a_{\gamma_\mu} \prod_{\kappa=1}^{B} b_{\eta_\kappa}$, where the indices $\alpha_\nu, \beta_\nu, \gamma_\mu, \eta_\kappa \in [m]$ are summed over. The leading factor m^{-r} comes from the r normalized inner products, and the factor $m^{-D/2}$ comes from the D occurrences of $M_{ij} = m^{-1/2} G_{ij}$. The integers D, A, B depend only on the fixed words, not on m.

For any fixed assignment pattern of the indices, the expectation of the Gaussian monomial is computed by Lemma F.12. Since the variables a, b, and G are independent, a nonzero Wick pairing can only pair a-variables with a-variables, b-variables with b-variables, and G-variables with G-variables having the same matrix-entry labels. Thus:

$$\mathbb{E}[a_i a_j] = \mathbf{1}_{\{i=j\}}, \quad \mathbb{E}[b_i b_j] = \mathbf{1}_{\{i=j\}}, \quad \mathbb{E}[G_{ij} G_{kl}] = \mathbf{1}_{\{i=k\}} \mathbf{1}_{\{j=l\}}.$$

Equivalently, in terms of M, $\mathbb{E}[M_{ij} M_{kl}] = m^{-1} \mathbf{1}_{\{i=k\}} \mathbf{1}_{\{j=l\}}$.

Each admissible pairing imposes a finite set of equality constraints among the summation indices. After imposing those Kronecker-delta constraints, suppose F free indices remain. Summing those free indices contributes m^F. Therefore each pairing pattern contributes a term of the form $C_\pi \, m^{-r-D/2+F_\pi}$, where C_π is a deterministic constant depending only on the pairing pattern. Since the number of Gaussian variables is fixed, the number of pairing patterns is finite. Hence $\mathbb{E} \prod_{\ell=1}^{r} Z_\ell^{(m)}$ is a finite Laurent expansion in powers of m. Moreover, every nonzero term has an integer power of m: the number of G-variables must be even in any nonzero Wick pairing, so $D/2$ is an integer for contributing terms.

It remains to show that no positive power of m can appear. By the fixed-word moment bounds proved in Lemma F.11, or directly by the same fixed-word Gaussian moment computation applied to norms, each normalized fixed-word inner product satisfies $\sup_{m \ge 1} \|Z_\ell^{(m)}\|_{L^r} < \infty$ for every fixed r. Hence, by Hölder's inequality,

$$\left| \mathbb{E} \prod_{\ell=1}^{r} Z_\ell^{(m)} \right| \le \prod_{\ell=1}^{r} \|Z_\ell^{(m)}\|_{L^r} \le C,$$

where C is independent of m. A finite Laurent expansion that is uniformly bounded as $m \to \infty$ cannot contain any positive power of m. Therefore the expansion has the form $c_0 + c_1 m^{-1} + \cdots + c_J m^{-J}$. Taking $c := c_0$ gives $\mathbb{E} \prod_{\ell=1}^{r} Z_\ell^{(m)} = c + O(m^{-1})$.

Finally, a fixed polynomial P is a finite linear combination of such products. Applying the product estimate term by term proves $\mathbb{E} P(Z^{(m)}) = P^\infty + O(m^{-1})$. □

Proposition F.14 (Expectation limit by word-closed induction for the three-layer linear model)**.** *For every fixed $0 \le t \le T$ and every $S \in \mathcal{S}_t^{\mathrm{lin}}$,*

$$\mathbb{E} S = S^\infty + O(m^{-1}), \qquad |S^\infty| \le C_T.$$

In particular, $\mathbb{E} g_{\theta_t}^{(m)} = \bar{g}_{\theta_t} + O(m^{-1})$. Moreover, at initialization the output bias is exactly zero: $\mathbb{E} g_{\theta_0}^{(m)} - \bar{g}_{\theta_0} = 0$.

Proof. We verify the hypotheses of Proposition E.2.

Step 1: base expectations. Write $a := a_0 = m \mathbf{v}_0$, $b := u_0$, $M := W_0$. Then $a, b \in \mathbb{R}^m$ are independent standard Gaussian vectors, and M has independent $N(0, m^{-1})$ entries, independent of a, b.

At time 0,

$$z_0 = Mb, \; d_0 = M^\top a, \; G_{0,0}^{xv} = \frac{1}{m} \langle a, Mb \rangle, \; g_{\theta_0}^{(m)} = \xi_0 G_{0,0}^{xv},$$

$$\chi_0 = g_{\theta_0}^{(m)} - y(\xi_0) = \xi_0 G_{0,0}^{xv} - y(\xi_0), \quad \rho_0 = \xi_0 \chi_0 = \xi_0^2 G_{0,0}^{xv} - \xi_0 y(\xi_0).$$

The first update gives

$$a_1 = a - \rho_0 z_0 = a - \rho_0 M b, \qquad u_1 = b - \rho_0 d_0 = b - \rho_0 M^\top a.$$

Using Lemma F.1, for $t = 0$,

$$z_1 = W_0 u_1 - \rho_0 a_0 K_{0,1}^{xx} = M u_1 - \rho_0 a K_{0,1}^{xx},$$

and therefore

$$z_1 = M(b - \rho_0 M^\top a) - \rho_0 a K_{0,1}^{xx} = Mb - \rho_0 M M^\top a - \rho_0 a K_{0,1}^{xx}.$$

Similarly,

$$d_1 = W_0^\top a_1 - \rho_0 u_0 K_{1,0}^{vv} = M^\top a_1 - \rho_0 b K_{1,0}^{vv},$$

so

$$d_1 = M^\top (a - \rho_0 M b) - \rho_0 b K_{1,0}^{vv} = M^\top a - \rho_0 M^\top M b - \rho_0 b K_{1,0}^{vv}.$$

It is useful to introduce

$$q_0 := b,\ q_1 := a,\ q_2 := Mb,\ q_3 := M^\top a,\ q_4 := MM^\top a,\ q_5 := M^\top Mb,\ q_6 := MM^\top Mb.$$

Then there is $z_0 = q_2$, $d_0 = q_3$, $W_0 d_0 = q_4$, $W_0 W_0^\top z_0 = q_6$. The concrete time-zero scalar observables are $K_{0,0}^{xx} = \frac{1}{m}\|q_0\|_2^2$ and

$$K_{0,0}^{vv} = \frac{1}{m}\|q_1\|_2^2, \quad G_{0,0}^{xv} = \frac{1}{m}\langle q_2, q_1\rangle, \quad Q_0^z = \frac{1}{m}\|q_2\|_2^2, \quad A_0^{vd} = \frac{1}{m}\langle q_1, q_4\rangle = \frac{1}{m}\|q_3\|_2^2,$$

$$H_0^{z,Wd} = \frac{1}{m}\langle q_2, q_4\rangle, \quad B_0^v = \frac{1}{m}\|q_4\|_2^2, \quad S_0^z = \frac{1}{m}\langle q_2, q_6\rangle, \quad T_0^z = \frac{1}{m}\langle q_4, q_6\rangle.$$

Every vector in $\mathcal{V}_0^\star$ is a finite linear combination of vectors of the form $R'a$, $R'b$, where R' is a word in M and $M^\top$ of length bounded by $L_* + 2$. The scalar coefficients in these linear combinations are fixed polynomials of finitely many normalized fixed-word inner products. The number of such vectors and all word lengths are bounded by constants depending only on T, not on m.

Now let R', R'' be any two fixed words in $M, M^\top$, and let $x, y \in \{a, b\}$. By Lemma F.13,

$$\mathbb{E}\frac{1}{m}\langle R'x, R''y\rangle = c(R', x; R'', y) + O(m^{-1}).$$

Moreover, the same lemma applies to any fixed polynomial in finitely many such normalized inner products. Therefore every coordinate of the initial word-closed scalar state satisfies $\mathbb{E}Y_{0,i}^{\mathrm{lin},(m)} = Y_{0,i}^{\mathrm{lin},\infty} + O(m^{-1})$ coordinatewisely.

We now record the most important concrete base expectations. Since a, b are standard Gaussian,

$$\mathbb{E}K_{0,0}^{xx} = \frac{1}{m}\sum_{i=1}^m \mathbb{E}b_i^2 = 1, \quad \mathbb{E}K_{0,0}^{vv} = \frac{1}{m}\sum_{i=1}^m \mathbb{E}a_i^2 = 1.$$

Also $G_{0,0}^{xv} = \frac{1}{m}\sum_{i,j=1}^m a_i M_{ij} b_j$. Each summand has mean zero because a_i, M_{ij}, and b_j are centered and independent. Hence

$$\mathbb{E}G_{0,0}^{xv} = 0, \quad \mathbb{E}g_{\theta_0}^{(m)} = \xi_0 \mathbb{E}G_{0,0}^{xv} = 0.$$

Consequently

$$\mathbb{E}\chi_0 = -y(\xi_0), \quad \mathbb{E}\rho_0 = -\xi_0 y(\xi_0).$$

Next, $Q_0^z = \frac{1}{m}\|Mb\|_2^2$. Conditioning on b, each coordinate $(Mb)_i$ has conditional variance $m^{-1}\|b\|_2^2$. Therefore $\mathbb{E}_M[\|Mb\|_2^2 \mid b] = \|b\|_2^2$, and hence $\mathbb{E}Q_0^z = \frac{1}{m}\mathbb{E}\|b\|_2^2 = 1$. Similarly, $A_0^{vd} = \frac{1}{m}\|M^\top a\|_2^2$, and conditioning on a gives $\mathbb{E}A_0^{vd} = 1$. For $H_0^{z,Wd} = \frac{1}{m}\langle Mb, MM^\top a\rangle$, condition on M, b. The expression is linear in a, and a is centered independent of M, b. Therefore $\mathbb{E}H_0^{z,Wd} = 0$.

We next compute the fourth-order base quantities. Since

$$B_0^v = \frac{1}{m}\|MM^\top a\|_2^2 = \frac{1}{m}a^\top MM^\top MM^\top a,$$

conditioning on M and using $\mathbb{E}[aa^\top] = I_m$ gives $\mathbb{E}_a[B_0^v \mid M] = \frac{1}{m}\operatorname{tr}\big((MM^\top)^2\big)$. Let $M = m^{-1/2}G$, where G has i.i.d. $N(0,1)$ entries. Then $\operatorname{tr}\big((MM^\top)^2\big) = \sum_{i,k=1}^m \left(\sum_{j=1}^m M_{ij}M_{kj}\right)^2$. If $i \neq k$, then

$$\mathbb{E}\left(\sum_j M_{ij}M_{kj}\right)^2 = \sum_j \mathbb{E}M_{ij}^2\,\mathbb{E}M_{kj}^2 = m \cdot m^{-2} = m^{-1}.$$

If $i = k$, then

$$\mathbb{E}\left(\sum_j M_{ij}^2\right)^2 = \sum_j \mathbb{E}M_{ij}^4 + \sum_{j\neq\ell}\mathbb{E}M_{ij}^2\,\mathbb{E}M_{i\ell}^2 = m \cdot 3m^{-2} + m(m-1)m^{-2} = 1 + \frac{2}{m}.$$

Therefore

$$\mathbb{E}\operatorname{tr}\big((MM^\top)^2\big) = m\left(1 + \frac{2}{m}\right) + m(m-1)\frac{1}{m} = 2m + 1.$$

Thus $\mathbb{E}B_0^v = \frac{1}{m}(2m+1) = 2 + m^{-1}$.

The computation for S_0^z is the same. Indeed,

$$S_0^z = \frac{1}{m}\langle Mb, MM^\top Mb\rangle = \frac{1}{m}b^\top M^\top MM^\top Mb = \frac{1}{m}b^\top (M^\top M)^2 b.$$

Conditioning on M and using $\mathbb{E}[bb^\top] = I_m$, we get

$$\mathbb{E}_b[S_0^z \mid M] = \frac{1}{m}\operatorname{tr}\big((M^\top M)^2\big) = \frac{1}{m}\operatorname{tr}\big((MM^\top)^2\big).$$

Hence $\mathbb{E}S_0^z = 2 + m^{-1}$. Finally, $T_0^z = \frac{1}{m}\langle MM^\top a, MM^\top Mb\rangle$. Conditioning on M, b, this is linear in a, and a is centered independent of M, b. Thus $\mathbb{E}T_0^z = 0$.

We also need $\mathbb{E}\big(G_{0,0}^{xv}\big)^2 = m^{-1}$. Indeed, conditioning on M, b, $G_{0,0}^{xv} = \frac{1}{m}\langle a, Mb\rangle$ is a centered Gaussian random variable in a with conditional variance $\frac{1}{m^2}\|Mb\|_2^2$. Therefore

$$\mathbb{E}\big[(G_{0,0}^{xv})^2 \mid M, b\big] = \frac{1}{m^2}\|Mb\|_2^2.$$

Taking expectation and using $\mathbb{E}\|Mb\|_2^2 = m$ gives $\mathbb{E}\big(G_{0,0}^{xv}\big)^2 = m^{-1}$.

Since

$$K_{0,1}^{xx} = K_{0,0}^{xx} - \rho_0 G_{0,0}^{xv}, \quad D_{0,0}^{x} = -\rho_0 G_{0,0}^{xv}, \quad K_{1,0}^{vv} = K_{0,0}^{vv} - \rho_0 G_{0,0}^{xv}, \quad \rho_0 = \xi_0^2 G_{0,0}^{xv} - \xi_0 y(\xi_0),$$

we have

$$\mathbb{E}[\rho_0 G_{0,0}^{xv}] = \xi_0^2 \mathbb{E}\big(G_{0,0}^{xv}\big)^2 - \xi_0 y(\xi_0)\mathbb{E}G_{0,0}^{xv} = \xi_0^2 m^{-1}.$$

Thus

$$\mathbb{E}K_{0,1}^{xx} = 1 - \xi_0^2 m^{-1}, \quad \mathbb{E}D_{0,0}^{x} = -\xi_0^2 m^{-1}, \quad \mathbb{E}K_{1,0}^{vv} = 1 - \xi_0^2 m^{-1}.$$

Finally, the exact zero initialization bias follows even more directly by conditioning $\mathbb{E}\left[g_{\theta_0}^{(m)} \,\middle|\, u_0, W_0\right] = \frac{\xi_0}{m}\mathbb{E}[\langle a_0, W_0 u_0\rangle \,|\, u_0, W_0]$. Since a_0 is centered and independent of u_0, W_0, $\mathbb{E}[a_0 \mid u_0, W_0] = 0$, and therefore

$$\mathbb{E}\left[g_{\theta_0}^{(m)} \,\middle|\, u_0, W_0\right] = \frac{\xi_0}{m}\,\langle 0, W_0 u_0\rangle = 0.$$

Hence $\mathbb{E}g_{\theta_0}^{(m)} = 0$. The deterministic infinite-width initialization output is also 0, so $\mathbb{E}g_{\theta_0}^{(m)} = 0 = \bar{g}_{\theta_0}$.

Step 2: moment hypothesis. By Lemma F.11, for every fixed $q \geq 2$, every coordinate S of every word-closed state $Y_t^{\mathrm{lin},(m)}$, $0 \leq t \leq T+1$, satisfies

$$\|S - \mathbb{E}S\|_{L^q} \leq C_{T,q} m^{-1/2}, \qquad \|S\|_{L^q} + |\mathbb{E}S| \leq C_{T,q}.$$

This is exactly the centered-fluctuation and bounded-moment hypothesis needed in Proposition E.2. In particular, whenever P is a fixed polynomial in coordinates of $Y_t^{\mathrm{lin},(m)}$, Lemma E.1 gives

$$\mathbb{E}P(Y_t^{\mathrm{lin},(m)}) = P(\mathbb{E}Y_t^{\mathrm{lin},(m)}) + O(m^{-1}).$$

Step 3: exact polynomial closure. By Lemma F.9, for every $0 \leq t \leq T$ there is a fixed polynomial map P_t^{lin}, independent of m, such that $Y_{t+1}^{\mathrm{lin},(m)} = P_t^{\mathrm{lin}}\big(Y_t^{\mathrm{lin},(m)}\big)$ exactly. Hence the closure hypothesis of Proposition E.2 holds with $R_{t,i}^{(m)} = 0$. The same lemma also shows that every $S \in \mathcal{S}_t^{\mathrm{lin}}$ is either a coordinate of $Y_t^{\mathrm{lin},(m)}$ or a fixed polynomial function of its coordinates.

Step 4: expectation induction. The base estimate from Step 1, the moment bounds from Step 2, and the exact polynomial closure from Step 3 verify all hypotheses of Proposition E.2 with $R_{t,i}^{(m)} = 0$. For completeness, we spell out the induction argument.

Assume that, for some fixed t, $\mathbb{E}Y_{t,i}^{\mathrm{lin},(m)} = Y_{t,i}^{\mathrm{lin},\infty} + O(m^{-1})$ for every coordinate i. Since $Y_{t+1,j}^{\mathrm{lin},(m)} = P_{t,j}^{\mathrm{lin}}(Y_t^{\mathrm{lin},(m)})$, we have $\mathbb{E}Y_{t+1,j}^{\mathrm{lin},(m)} = \mathbb{E}P_{t,j}^{\mathrm{lin}}(Y_t^{\mathrm{lin},(m)})$. By Lemma E.1 and the moment bounds from Step 2,

$$\mathbb{E}P_{t,j}^{\mathrm{lin}}(Y_t^{\mathrm{lin},(m)}) = P_{t,j}^{\mathrm{lin}}(\mathbb{E}Y_t^{\mathrm{lin},(m)}) + O(m^{-1}).$$

By the induction hypothesis, $\mathbb{E}Y_t^{\mathrm{lin},(m)} = Y_t^{\mathrm{lin},\infty} + O(m^{-1})$ coordinatewisely. Since $P_{t,j}^{\mathrm{lin}}$ is a fixed polynomial and the deterministic trajectory is bounded on the fixed horizon T,

$$P_{t,j}^{\mathrm{lin}}(\mathbb{E}Y_t^{\mathrm{lin},(m)}) = P_{t,j}^{\mathrm{lin}}(Y_t^{\mathrm{lin},\infty}) + O(m^{-1}).$$

The deterministic limit trajectory is defined by $Y_{t+1,j}^{\mathrm{lin},\infty} = P_{t,j}^{\mathrm{lin}}(Y_t^{\mathrm{lin},\infty})$. Therefore

$$\mathbb{E}Y_{t+1,j}^{\mathrm{lin},(m)} = Y_{t+1,j}^{\mathrm{lin},\infty} + O(m^{-1}).$$

Together with the base estimate from Step 1, this proves $\mathbb{E}S = S^\infty + O(m^{-1})$ for every coordinate S of $Y_t^{\mathrm{lin},(m)}$, for every fixed $0 \leq t \leq T+1$.

Every observable in $\mathcal{S}_t^{\mathrm{lin}}$ is either a coordinate of $Y_t^{\mathrm{lin},(m)}$, or a fixed polynomial function of coordinates of $Y_t^{\mathrm{lin},(m)}$. Thus, applying Lemma E.1 once more if necessary, $\mathbb{E}S = S^\infty + O(m^{-1})$, $S \in \mathcal{S}_t^{\mathrm{lin}}$.

The deterministic bound $|S^\infty| \leq C_T$ follows from the uniform expectation bounds in Step 2. Indeed, $|\mathbb{E}S| \leq \|S\|_{L^1} \leq \|S\|_{L^q} \leq C_{T,q}$ for any fixed $q \geq 2$. Passing to the limit along $S^\infty = \lim_{m\to\infty} \mathbb{E}S$ gives $|S^\infty| \leq C_T$.

Finally, for any input ξ, $g_{\theta_t}^{(m)}(\xi) = \xi\, G_{t,t}^{xv}$, $G_{t,t}^{xv} = \frac{1}{m}\langle z_t, a_t\rangle$. Since $G_{t,t}^{xv}$ is a scalar coordinate of the word-closed state, $\mathbb{E}G_{t,t}^{xv} = (G_{t,t}^{xv})^\infty + O(m^{-1})$. Therefore

$$\mathbb{E}g_{\theta_t}^{(m)}(\xi) = \xi\, \mathbb{E}G_{t,t}^{xv} = \xi\, (G_{t,t}^{xv})^\infty + O(m^{-1}) = \bar{g}_{\theta_t}(\xi) + O(m^{-1}).$$

At the training input $\xi = \xi_t$, this gives $\mathbb{E}g_{\theta_t}^{(m)} = \bar{g}_{\theta_t} + O(m^{-1})$.

At $t = 0$, Step 1 showed $\mathbb{E}g_{\theta_0}^{(m)} = 0 = \bar{g}_{\theta_0}$. Hence $\mathbb{E}g_{\theta_0}^{(m)} - \bar{g}_{\theta_0} = 0$. This completes the word-closed scalar-state expectation induction. □

Lemma F.15 (Agreement with the Tensor Program output limit)**.** *Fix a finite horizon $T < \infty$. Consider the scalar-input three-layer linear μP model under the recursion in Section 3, with a deterministic training sequence, or conditionally on a training sequence independent of the Gaussian initialization. Let $g_{\theta_t}^{\mathrm{TP}}(\xi)$ denote the standard Tensor Program infinite-width output limit obtained by applying the Tensor Program theorem to the corresponding compiled finite-time μP computation.*

Let $\bar{g}^{\mathrm{wc}}_{\theta_t}(\xi)$ denote the word-closed deterministic comparison output constructed in Appendix F: $\bar{g}^{\mathrm{wc}}_{\theta_t}(\xi) := \xi\,(G^{xv}_{t,t})^{\infty}$, where $(G^{xv}_{t,t})^{\infty}$ is the corresponding coordinate of the deterministic scalar state $Y_t^{\mathrm{lin},\infty}$. Then, for every fixed $0 \le t \le T$ and every fixed bounded input ξ,

$$g^{\mathrm{TP}}_{\theta_t}(\xi) = \bar{g}^{\mathrm{wc}}_{\theta_t}(\xi).$$

Consequently, in the three-layer linear setting, the Tensor Program output limit and the word-closed deterministic comparison output coincide.

Proof. In the three-layer linear model, $g^{(m)}_{\theta_t}(\xi) = \xi\,\frac{1}{m}\langle m\mathbf{v}_t, z_t\rangle = \xi\,G^{xv}_{t,t}$.

By the Tensor Program theorem applied to the compiled finite-time μP computation, the width-m output converges almost surely to its Tensor Program limit:

$$g^{(m)}_{\theta_t}(\xi) \xrightarrow[m\to\infty]{a.s.} g^{\mathrm{TP}}_{\theta_t}(\xi), \quad g^{(m)}_{\theta_t}(\xi) \xrightarrow[m\to\infty]{\mathbb{P}} g^{\mathrm{TP}}_{\theta_t}(\xi).$$

On the other hand, Appendix F gives convergence of the same finite-width output to the word-closed deterministic comparison output. Indeed, Theorem F.8, applied to the scalar observable $G^{xv}_{t,t}$, gives $G^{xv}_{t,t} - \mathbb{E}G^{xv}_{t,t} \xrightarrow[m\to\infty]{\mathbb{P}} 0$. Moreover, Proposition F.14 gives $\mathbb{E}G^{xv}_{t,t} = (G^{xv}_{t,t})^{\infty} + O(m^{-1})$. Therefore, $G^{xv}_{t,t} \xrightarrow[m\to\infty]{\mathbb{P}} (G^{xv}_{t,t})^{\infty}$. Multiplying by the fixed scalar ξ, we obtain

$$g^{(m)}_{\theta_t}(\xi) = \xi\,G^{xv}_{t,t} \xrightarrow[m\to\infty]{\mathbb{P}} \xi\,(G^{xv}_{t,t})^{\infty} = \bar{g}^{\mathrm{wc}}_{\theta_t}(\xi).$$

Thus the same sequence $g^{(m)}_{\theta_t}(\xi)$ converges in probability to both $g^{\mathrm{TP}}_{\theta_t}(\xi)$ and $\bar{g}^{\mathrm{wc}}_{\theta_t}(\xi)$. Since limits in probability are unique, the two limits must agree: $g^{\mathrm{TP}}_{\theta_t}(\xi) = \bar{g}^{\mathrm{wc}}_{\theta_t}(\xi)$. This proves the claim. □

Theorem F.16 (Simultaneous deterministic-limit concentration for GD)**.** *Fix $T < \infty$. Then there exists a constant $C_T \ge 1$, depending only on T and M_T^{data}, such that for every $\delta \in (0, 1/2]$, with probability at least $1-\delta$, the following hold simultaneously for every $0 \le t \le T$ and every $S \in \mathcal{S}_t^{\mathrm{lin}}$:*

$$|S - \mu_S^{(m)}| \le C_T m^{-1/2} L_\delta^{\alpha_t}, \quad |\mu_S^{(m)} - S^{\infty}| \le C_T m^{-1},$$

and hence

$$|S - S^{\infty}| \le C_T m^{-1/2} L_\delta^{\alpha_t}.$$

Moreover, there is

$$|S^{\infty}| \le C_T, \quad |S| + |S^{\infty}| \le C_T L_\delta^{\alpha_t}, \quad |\mathbb{E}g^{(m)}_{\theta_t} - \bar{g}_{\theta_t}| \le C_T m^{-1}, \quad |g^{(m)}_{\theta_t} - \bar{g}_{\theta_t}| \le C_T m^{-1/2} L_\delta^{\alpha_t}.$$

Remark F.17 (Deterministic limit used in the three-layer linear theorem). In the three-layer linear setting, the deterministic comparison trajectory used in the proof is the word-closed trajectory

$$\bar{g}^{\mathrm{wc}}_{\theta_t}(\xi) := \xi\,(G^{xv}_{t,t})^{\infty}.$$

The concentration theorem is first proved with respect to this word-closed trajectory. Lemma F.15 then identifies $\bar{g}^{\mathrm{wc}}_{\theta_t}$ with the standard Tensor Program infinite-width output limit, so the same deterministic limit may be written simply as $\bar{g}_{\theta_t}$ in the main theorem.

Proof. The centered estimate $|S - \mu_S^{(m)}| \le C_T m^{-1/2} L_\delta^{\alpha_t}$ is Theorem F.8, and the expectation estimate $|\mu_S^{(m)} - S^{\infty}| \le C_T m^{-1}$ is Proposition F.14. Since $m^{-1} \le m^{-1/2} L_\delta^{\alpha_t}$ $(m \ge 1)$, the bias term is absorbed into the same high-probability scale, giving

$$|S - S^{\infty}| \le |S - \mu_S^{(m)}| + |\mu_S^{(m)} - S^{\infty}| \le C_T m^{-1/2} L_\delta^{\alpha_t}.$$

The bounds on $|S^{\infty}|$ and $|S| + |S^{\infty}|$ follow after increasing C_T. The output estimates are the special case $S = g^{(m)}_{\theta_t}$. □

Corollary F.18 (Three-layer mean-square error and bias)**.** *In the three-layer linear setting of Theorem 4.1, for fixed T and some C_T independent of m,*

$$\mathbb{E}\max_{0\le t\le T} |g^{(m)}_{\theta_t} - \bar{g}_{\theta_t}|^2 \le \frac{C_T}{m}, \qquad \max_{0\le t\le T} |\mathbb{E}g^{(m)}_{\theta_t} - \bar{g}_{\theta_t}| \le \frac{C_T}{m}.$$

Proof of Corollary F.18. Lemmas F.7 and F.10 give $\mathrm{Var}(g_{\theta_t}^{(m)}) \le C_T/m$. Proposition F.14 gives the bias bound. Consequently,

$$\mathbb{E} \max_{0\le t\le T} |g_{\theta_t}^{(m)} - \bar{g}_{\theta_t}|^2 \le \sum_{t=0}^{T} \left(\mathrm{Var}(g_{\theta_t}^{(m)}) + |\mathbb{E} g_{\theta_t}^{(m)} - \bar{g}_{\theta_t}|^2 \right) \le \frac{C_T}{m},$$

after enlarging C_T. For random training sequences, all expectations are conditional on the sequence, as in Theorem 4.1. □

G Proof of the KL-Regularized Domain-Weight Stability Theorem

Loss normalization. In this appendix we use the empirical domain-loss convention

$$\ell_{i,t}^{(m)} := \frac{1}{n_i} \sum_{j=1}^{n_i} \big(g_t^{(m)}(\xi_{i,j}) - y_{i,j}\big)^2.$$

This differs by a factor of 2 from the single-sample loss convention $\frac{1}{2}(g-y)^2$ used in the main theoretical setup. Consequently, the adaptive coefficients $A_t^{(m)}$ and $A_t^{(\infty)}$ below contain a factor 2. If empirical domain losses are instead defined with the factor $1/2$, then this factor 2 should be removed from both coefficients. All concentration bounds are unchanged up to constants.

Setting. Let $\mathcal{D} := \bigcup_{i=1}^K \mathcal{D}_i$, $\mathcal{D}_i := \{(\xi_{i,j}, y_{i,j})\}_{j=1}^{n_i}$, $N := \sum_{i=1}^K n_i$, and assume $|\xi_{i,j}| \le 1$, $|y_{i,j}| \le Y$. We consider a two-layer scalar-input model with either $\sigma(z) = z$ or $\sigma(z) = z^2$.

Let $U_{t,\alpha} \in \mathbb{R}$ denote the hidden weight of neuron α, let $X_{t,\alpha}(\xi) := \sigma(U_{t,\alpha}\xi)$, let $a_{t,\alpha}$ be the actual output weight, and define the μP base output variable $V_{t,\alpha} := m a_{t,\alpha}$. The finite-width output is $g_t^{(m)}(\xi) := g_{\theta_t}^{(m)}(\xi) = m^{-1} \sum_{\alpha=1}^m V_{t,\alpha} X_{t,\alpha}(\xi)$.

We use the μP learning-rate multipliers $\eta_a = m^{-1}$ and $\eta_U = m$, with base learning rate equal to 1. Equivalently, a fixed scalar base learning rate can be inserted everywhere and absorbed into the constants below.

For a single example, write $\chi_t := g_t^{(m)}(\xi_t) - y(\xi_t)$. Gradient descent gives

$$V_{t+1,\alpha} = V_{t,\alpha} - \chi_t X_{t,\alpha}(\xi_t), \quad U_{t+1,\alpha} = U_{t,\alpha} - m\chi_t a_{t,\alpha} \sigma'(U_{t,\alpha}\xi_t)\xi_t = U_{t,\alpha} - \chi_t V_{t,\alpha} \sigma'(U_{t,\alpha}\xi_t)\xi_t.$$

Thus, for $\sigma(z) = z$, $X_{t,\alpha}(\xi_t) = \xi_t U_{t,\alpha}$, $\sigma'(U_{t,\alpha}\xi_t) = 1$, and the single-example recursion becomes

$$V_{t+1,\alpha} = V_{t,\alpha} - \chi_t \xi_t U_{t,\alpha}, \qquad U_{t+1,\alpha} = U_{t,\alpha} - \chi_t \xi_t V_{t,\alpha}.$$

For $\sigma(z) = z^2$, $X_{t,\alpha}(\xi_t) = \xi_t^2 U_{t,\alpha}^2$, $\sigma'(U_{t,\alpha}\xi_t) = 2U_{t,\alpha}\xi_t$, and the single-example recursion becomes

$$V_{t+1,\alpha} = V_{t,\alpha} - \chi_t \xi_t^2 U_{t,\alpha}^2, \qquad U_{t+1,\alpha} = U_{t,\alpha} - 2\chi_t \xi_t^2 V_{t,\alpha} U_{t,\alpha}.$$

These are the two-block μP recursions used below.

For each empirical example, define the residual $r_{i,j,t}^{(m)} := g_t^{(m)}(\xi_{i,j}) - y_{i,j}$. For each domain, define the empirical square loss $\ell_{i,t}^{(m)} := \frac{1}{n_i} \sum_{j=1}^{n_i} \left(r_{i,j,t}^{(m)} \right)^2$. The KL-regularized domain weights are $\pi_{i,t}^{(m)} = \frac{\exp(\lambda \ell_{i,t}^{(m)})}{\sum_{k=1}^K \exp(\lambda \ell_{k,t}^{(m)})}$, where $\lambda > 0$.

For the linear activation, define $A_t^{(m)} := 2 \sum_{i=1}^K \frac{\pi_{i,t}^{(m)}}{n_i} \sum_{j=1}^{n_i} r_{i,j,t}^{(m)} \xi_{i,j}$. Then the KL-reweighted finite dynamics are

$$V_{t+1,\alpha}^{(m)} = V_{t,\alpha}^{(m)} - A_t^{(m)} U_{t,\alpha}^{(m)}, \quad U_{t+1,\alpha}^{(m)} = U_{t,\alpha}^{(m)} - A_t^{(m)} V_{t,\alpha}^{(m)}.$$

For the quadratic activation, define $A_t^{(m)} := 2 \sum_{i=1}^K \frac{\pi_{i,t}^{(m)}}{n_i} \sum_{j=1}^{n_i} r_{i,j,t}^{(m)} \xi_{i,j}^2$. Then the full-batch KL-reweighted finite dynamics are

$$V_{t+1,\alpha}^{(m)} = V_{t,\alpha}^{(m)} - A_t^{(m)} \left(U_{t,\alpha}^{(m)} \right)^2, \quad U_{t+1,\alpha}^{(m)} = U_{t,\alpha}^{(m)} - 2A_t^{(m)} V_{t,\alpha}^{(m)} U_{t,\alpha}^{(m)}.$$

This factor 2 is due to the loss-normalization convention stated above.

For $\sigma(z) = z$, assume the initialization $(V_{0,\alpha}, U_{0,\alpha})_{\alpha=1}^m$ is i.i.d. centered Gaussian with deterministic covariance, as in section C. For $\sigma(z) = z^2$, assume $U_{0,\alpha} \sim \mathcal{N}(0,1)$, $V_{0,\alpha} \sim \mathcal{N}(0,1)$, with all coordinates mutually independent, as in section D.

Lemma G.1 (KL inner maximization and log-sum-exp objective)**.** *For fixed losses $\ell = (\ell_1, \dots, \ell_K)$ and $\mathbf{u} = (1/K, \dots, 1/K)$, one has*

$$\max_{\pi \in \Delta^K} \left\{ \sum_{i=1}^K \pi_i \ell_i - \frac{1}{\lambda} D_{\mathrm{KL}}(\pi \| \mathbf{u}) \right\} = \frac{1}{\lambda} \log \left(\frac{1}{K} \sum_{i=1}^K e^{\lambda \ell_i} \right),$$

and the unique optimizer is $\pi_i^*(\ell) = \frac{e^{\lambda \ell_i}}{\sum_{k=1}^K e^{\lambda \ell_k}}$. Consequently, for $\mathcal{R}_m(\theta) := \frac{1}{\lambda} \log \left(\frac{1}{K} \sum_{i=1}^K e^{\lambda \ell_i^{(m)}(\theta)} \right)$, we have $\nabla_\theta \mathcal{R}_m(\theta) = \sum_{i=1}^K \pi_i^{(m)}(\theta) \nabla_\theta \ell_i^{(m)}(\theta)$.

Proof. Define

$$Z(\ell) := \sum_{i=1}^K u_i e^{\lambda \ell_i} = \frac{1}{K} \sum_{i=1}^K e^{\lambda \ell_i}, \quad \pi_i^*(\ell) := \frac{u_i e^{\lambda \ell_i}}{Z(\ell)}.$$

Then $\log \frac{\pi_i^*(\ell)}{u_i} = \lambda \ell_i - \log Z(\ell)$. Therefore, for any $\pi \in \Delta^K$,

$$D_{\mathrm{KL}}(\pi \| \pi^*) = \sum_i \pi_i \log \frac{\pi_i}{\pi_i^*} = D_{\mathrm{KL}}(\pi \| \mathbf{u}) - \lambda \sum_i \pi_i \ell_i + \log Z(\ell).$$

Rearranging gives

$$\sum_i \pi_i \ell_i - \frac{1}{\lambda} D_{\mathrm{KL}}(\pi \| \mathbf{u}) = \frac{1}{\lambda} \log Z(\ell) - \frac{1}{\lambda} D_{\mathrm{KL}}(\pi \| \pi^*).$$

Since $D_{\mathrm{KL}}(\pi \| \pi^*) \geq 0$, the maximum is attained at $\pi = \pi^*$, and the value is $\frac{1}{\lambda} \log Z(\ell) = \frac{1}{\lambda} \log \left(\frac{1}{K} \sum_{i=1}^K e^{\lambda \ell_i} \right)$. Differentiating the log-sum-exp objective gives

$$\nabla_\theta \mathcal{R}_m(\theta) = \sum_{i=1}^K \frac{e^{\lambda \ell_i^{(m)}(\theta)}}{\sum_k e^{\lambda \ell_k^{(m)}(\theta)}} \nabla_\theta \ell_i^{(m)}(\theta),$$

which is the claimed formula. □

Lemma G.2 (Softmax Lipschitz bound)**.** *Let $\Phi_\lambda(s)_i := \frac{e^{\lambda s_i}}{\sum_{k=1}^K e^{\lambda s_k}}$. Then*

$$\|\Phi_\lambda(s) - \Phi_\lambda(r)\|_\infty \leq \lambda \|s - r\|_\infty.$$

Proof. For every i, j,

$$\partial_j \Phi_\lambda(s)_i = \lambda \Phi_\lambda(s)_i \left(\mathbf{1}_{\{i=j\}} - \Phi_\lambda(s)_j \right).$$

Therefore,

$$\sum_{j=1}^K |\partial_j \Phi_\lambda(s)_i| = 2\lambda \Phi_\lambda(s)_i \left(1 - \Phi_\lambda(s)_i\right) \leq \lambda.$$

The mean-value theorem gives

$$|\Phi_\lambda(s)_i - \Phi_\lambda(r)_i| \leq \lambda \|s - r\|_\infty$$

for every i. Taking the maximum over i proves the claim. □

Reference dynamics. For both activations, initialize the reference process by the same particle initialization: $U_{0,\alpha}^{\mathrm{ref}} := U_{0,\alpha}^{(m)}$, $V_{0,\alpha}^{\mathrm{ref}} := V_{0,\alpha}^{(m)}$. The reference dynamics use the deterministic population coefficient $A_t^{(\infty)}$ defined below. We denote the corresponding infinite-width output by $\bar{g}_{\theta_t}(\xi)$ and abbreviate it by $\bar{g}_t(\xi)$ within this subsection.

For $\sigma(z) = z$, define the deterministic reference output $\bar{g}_t(\xi) := \xi\,\mathbb{E}[V_{t,\alpha}^{\mathrm{ref}} U_{t,\alpha}^{\mathrm{ref}}]$. For $\sigma(z) = z^2$, define $\bar{g}_t(\xi) := \xi^2 \mathbb{E}[V_{t,\alpha}^{\mathrm{ref}} (U_{t,\alpha}^{\mathrm{ref}})^2]$. Also define

$$r_{i,j,t}^{(\infty)} := \bar{g}_t(\xi_{i,j}) - y_{i,j}, \quad \ell_{i,t}^{(\infty)} := \frac{1}{n_i} \sum_{j=1}^{n_i} \left(r_{i,j,t}^{(\infty)} \right)^2, \quad \pi_{i,t}^{(\infty)} = \frac{\exp(\lambda \ell_{i,t}^{(\infty)})}{\sum_{k=1}^{K} \exp(\lambda \ell_{k,t}^{(\infty)})}.$$

Set $B_T := \max_{i \in [K]} \max_{0 \le t \le T} \sqrt{\ell_{i,t}^{(\infty)}}$.

Define $A_t^{(\infty)}$ by the same formula as $A_t^{(m)}$, replacing $\pi_{i,t}^{(m)}$, $r_{i,j,t}^{(m)}$ by $\pi_{i,t}^{(\infty)}$, $r_{i,j,t}^{(\infty)}$. Thus, for $\sigma(z) = z$, $A_t^{(\infty)} = 2 \sum_{i=1}^{K} \frac{\pi_{i,t}^{(\infty)}}{n_i} \sum_{j=1}^{n_i} r_{i,j,t}^{(\infty)} \xi_{i,j}$, and for $\sigma(z) = z^2$, $A_t^{(\infty)} = 2 \sum_{i=1}^{K} \frac{\pi_{i,t}^{(\infty)}}{n_i} \sum_{j=1}^{n_i} r_{i,j,t}^{(\infty)} \xi_{i,j}^2$.

The reference dynamics are obtained from the finite dynamics by replacing $A_t^{(m)}$ with $A_t^{(\infty)}$. Thus, in the linear case,

$$V_{t+1,\alpha}^{\mathrm{ref}} = V_{t,\alpha}^{\mathrm{ref}} - A_t^{(\infty)} U_{t,\alpha}^{\mathrm{ref}}, \quad U_{t+1,\alpha}^{\mathrm{ref}} = U_{t,\alpha}^{\mathrm{ref}} - A_t^{(\infty)} V_{t,\alpha}^{\mathrm{ref}}.$$

In the quadratic case,

$$V_{t+1,\alpha}^{\mathrm{ref}} = V_{t,\alpha}^{\mathrm{ref}} - A_t^{(\infty)} \left(U_{t,\alpha}^{\mathrm{ref}} \right)^2, \quad U_{t+1,\alpha}^{\mathrm{ref}} = U_{t,\alpha}^{\mathrm{ref}} - 2 A_t^{(\infty)} V_{t,\alpha}^{\mathrm{ref}} U_{t,\alpha}^{\mathrm{ref}}.$$

Lemma G.3 (Output error implies loss, weight, and coefficient error)**.** *Define*

$$\Delta_t := \max_{i \in [K]} \max_{1 \le j \le n_i} \left| g_t^{(m)}(\xi_{i,j}) - \bar{g}_t(\xi_{i,j}) \right|, \quad E_t := \|\ell_t^{(m)} - \ell_t^{(\infty)}\|_\infty, \quad P_t := \|\pi_t^{(m)} - \pi_t^{(\infty)}\|_\infty,$$

and $\alpha_t := |A_t^{(m)} - A_t^{(\infty)}|$. *Then, for every* $0 \le t \le T$,

$$E_t \le 2 B_T \Delta_t + \Delta_t^2, \quad P_t \le \lambda (2 B_T \Delta_t + \Delta_t^2), \quad \alpha_t \le C_{\lambda, K, B_T} (\Delta_t + \Delta_t^2),$$

where C_{λ, K, B_T} *depends only on* λ, K, *and* B_T.

Proof. For every i, j, t, define $d_{i,j,t} := g_t^{(m)}(\xi_{i,j}) - \bar{g}_t(\xi_{i,j})$. Then $r_{i,j,t}^{(m)} = r_{i,j,t}^{(\infty)} + d_{i,j,t}$. Therefore,

$$\left(r_{i,j,t}^{(m)} \right)^2 - \left(r_{i,j,t}^{(\infty)} \right)^2 = d_{i,j,t}^2 + 2 d_{i,j,t} r_{i,j,t}^{(\infty)}.$$

Averaging over domain i, we get

$$\ell_{i,t}^{(m)} - \ell_{i,t}^{(\infty)} = \frac{1}{n_i} \sum_j d_{i,j,t}^2 + \frac{2}{n_i} \sum_j d_{i,j,t} r_{i,j,t}^{(\infty)}.$$

Taking absolute values and applying Cauchy–Schwarz gives

$$\left| \ell_{i,t}^{(m)} - \ell_{i,t}^{(\infty)} \right| \le \Delta_t^2 + 2 \Delta_t \left(\frac{1}{n_i} \sum_j \left(r_{i,j,t}^{(\infty)} \right)^2 \right)^{1/2}.$$

Since

$$\left(\frac{1}{n_i} \sum_j \left(r_{i,j,t}^{(\infty)} \right)^2 \right)^{1/2} = \sqrt{\ell_{i,t}^{(\infty)}} \le B_T,$$

we obtain $E_t \le 2 B_T \Delta_t + \Delta_t^2$. By Lemma G.2,

$$P_t \le \lambda E_t \le \lambda (2 B_T \Delta_t + \Delta_t^2).$$

It remains to bound α_t. In both activation cases,

$$A_t^{(m)} = 2\sum_{i=1}^{K} \frac{\pi_{i,t}^{(m)}}{n_i} \sum_{j=1}^{n_i} r_{i,j,t}^{(m)} \psi(\xi_{i,j}),$$

where $\psi(\xi) = \xi$for $\sigma(z) = z$, $\psi(\xi) = \xi^2$ for $\sigma(z) = z^2$. Since $|\xi_{i,j}| \le 1$, we have $|\psi(\xi_{i,j})| \le 1$. Add and subtract the mixed term with $\pi_{i,t}^{(m)} r_{i,j,t}^{(\infty)}$. Then

$$A_t^{(m)} - A_t^{(\infty)} = 2\sum_i \frac{\pi_{i,t}^{(m)}}{n_i} \sum_j \left(r_{i,j,t}^{(m)} - r_{i,j,t}^{(\infty)} \right) \psi(\xi_{i,j}) + 2\sum_i \frac{\pi_{i,t}^{(m)} - \pi_{i,t}^{(\infty)}}{n_i} \sum_j r_{i,j,t}^{(\infty)} \psi(\xi_{i,j}).$$

Hence

$$\alpha_t \le 2\sum_i \pi_{i,t}^{(m)} \Delta_t + 2\sum_i \left| \pi_{i,t}^{(m)} - \pi_{i,t}^{(\infty)} \right| \frac{1}{n_i} \sum_j \left| r_{i,j,t}^{(\infty)} \right|.$$

The first term equals $2\Delta_t$. For the second term,

$$\frac{1}{n_i} \sum_j \left| r_{i,j,t}^{(\infty)} \right| \le \sqrt{\ell_{i,t}^{(\infty)}} \le B_T.$$

Thus

$$\alpha_t \le 2\Delta_t + 2KB_T P_t.$$

Using the previous bound on P_t, we obtain

$$\alpha_t \le 2\Delta_t + 2KB_T\lambda(2B_T\Delta_t + \Delta_t^2) \le C_{\lambda,K,B_T}(\Delta_t + \Delta_t^2).$$

□

Lemma G.4 (Linear KL-adaptive bootstrap; one-hidden-layer linear variant)**.** *Assume $\sigma(z) = z$. Define*

$$\varepsilon_t := \max_{\alpha \le m} \left(|V_{t,\alpha}^{(m)} - V_{t,\alpha}^{\mathrm{ref}}| + |U_{t,\alpha}^{(m)} - U_{t,\alpha}^{\mathrm{ref}}| \right),$$

and define the empirical-support output error

$$\Delta_t := \max_{i \in [K]} \max_{1 \le j \le n_i} \left| g_t^{(m)}(\xi_{i,j}) - \bar{g}_t(\xi_{i,j}) \right|.$$

Let

$$L := 1 + \log \frac{2mN(T+1)}{\delta}.$$

There is an event of probability at least $1 - \delta$, obtained by union bounding the reference Gaussian maximum and empirical-product concentration bounds over all $0 \le t \le T$ and all empirical inputs $\{\xi_{i,j}\}$, on which

$$\varepsilon_0 = 0, \qquad \Delta_0 \le C_0 m^{-1/2} L,$$

and, for every $0 \le t < T$,

$$\varepsilon_{t+1} \le C_t \left[(1 + \Delta_t + \Delta_t^2)\varepsilon_t + (\Delta_t + \Delta_t^2) L^{1/2} \right],$$

$$\Delta_t \le C_t m^{-1/2} L + C_t L^{1/2} \varepsilon_t + \varepsilon_t^2.$$

Proof. The proof follows the perturbation argument of Lemma C.6, with the scalar residual coefficient replaced by the KL-adaptive coefficient $A_t^{(m)}$.

The finite and reference recursions are

$$V_{t+1,\alpha}^{(m)} = V_{t,\alpha}^{(m)} - A_t^{(m)} U_{t,\alpha}^{(m)}, \qquad U_{t+1,\alpha}^{(m)} = U_{t,\alpha}^{(m)} - A_t^{(m)} V_{t,\alpha}^{(m)},$$

$$V_{t+1,\alpha}^{\mathrm{ref}} = V_{t,\alpha}^{\mathrm{ref}} - A_t^{(\infty)} U_{t,\alpha}^{\mathrm{ref}}, \qquad U_{t+1,\alpha}^{\mathrm{ref}} = U_{t,\alpha}^{\mathrm{ref}} - A_t^{(\infty)} V_{t,\alpha}^{\mathrm{ref}}.$$

Subtracting gives

$$V_{t+1,\alpha}^{(m)} - V_{t+1,\alpha}^{\mathrm{ref}} = V_{t,\alpha}^{(m)} - V_{t,\alpha}^{\mathrm{ref}} - A_t^{(\infty)} \left(U_{t,\alpha}^{(m)} - U_{t,\alpha}^{\mathrm{ref}} \right) + \left(A_t^{(\infty)} - A_t^{(m)} \right) U_{t,\alpha}^{(m)}.$$

The same identity holds for the U-coordinate after interchanging U and V.

On the reference maximum event from Lemma C.2, $|U^{\mathrm{ref}}_{t,\alpha}| + |V^{\mathrm{ref}}_{t,\alpha}| \le C_t L^{1/2}$. Therefore

$$|U^{(m)}_{t,\alpha}| + |V^{(m)}_{t,\alpha}| \le C_t L^{1/2} + \varepsilon_t.$$

Since $A^{(\infty)}_t$ is deterministic and bounded for fixed $t \le T$, and since $|A^{(m)}_t - A^{(\infty)}_t| \le C_{\lambda,K,B_T}(\Delta_t + \Delta_t^2)$, by Lemma G.3, we get

$$\varepsilon_{t+1} \le C_t \left[(1 + \Delta_t + \Delta_t^2)\varepsilon_t + (\Delta_t + \Delta_t^2) L^{1/2} \right].$$

By the output decomposition in Lemma C.5,

$$g^{(m)}_t(\xi) - \bar{g}_t(\xi) = \xi \left[\frac{1}{m} \sum_{\alpha=1}^{m} \left(V^{\mathrm{ref}}_{t,\alpha} U^{\mathrm{ref}}_{t,\alpha} - \mathbb{E}[V^{\mathrm{ref}}_{t,\alpha} U^{\mathrm{ref}}_{t,\alpha}] \right) \right] + \mathrm{Err}_t(\xi).$$

The empirical average is bounded by $C_t m^{-1/2} L$ uniformly over the finite support. The error satisfies

$$|\,\mathrm{Err}_t(\xi)| \le C_t L^{1/2} \varepsilon_t + \varepsilon_t^2.$$

Taking the maximum over the finite support gives

$$\Delta_t \le C_t m^{-1/2} L + C_t L^{1/2} \varepsilon_t + \varepsilon_t^2.$$

At initialization, the finite and reference particles coincide, so $\varepsilon_0 = 0$. The initialization bound for Δ_0 follows from the centered Gaussian-product concentration in Lemma C.2. □

Lemma G.5 (Quadratic KL-adaptive bootstrap; one-hidden-layer quadratic variant)**.** *Assume $\sigma(z) = z^2$. Define*

$$M_t := 1 + \max_{\alpha \le m} \left(|U^{(m)}_{t,\alpha}| + |V^{(m)}_{t,\alpha}| + |U^{\mathrm{ref}}_{t,\alpha}| + |V^{\mathrm{ref}}_{t,\alpha}| \right),$$

$$\varepsilon_t := \max_{\alpha \le m} \left(|U^{(m)}_{t,\alpha} - U^{\mathrm{ref}}_{t,\alpha}| + |V^{(m)}_{t,\alpha} - V^{\mathrm{ref}}_{t,\alpha}| \right),$$

and define the empirical-support output error

$$\Delta_t := \max_{i \in [K]} \max_{1 \le j \le n_i} \left| g^{(m)}_t(\xi_{i,j}) - \bar{g}_t(\xi_{i,j}) \right|.$$

Let

$$L := 1 + \log \frac{2mN(T+1)}{\delta}.$$

There is an event of probability at least $1 - \delta$, obtained by union bounding the reference-polynomial maximum and empirical-fluctuation bounds over all $0 \le t \le T$ and all empirical inputs $\{\xi_{i,j}\}$, on which

$$M_0 \le C_0 L, \qquad \varepsilon_0 = 0, \qquad \Delta_0 \le C_0 m^{-1/2} L^3,$$

and, for every $0 \le t < T$,

$$M_{t+1} \le C_t M_t^5,$$

$$\varepsilon_{t+1} \le C_t \left[M_t \varepsilon_t + M_t^2 (\Delta_t + \Delta_t^2) \right],$$

$$\Delta_t \le C_t m^{-1/2} L^{3 \cdot 2^t} + C_t M_t^2 \varepsilon_t.$$

Proof. This follows the quadratic bootstrap of Lemma D.2, with the scalar residual coefficient replaced by the adaptive coefficient $A^{(m)}_t$.

For $\sigma(z) = z^2$, the finite width recursion is

$$V^{(m)}_{t+1,\alpha} = V^{(m)}_{t,\alpha} - A^{(m)}_t \left(U^{(m)}_{t,\alpha} \right)^2, \quad U^{(m)}_{t+1,\alpha} = U^{(m)}_{t,\alpha} - 2A^{(m)}_t V^{(m)}_{t,\alpha} U^{(m)}_{t,\alpha}.$$

The reference recursion uses $A^{(\infty)}_t$. Since $A^{(\infty)}_t$ is deterministic, the reference particles remain i.i.d. bounded-degree Gaussian polynomials of the initialization, exactly as in Lemma D.1. Hence the same reference maximum and empirical fluctuation bounds apply.

The deterministic growth estimate is unchanged: $M_{t+1} \le C_t M_t^5$. Subtracting the finite and reference recursions gives

$$\varepsilon_{t+1} \le C_t \left(M_t \varepsilon_t + M_t^2 \alpha_t \right).$$

By Lemma G.3, there is $\alpha_t \le C_{\lambda,K,B_T}(\Delta_t + \Delta_t^2)$. Therefore

$$\varepsilon_{t+1} \le C_t \left[M_t \varepsilon_t + M_t^2 (\Delta_t + \Delta_t^2) \right].$$

By the output decomposition used in Lemma D.2,

$$\Delta_t \le \max_{i,j} \left| \frac{1}{m} \sum_{\alpha=1}^{m} \left(Y_{t,\alpha}^{\mathrm{ref}}(\xi_{i,j}) - \mathbb{E} Y_{t,\alpha}^{\mathrm{ref}}(\xi_{i,j}) \right) \right| + C_t M_t^2 \varepsilon_t,$$

where $Y_{t,\alpha}^{\mathrm{ref}}(\xi) = V_{t,\alpha}^{\mathrm{ref}} \xi^2 \left(U_{t,\alpha}^{\mathrm{ref}} \right)^2$. Hypercontractivity for bounded-degree Gaussian polynomials gives

$$\max_{i,j} \left| \frac{1}{m} \sum_{\alpha=1}^{m} \left(Y_{t,\alpha}^{\mathrm{ref}}(\xi_{i,j}) - \mathbb{E} Y_{t,\alpha}^{\mathrm{ref}}(\xi_{i,j}) \right) \right| \le C_t m^{-1/2} L^{3 \cdot 2^t}.$$

Thus

$$\Delta_t \le C_t m^{-1/2} L^{3 \cdot 2^t} + C_t M_t^2 \varepsilon_t.$$

At initialization, $\varepsilon_0 = 0$, Gaussian maxima give $M_0 \le C_0 L$, and the degree-three Gaussian-polynomial concentration gives $\Delta_0 \le C_0 m^{-1/2} L^3$. □

Theorem G.6 (Finite-time μP stability of KL-regularized domain weights for two-layer dynamics)**.** *Fix an integer $T \ge 0$, $\lambda > 0$, and $\delta \in (0,1)$. Assume the setting above and assume $B_T < \infty$. Then there exist constants $C_T, a_T > 0$, such that, with probability at least $1 - \delta$,*

$$\max_{0 \le t \le T} \max_{i \in [K]} \max_{1 \le j \le n_i} \left| g_t^{(m)}(\xi_{i,j}) - \bar{g}_t(\xi_{i,j}) \right| \le \varepsilon_{m,T}(\delta),$$

where $\varepsilon_{m,T}(\delta) := C_T m^{-1/2} \left(1 + \log \frac{2mN(T+1)}{\delta} \right)^{a_T}$. Consequently,

$$\max_{0 \le t \le T} \max_{i \in [K]} \left| \pi_{i,t}^{(m)} - \pi_{i,t}^{(\infty)} \right| \le \lambda \left(2 B_T \varepsilon_{m,T}(\delta) + \varepsilon_{m,T}(\delta)^2 \right).$$

In particular, if $B_T = O(1)$, $\lambda = O(1)$, and T is fixed, then $\max_{0 \le t \le T} \max_{i \in [K]} \left| \pi_{i,t}^{(m)} - \pi_{i,t}^{(\infty)} \right| = \widetilde{O}(m^{-1/2})$.

Proof. The KL identity in Lemma G.1 shows that the KL-reweighted update is gradient descent on the log-sum-exp objective $\mathcal{R}_m(\theta) = \frac{1}{\lambda} \log \left(\frac{1}{K} \sum_{i=1}^{K} e^{\lambda \ell_i^{(m)}(\theta)} \right)$, with the μP coordinate convention stated above. The point is not to appeal to a separate black-box output-concentration theorem for this log-sum-exp trajectory. Instead, we prove output concentration directly by the same finite-time induction used in Theorems C.7 and D.3.

First consider the linear case $\sigma(z) = z$, write $L := 1 + \log \frac{2mN(T+1)}{\delta}$, by Lemma G.4, $\varepsilon_0 = 0$, $\Delta_0 \le C_0 m^{-1/2} L$, and, for $t < T$,

$$\varepsilon_{t+1} \le C_t \left[(1 + \Delta_t + \Delta_t^2) \varepsilon_t + (\Delta_t + \Delta_t^2) L^{1/2} \right], \quad \Delta_t \le C_t m^{-1/2} L + C_t L^{1/2} \varepsilon_t + \varepsilon_t^2.$$

Define exponents recursively by $\gamma_0^{\mathrm{lin}} := 0$, $\beta_0^{\mathrm{lin}} := 1$,

$$\gamma_{t+1}^{\mathrm{lin}} := 1 + \max \left\{ \gamma_t^{\mathrm{lin}}, \beta_t^{\mathrm{lin}} + \gamma_t^{\mathrm{lin}}, 2\beta_t^{\mathrm{lin}} + \gamma_t^{\mathrm{lin}}, \beta_t^{\mathrm{lin}} + \frac{1}{2}, 2\beta_t^{\mathrm{lin}} + \frac{1}{2} \right\},$$

$$\beta_{t+1}^{\mathrm{lin}} := 1 + \max \left\{ 1, \gamma_{t+1}^{\mathrm{lin}} + \frac{1}{2}, 2\gamma_{t+1}^{\mathrm{lin}} \right\}.$$

We claim that, for $0 \le t \le T$,

$$\varepsilon_t \le C_t m^{-1/2} L^{\gamma_t^{\mathrm{lin}}}, \qquad \Delta_t \le C_t m^{-1/2} L^{\beta_t^{\mathrm{lin}}}.$$

The base case follows from $\varepsilon_0 = 0$ and $\Delta_0 \le C_0 m^{-1/2} L$. Assume the bounds hold at time t. Substituting them into the ε_{t+1} recursion and using $m^{-1} \le m^{-1/2}$, we obtain

$$\varepsilon_{t+1} \le C_{t+1} m^{-1/2} \left(L^{\gamma_t^{\mathrm{lin}}} + L^{\beta_t^{\mathrm{lin}} + \gamma_t^{\mathrm{lin}}} + L^{2\beta_t^{\mathrm{lin}} + \gamma_t^{\mathrm{lin}}} + L^{\beta_t^{\mathrm{lin}} + 1/2} + L^{2\beta_t^{\mathrm{lin}} + 1/2} \right).$$

The definition of $\gamma_{t+1}^{\mathrm{lin}}$ gives

$$\varepsilon_{t+1} \le C_{t+1} m^{-1/2} L^{\gamma_{t+1}^{\mathrm{lin}}}.$$

Then the output bootstrap gives

$$\Delta_{t+1} \le C_{t+1} m^{-1/2} L + C_{t+1} L^{1/2} \varepsilon_{t+1} + \varepsilon_{t+1}^2.$$

Using the bound for ε_{t+1}, and again using $m^{-1} \le m^{-1/2}$, yields

$$\Delta_{t+1} \le C_{t+1} m^{-1/2} \left(L + L^{\gamma_{t+1}^{\mathrm{lin}} + 1/2} + L^{2\gamma_{t+1}^{\mathrm{lin}}} \right).$$

The definition of $\beta_{t+1}^{\mathrm{lin}}$ gives

$$\Delta_{t+1} \le C_{t+1} m^{-1/2} L^{\beta_{t+1}^{\mathrm{lin}}}.$$

This closes the linear induction. Hence $\max_{0 \le t \le T} \Delta_t \le C_T m^{-1/2} L^{\beta_T^{\mathrm{lin}}}$.

Now consider the quadratic case $\sigma(z) = z^2$. By Lemma G.5, $M_0 \le C_0 L$, $\varepsilon_0 = 0$, $\Delta_0 \le C_0 m^{-1/2} L^3$, and

$$M_{t+1} \le C_t M_t^5, \quad \varepsilon_{t+1} \le C_t \left[M_t \varepsilon_t + M_t^2 (\Delta_t + \Delta_t^2) \right], \quad \Delta_t \le C_t m^{-1/2} L^{3 \cdot 2^t} + C_t M_t^2 \varepsilon_t.$$

Define exponents recursively by

$$a_0^{\mathrm{quad}} := 1, \quad \gamma_0^{\mathrm{quad}} := 0, \quad \beta_0^{\mathrm{quad}} := 3, \quad a_{t+1}^{\mathrm{quad}} := 1 + 5a_t^{\mathrm{quad}},$$

$$\gamma_{t+1}^{\mathrm{quad}} := 1 + \max \left\{ a_t^{\mathrm{quad}} + \gamma_t^{\mathrm{quad}}, 2a_t^{\mathrm{quad}} + \beta_t^{\mathrm{quad}}, 2a_t^{\mathrm{quad}} + 2\beta_t^{\mathrm{quad}} \right\},$$

$$\beta_{t+1}^{\mathrm{quad}} := 1 + \max \left\{ 3 \cdot 2^{t+1}, 2a_{t+1}^{\mathrm{quad}} + \gamma_{t+1}^{\mathrm{quad}} \right\}.$$

We claim that, for $0 \le t \le T$,

$$M_t \le C_t L^{a_t^{\mathrm{quad}}}, \quad \varepsilon_t \le C_t m^{-1/2} L^{\gamma_t^{\mathrm{quad}}}, \quad \Delta_t \le C_t m^{-1/2} L^{\beta_t^{\mathrm{quad}}}.$$

The base case follows from the bounds on M_0, ε_0, and Δ_0. Assume the three bounds hold at time t. Since $M_{t+1} \le C_t M_t^5$, we have

$$M_{t+1} \le C_{t+1} L^{5a_t^{\mathrm{quad}}} \le C_{t+1} L^{a_{t+1}^{\mathrm{quad}}}.$$

Next, there is

$$\varepsilon_{t+1} \le C_t \left[M_t \varepsilon_t + M_t^2 \Delta_t + M_t^2 \Delta_t^2 \right].$$

Substituting the induction hypotheses and using $m^{-1} \le m^{-1/2}$, we get

$$\varepsilon_{t+1} \le C_{t+1} m^{-1/2} \left(L^{a_t^{\mathrm{quad}} + \gamma_t^{\mathrm{quad}}} + L^{2a_t^{\mathrm{quad}} + \beta_t^{\mathrm{quad}}} + L^{2a_t^{\mathrm{quad}} + 2\beta_t^{\mathrm{quad}}} \right).$$

The definition of $\gamma_{t+1}^{\mathrm{quad}}$ gives

$$\varepsilon_{t+1} \le C_{t+1} m^{-1/2} L^{\gamma_{t+1}^{\mathrm{quad}}}.$$

Finally, there is

$$\Delta_{t+1} \le C_{t+1} m^{-1/2} L^{3 \cdot 2^{t+1}} + C_{t+1} M_{t+1}^2 \varepsilon_{t+1}.$$

Using the already proved bounds for M_{t+1} and ε_{t+1}, we obtain

$$\Delta_{t+1} \le C_{t+1} m^{-1/2} \left(L^{3 \cdot 2^{t+1}} + L^{2a_{t+1}^{\mathrm{quad}} + \gamma_{t+1}^{\mathrm{quad}}} \right).$$

The definition of $\beta_{t+1}^{\mathrm{quad}}$ gives

$$\Delta_{t+1} \le C_{t+1} m^{-1/2} L^{\beta_{t+1}^{\mathrm{quad}}}.$$

This closes the quadratic induction. Hence

$$\max_{0\le t\le T} \Delta_t \le C_T m^{-1/2} L^{\beta_T^{\mathrm{quad}}}.$$

Combining the linear and quadratic cases, define $a_T := \max\left\{\beta_T^{\mathrm{lin}}, \beta_T^{\mathrm{quad}}\right\}$. Then, for either activation, $\max_{0\le t\le T} \Delta_t \le C_T m^{-1/2} L^{a_T}$. Equivalently,

$$\max_{0\le t\le T} \max_{i\in[K]} \max_{1\le j\le n_i} \left| g_t^{(m)}(\xi_{i,j}) - \bar{g}_t(\xi_{i,j}) \right| \le C_T m^{-1/2} \left(1 + \log \frac{2mN(T+1)}{\delta}\right)^{a_T}.$$

It remains to convert output concentration into domain-weight concentration. By Lemma G.3, we have $P_t \le \lambda(2B_T\Delta_t + \Delta_t^2)$. Taking the maximum over $0 \le t \le T$, and using $\max_{0\le t\le T} \Delta_t \le \varepsilon_{m,T}(\delta)$, gives

$$\max_{0\le t\le T} \|\pi_t^{(m)} - \pi_t^{(\infty)}\|_\infty \le \lambda\left(2B_T\varepsilon_{m,T}(\delta) + \varepsilon_{m,T}(\delta)^2\right).$$

Since

$$\|\pi_t^{(m)} - \pi_t^{(\infty)}\|_\infty = \max_{i\in[K]} |\pi_{i,t}^{(m)} - \pi_{i,t}^{(\infty)}|,$$

we obtain

$$\max_{0\le t\le T} \max_{i\in[K]} |\pi_{i,t}^{(m)} - \pi_{i,t}^{(\infty)}| \le \lambda\left(2B_T\varepsilon_{m,T}(\delta) + \varepsilon_{m,T}(\delta)^2\right).$$

If $B_T = O(1)$, $\lambda = O(1)$, and T is fixed, then $\varepsilon_{m,T}(\delta) = \widetilde{O}(m^{-1/2})$, and hence

$$\lambda\left(2B_T\varepsilon_{m,T}(\delta) + \varepsilon_{m,T}(\delta)^2\right) = \widetilde{O}(m^{-1/2}).$$

This proves the theorem. □

H Proofs of the Horizon-Dependence Examples

H.1 Worst-Case Exponential Dependence under Square Loss

Proof of Theorem 4.3. Let $Z_t^{(m)} := \begin{pmatrix} q_t^{(m)} \\ s_t^{(m)} \end{pmatrix}$. For $\beta = 2\sqrt{3}$, the reference trajectory satisfies

$$\bar{Z}_0 = (0,1), \qquad \bar{Z}_t = (2\sqrt{3}, 4), \quad t \ge 1.$$

At initialization,

$$\sqrt{m}\begin{pmatrix} q_0^{(m)} \\ s_0^{(m)} - 1 \end{pmatrix} \Longrightarrow N(0, I_2),$$

because

$$\mathrm{Var}(V_0U_0) = 1, \qquad \mathrm{Var}\left[\frac{1}{2}(V_0^2 + U_0^2) - 1\right] = 1,$$

and the covariance is zero.

The Jacobian of (3) at $(0,1)$ is

$$J_0 = \begin{pmatrix} 3 & 2\sqrt{3} \\ \sqrt{3} & 4 \end{pmatrix},$$

while at the stationary point $(2\sqrt{3}, 4)$,

$$J_\star = \begin{pmatrix} -3 & 0 \\ -2\sqrt{3} & 1 \end{pmatrix}.$$

Hence

$$Dq_T(\bar{Z}_0) = (1,0)J_\star^{T-1}J_0 = (-3)^{T-1}\begin{pmatrix} 3 & 2\sqrt{3} \end{pmatrix}.$$

The delta method therefore gives

$$\sqrt{m}\left(q_T^{(m)} - 2\sqrt{3}\right) \Longrightarrow N\left(0, 21\,9^{T-1}\right).$$

Since the state recursion depends only on $\xi_t^2 = 1/2$, $\xi_T \perp q_T^{(m)}$, $\xi_T^2 = \frac{1}{2}$. Therefore

$$\sqrt{m}\left(g_T^{(m)} - \bar{g}_T\right) \Longrightarrow N\left(0, \frac{21}{2}9^{T-1}\right).$$

Let $\sigma_T = \sqrt{\frac{21}{2}}\,3^{T-1}$. If

$$\liminf_{m\to\infty} \mathbb{P}\left(\sqrt{m}|g_T^{(m)} - \bar{g}_T| \le B_T(\delta)\right) \ge 1 - \delta,$$

then convergence in distribution implies

$$\mathbb{P}(|N(0, \sigma_T^2)| \le B_T(\delta)) \ge 1 - \delta,$$

and hence

$$B_T(\delta) \ge \sigma_T z_{1-\delta/2} = \sqrt{\frac{21}{2}}\, z_{1-\delta/2}\, 3^{T-1}.$$

Thus, in this square-loss construction, the fluctuation prefactor is at least $\Omega(3^T)$ at any fixed confidence. □

Define

$$s_t^{(m)} := \frac{1}{2m}\sum_{\alpha=1}^{m}\left(V_{t,\alpha}^2 + U_{t,\alpha}^2\right).$$

Since

$$\xi_t^2 = \frac{1}{2}, \qquad \xi_t\chi_t = \frac{1}{2}(q_t^{(m)} - \beta),$$

set

$$a_t := \frac{1}{2}(q_t^{(m)} - \beta).$$

Direct expansion gives

$$\begin{aligned} q_{t+1}^{(m)} &= (1 + a_t^2)q_t^{(m)} - 2a_t s_t^{(m)}, \\ s_{t+1}^{(m)} &= (1 + a_t^2)s_t^{(m)} - 2a_t q_t^{(m)}. \end{aligned} \tag{3}$$

At infinite width, $(\bar{q}_0, \bar{s}_0) = (0, 1)$, and therefore

$$(\bar{q}_1, \bar{s}_1) = \left(\beta, 1 + \frac{\beta^2}{4}\right),$$

which is stationary for every $t \ge 1$.

For $t \ge 1$, define

$$e_t := q_t^{(m)} - \beta, \qquad d_t := s_t^{(m)} - \left(1 + \frac{\beta^2}{4}\right).$$

Then (3) gives

$$\begin{aligned} e_{t+1} &= e_t\left(-\frac{\beta^2}{4} - d_t + \frac{\beta}{4}e_t + \frac{1}{4}e_t^2\right), \\ d_{t+1} &= d_t - \beta e_t + \frac{\beta^2 - 12}{16}e_t^2 + \frac{1}{4}e_t^2 d_t. \end{aligned} \tag{4}$$

H.2 Time-Uniform Bound under Label-Smoothed Cross-Entropy

Proof of Theorem 4.4. Let $u_{t,\alpha} := U_{t,\alpha}^\top v$, and define $q_t^{(m)} := \frac{1}{m}\sum_{\alpha=1}^m V_{t,\alpha}u_{t,\alpha}$, $s_t^{(m)} := \frac{1}{2m}\sum_{\alpha=1}^m \left(V_{t,\alpha}^2 + u_{t,\alpha}^2\right)$. Since $\xi_i = y_i v$, $y_i g_t^{(m)}(\xi_i) = q_t^{(m)}$. Thus the empirical cross-entropy depends only on $q_t^{(m)}$, and $\ell'(q) = \sigma(q) - p$.

Set $a_t := \eta_\star\big(\sigma(q_t^{(m)}) - p\big)$. The μP updates of both layers reduce exactly to

$$V_{t+1,\alpha} = V_{t,\alpha} - a_t u_{t,\alpha}, \quad u_{t+1,\alpha} = u_{t,\alpha} - a_t V_{t,\alpha}.$$

Hence

$$q_{t+1}^{(m)} = (1+a_t^2)q_t^{(m)} - 2a_t s_t^{(m)}, \quad s_{t+1}^{(m)} = (1+a_t^2)s_t^{(m)} - 2a_t q_t^{(m)}. \tag{5}$$

At infinite width, $\bar{q}_0 = 0$, $\bar{s}_0 = 1$. Since $\eta_\star\left(\frac{1}{2} - p\right) = -\frac{q_\star}{2}$, the first update gives

$$\bar{q}_1 = q_\star, \qquad \bar{s}_1 = 1 + \frac{q_\star^2}{4} =: s_\star.$$

Because $\sigma(q_\star) = p$, the reference update coefficient is zero thereafter. Hence

$$\bar{q}_t = q_\star, \quad \bar{s}_t = s_\star, \quad t \geq 1. \tag{6}$$

Now define

$$e_t := q_t^{(m)} - q_\star, \qquad d_t := s_t^{(m)} - s_\star.$$

Linearizing (5) at $(q_\star, s_\star)$ gives

$$\begin{pmatrix} e_{t+1} \\ d_{t+1} \end{pmatrix} = \begin{pmatrix} \lambda_\star & 0 \\ -2c_\star q_\star & 1 \end{pmatrix} \begin{pmatrix} e_t \\ d_t \end{pmatrix} + O(e_t^2 + |e_t d_t|),$$

where $c_\star = \eta_\star p(1-p)$, and $\lambda_\star = 1 - 2s_\star c_\star$. For $\varepsilon = 0.1$, $p = 0.95$, $q_\star = \log 19$, and $\lambda_\star \approx 0.0156$. In particular, $|\lambda_\star| < 1$.

The output-error update vanishes at $e_t = 0$ for every d_t. Thus, by continuity, there exist $r > 0$, $C_1 < \infty$, and $0 < \rho < 1$ such that whenever $|e_t| \leq r$, $|d_t| \leq r$, one has

$$|e_{t+1}| \leq \rho|e_t|, \quad |d_{t+1} - d_t| \leq C_1|e_t|. \tag{7}$$

Choose the local initialization neighborhood sufficiently small so that $|e_1| \leq r, \quad |d_1| + \frac{C_1}{1-\rho}|e_1| \leq r$. Then, inductively using (7), $|e_t| \leq \rho^{t-1}|e_1|$, and

$$|d_t| \leq |d_1| + C_1 \sum_{j=1}^{t-1} |e_j| \leq |d_1| + \frac{C_1}{1-\rho}|e_1| \leq r.$$

Hence the trajectory remains in the local region for all t, and the bootstrap closes.

It remains to show that the finite-width trajectory enters this stable region. Write $x_m := q_0^{(m)}$, $u_m := s_0^{(m)} - 1$. Set $h_\delta = \log(4/\delta)$ and $S_\pm = m^{-1}\sum_\alpha ((V_{0,\alpha} \pm u_{0,\alpha})/\sqrt{2})^2$. Each $mS_\pm$ has the χ_m^2 distribution, so the chi-square tail bound and a union bound give

$$\max\{|S_+ - 1|, |S_- - 1|\} \leq b_{m,\delta} := 2\sqrt{h_\delta/m} + 2h_\delta/m$$

with probability at least $1 - \delta$. Since $x_m = (S_+ - S_-)/2$ and $u_m = (S_+ + S_-)/2 - 1$, on this event

$$x_m^2 + u_m^2 \leq b_{m,\delta}^2. \tag{8}$$

The first-step map is smooth near $(0,1)$, and maps $(0,1) \mapsto (q_\star, s_\star)$. Thus, for some constant K_0,

$$|e_1| + |d_1| \leq K_0(|x_m| + |u_m|).$$

Combining with (8), $|e_1| + |d_1| \leq C_2 b_{m,\delta}$. Choose $0 < r_0 \leq 1$ small enough that $x_m^2 + u_m^2 \leq r_0^2$ implies both first-step conditions above. Such a choice exists by continuity of the first-step map and depends only on the fixed model and learning rate. For

$$m \geq m_0(\delta) := \left\lceil \frac{16 h_\delta}{r_0^2} \right\rceil,$$

we have $b_{m,\delta} \le r_0/2 + r_0^2/8 < r_0$, so the initialization event in (8) implies these conditions. Also $b_{m,\delta} \le (5/2)\sqrt{h_\delta/m}$ for these widths. On this same event, of probability at least $1-\delta$, the induction above controls every $t \ge 1$ with $|e_t| \le \rho^{t-1}|e_1|$.

Finally, since $g_t^{(m)}(\xi_i) - \bar{g}_t(\xi_i) = y_i e_t$, we have

$$\sup_{t\ge 1} \max_i |g_t^{(m)}(\xi_i) - \bar{g}_t(\xi_i)| \le |e_1| \le C\sqrt{\frac{h_\delta}{m}}.$$

The initialization error satisfies the same order by (8). Therefore

$$\sup_{t\ge 0} \max_i |g_t^{(m)}(\xi_i) - \bar{g}_t(\xi_i)| \le C\sqrt{\frac{h_\delta}{m}}.$$

The constant C is independent of m, δ, T, and the width threshold $m_0(\delta) = \lceil C_0 \log(4/\delta) \rceil$, with $C_0 = 16/r_0^2$, is independent of T. □

I A Sharper Bound for Two-Layer Linear Networks

For two-layer linear networks, the empirical covariance of the neuron parameters closes under the exact discrete updates. This gives a vector-input refinement of Theorem 4.1 without horizon-dependent logarithmic powers. The time and learning-rate dependence is retained explicitly in the sensitivity factors below.

Proposition I.1 (Vector-input covariance bound)**.** *Consider $g_t^{(m)}(x) = m^{-1}\sum_{\alpha=1}^m V_{t,\alpha} U_{t,\alpha}^\top x$, with independent $V_{0,\alpha} \sim N(0,1)$ and $U_{0,\alpha} \sim N(0, I_d)$. Let $\|x_t\|_2 \le X$, $|y_t| \le Y$, and use the square-loss updates*

$$V_{t+1,\alpha} = V_{t,\alpha} - \eta r_t U_{t,\alpha}^\top x_t, \qquad U_{t+1,\alpha} = U_{t,\alpha} - \eta r_t V_{t,\alpha} x_t, \qquad r_t = g_t^{(m)}(x_t) - y_t,$$

where $\eta \ge 0$. The training sequence is deterministic, or independent of the initialization and conditioned upon. Set $R_0 = 2$, $D_0 = 1$, and

$$a_t = 1 + \eta X(XR_t + Y), \qquad R_{t+1} = a_t^2 R_t, \qquad L_t = a_t^2 + 2\eta X^2 R_t a_t, \qquad D_{t+1} = L_t D_t.$$

For $\delta \in (0,1)$, $h = (d+1)\log 9 + \log(2/\delta)$ and $m \ge 64h$, with probability at least $1-\delta$,

$$\max_{0\le t\le T} \sup_{\|x\|_2 \le X} |g_t^{(m)}(x) - \bar{g}_t(x)| \le 4XD_T\left(\sqrt{\frac{h}{m}} + \frac{h}{m}\right). \tag{9}$$

Here $\bar{g}_t$ is the infinite-width limit defined by the covariance recursion below, starting from I_{d+1}. In particular, the logarithmic powers do not depend on T.

Proof. **Exact covariance recursion.** Write $w_{t,\alpha} = (V_{t,\alpha}, U_{t,\alpha}^\top)^\top$ and $\Sigma_t^{(m)} = m^{-1}\sum_\alpha w_{t,\alpha} w_{t,\alpha}^\top$. For a symmetric matrix Σ, let $q(\Sigma) \in \mathbb{R}^d$ denote its lower-left block, and define

$$B(x) = \begin{pmatrix} 0 & x^\top \\ x & 0_{d\times d} \end{pmatrix}, \quad A_t(\Sigma) = I_{d+1} - \eta\big(x_t^\top q(\Sigma) - y_t\big)B(x_t), \quad F_t(\Sigma) = A_t(\Sigma)\Sigma A_t(\Sigma)^\top.$$

The parameter updates give $\Sigma_{t+1}^{(m)} = F_t(\Sigma_t^{(m)})$ exactly, and $g_t^{(m)}(x) = x^\top q(\Sigma_t^{(m)})$. Define $\bar{\Sigma}_0 = I_{d+1}$, $\bar{\Sigma}_{t+1} = F_t(\bar{\Sigma}_t)$, and $\bar{g}_t(x) = x^\top q(\bar{\Sigma}_t)$. The strong law for $\Sigma_0^{(m)}$ and continuity of each F_t identify this as the infinite-width limit at every fixed time.

Initialization concentration. For a unit vector $v \in \mathbb{R}^{d+1}$, $mv^\top \Sigma_0^{(m)} v$ has the χ^2_m distribution. Its moment-generating function $(1-2s)^{-m/2}$ gives the standard bound

$$\mathbb{P}\left(|v^\top(\Sigma_0^{(m)} - I_{d+1})v| > 2\sqrt{u/m} + 2u/m\right) \le 2e^{-u}, \qquad u > 0.$$

Take a $1/4$-net of the unit sphere with at most 9^{d+1} elements. For any symmetric H, approximation by this net gives $\|H\|_{\mathrm{op}} \le 2\max_v |v^\top H v|$, with the maximum over the net. A union bound with $u = h$ therefore yields the event

$$\|\Sigma_0^{(m)} - I_{d+1}\|_{\mathrm{op}} \le \epsilon_m := 4\left(\sqrt{h/m} + h/m\right) \tag{10}$$

with probability at least $1-\delta$. Since $m \geq 64h$, $\epsilon_m \leq 9/16 < 1$, so both initial covariances have operator norm at most $R_0 = 2$.

Propagation of the covariance error. For $\|\Sigma\|_{\mathrm{op}} \leq R_t$, the inequalities $\|B(x_t)\|_{\mathrm{op}} \leq X$ and $|x_t^\top q(\Sigma) - y_t| \leq XR_t + Y$ imply $\|A_t(\Sigma)\|_{\mathrm{op}} \leq a_t$ and $\|F_t(\Sigma)\|_{\mathrm{op}} \leq a_t^2 R_t = R_{t+1}$. Thus both covariance trajectories stay within these bounds. For two covariances $\Sigma, \widetilde{\Sigma}$ in the same ball,

$$\|A_t(\Sigma) - A_t(\widetilde{\Sigma})\|_{\mathrm{op}} \leq \eta X^2 \|\Sigma - \widetilde{\Sigma}\|_{\mathrm{op}}.$$

Expanding the difference of the two products defining F_t then gives

$$\|F_t(\Sigma) - F_t(\widetilde{\Sigma})\|_{\mathrm{op}} \leq \big(a_t^2 + 2\eta X^2 R_t a_t\big)\|\Sigma - \widetilde{\Sigma}\|_{\mathrm{op}}.$$

Consequently, $\|\Sigma_t^{(m)} - \bar{\Sigma}_t\|_{\mathrm{op}} \leq D_t \epsilon_m$. Since $L_t \geq 1$, $D_t \leq D_T$ for $t \leq T$. Finally, $|x^\top q(\Sigma_t^{(m)} - \bar{\Sigma}_t)| \leq X\|\Sigma_t^{(m)} - \bar{\Sigma}_t\|_{\mathrm{op}}$, which proves (9). □

Remark. The explicit D_T above is a worst-case sensitivity bound, not a claim of time-uniform control. A bounded reference trajectory gives a more informative version. If $\max_{t \leq T} \|\bar{\Sigma}_t\|_{\mathrm{op}} \leq R - 1$, with $R \geq 2$, set $a = 1 + \eta X(XR + Y)$ and $L = a^2 + 2\eta X^2 R a$. Whenever $m \geq 64hL^{2T}$, the same initialization event gives $\epsilon_m L^T < 1$. Induction in the radius-one neighborhood of the reference trajectory then gives (9) with $D_T = L^T$. This separates initialization concentration from dynamical amplification and exhibits the learning-rate dependence. These refinements concern the linear model; the quadratic and three-layer bounds in Theorem 4.1 retain their stated scope, as does the adaptive domain-weight theorem.

J Experimental Details

J.1 Synthetic Experiments

Architecture and training. The two-layer student and teacher networks have weight matrices $W_1 \in \mathbb{R}^{m \times d}$ and $W_2 \in \mathbb{R}^{1 \times m}$, with $d = 500$. We use widths

$$m \in \{100, 311, 522, 733, 944, 1155, 1366, 1577, 1788, 2000\},$$

with the width-2000 model serving as the teacher. The ReLU models add an activation after the first linear layer and otherwise use the same architecture and width sweep. The regularization parameter is $\lambda = 1.0$, and each model is trained for $T = 500$ steps. Results are averaged over seeds 0, 1, and 42; shaded bands in Figures 1 and 2 indicate standard deviation.

Learning-rate selection under μP. For linear networks, we select the learning rate by grid search over $[1.0 \times 10^{-4}, 5.0 \times 10^{-4}]$ and use $\eta = 2.0 \times 10^{-4}$ at every width. ReLU networks use the same search range, yielding $\eta = 3.0 \times 10^{-4}$ at every width. Both settings keep $\lambda = 1.0$.

Standard-parameterization ablation. The ReLU ablation uses Kaiming initialization (He et al., 2015) and an unscaled linear output. Data generation, architectures, widths, regularization, training budget, and seeds match the μP experiment. We tune the learning rate separately for each width by grid search over $[5 \times 10^{-5}, 1 \times 10^{-2}]$. The selected rates are listed below, grouping widths that share the same rate.

Widths m	Selected learning rate η
$100, 311, 522$	5×10^{-4}
$733, 944, 1366, 1577, 1788, 2000$	1×10^{-4}
1155	5×10^{-5}

The additional quadratic-activation comparison appears in Figure 5.

Training-horizon experiment. For Figure 6, we use two-layer ReLU networks under μP with student width 100 and teacher width 2000. We vary T over $[100, 5000]$, tune the student learning rate for each training budget, and average the student–teacher KL divergence over three random seeds. This comparison measures cross-width stability across training budgets with a learning rate selected for each budget.

J.2 GPT-STYLE EXPERIMENTS

Dataset and architecture. We use SlimPajama-6B (Soboleva et al., 2023), sourced from `https://huggingface.co/datasets/cerebras/SlimPajama-627B` under the Apache-2.0 license. Its seven domains are ArXiv, Books, CommonCrawl, Github, StackExchange, C4, and Wikipedia. Student and teacher models have 4 layers and 8 attention heads and are initialized under μP. Student hidden widths are $\{64, 128, 256, 512, 1024\}$, and the teacher width is 1024.

Optimization and computational resources. We use MuAdamW (Yang et al., 2021) and the exponentiated-gradient domain-weight update on the KL-regularized weighted cross-entropy objective described in Section 7. The model learning rate η and domain-weight learning rate η_π are tuned over $[1.0 \times 10^{-4}, 1.0 \times 10^{-2}]$; both are set to 1.0×10^{-3}. The KL regularization parameter is $\lambda = 10.0$. The batch size is 4 per domain and 28 in total, with maximum sequence length 1024. Each model is trained for $T = 5000$ steps on 8 RTX A6000 GPUs using distributed data parallel training.

K ADDITIONAL SYNTHETIC EXPERIMENTS

Figures 5 and 6 report the quadratic-activation ablation and the training-budget comparison described in Section 7. Experimental settings appear in Appendix J.1.

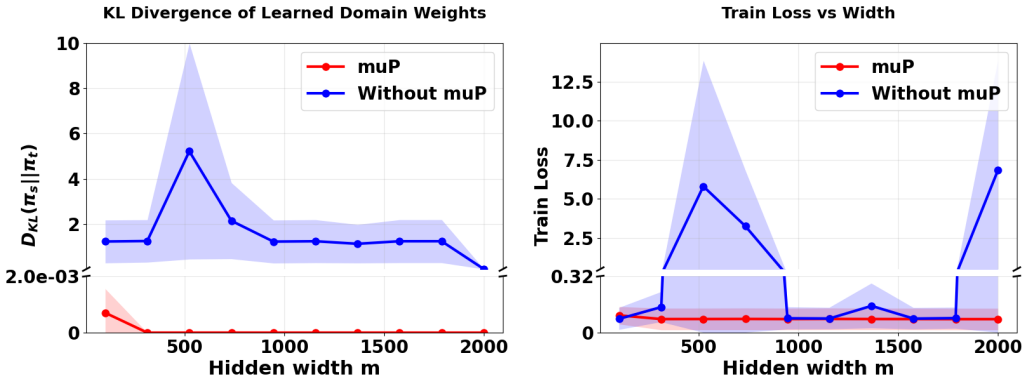


Figure 5: 2-layer network with quadratic activation function (with a two-scale y-axis, where the lower section uses a smaller y-axis scale).

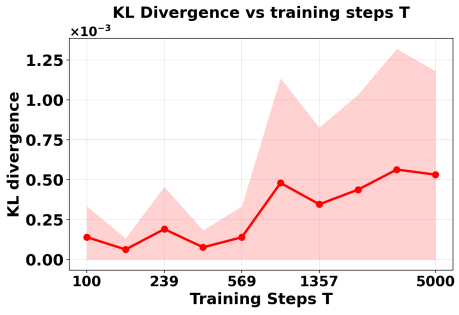


Figure 6: KL divergence of learned domain weights vs training horizon T. The student learning rate is tuned separately for each training budget.