A QUALITATIVE TEST-RISK MECHANISM FOR SCALING BEHAVIOR IN NORMALIZED RESIDUAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

The scaling behavior, in which test performance often improves as model size and data increase, is a central empirical phenomenon in modern deep learning, yet its theoretical basis remains incomplete. In this paper, we study depth expansion in normalized residual networks: starting from a trained model in an old hypothesis class, we insert a new residual block at an intermediate layer and ask when such an expansion can yield a provable improvement in test risk. We develop a unified framework that decomposes this question into representational gain, optimization gain, and generalization transfer. First, under a first-order descent condition near zero initialization, we prove that the expanded hypothesis class contains an auxiliary jumpboard model with strictly smaller test risk than the original model with high probability. Second, under norm control tailored to post-normalized residual architectures, we establish a norm-based Rademacher complexity bound for the expanded model class. These ingredients lead to a test-risk guarantee, which works directly at the train/test level. Together, these results provide a theorem-driven mechanism under which residual depth expansion can improve test performance in normalized residual networks. More broadly, they suggest that scaling is inherently joint: depth creates new improving directions, width enhances the finite-sample observability of weak signals, and data determines whether the statistical cost of expansion can be controlled.

1 INTRODUCTION

Large-scale deep learning systems exhibit a striking empirical regularity: as model size and training data increase, test performance often improves predictably. This regularity, often summarized by empirical scaling laws, has become central in modern machine learning, yet its theoretical basis remains incomplete. In particular, explaining when increasing depth improves *test risk* requires more than observing that a deeper model class is larger: one must also show that the enlarged class contains a genuinely better function, that optimization can realize this gain, and that the gain survives finite-sample error.

In this paper, we study depth expansion in normalized residual networks. Starting from a trained reference model f^*_{old}, we insert a new residual block after an intermediate layer and ask when this expansion can be converted into a provable improvement in test risk. Our analysis is built on the decomposition

$$\text{jumpboard model} \longrightarrow \text{optimization gain} \longrightarrow \text{generalization transfer}.$$

At the representation level, we show that if the inserted block creates a nontrivial first-order descent direction near zero initialization, then the expanded hypothesis class contains a *jumpboard model* with strictly smaller test risk than f^*_{old} with high probability. At the optimization level, we ask whether training can realize this potential gain. At the generalization level, we control the statistical cost of the enlarged class and translate the population-level gain into a test-risk guarantee.

Under this framework, we obtain a qualitative scaling-mechanism theorem. The proof route of this theorem works directly at the train/test level, and is therefore more robust in degenerate regimes where the population margin becomes too small to transfer sharply. Together, these results show that when representational gain and optimization gain dominate generalization cost, increasing depth does not worsen test performance; under stronger conditions, it yields strict improvement.

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.

Our analysis also clarifies the distinct roles of width and data. Width improves the finite-sample observability of weak improvement signals, especially near the deepest-model regime, while data controls the decay of the generalization penalty. Thus, the relevant scaling picture is inherently joint: depth creates new improving directions, width stabilizes their observability, and data pays for the statistical cost of model expansion.

Our contributions are threefold. First, we propose a unified analytical framework that decomposes depth scaling into representational gain, optimization gain, and generalization transfer, thereby avoiding the common shortcut of equating a stronger hypothesis class with better test performance. Second, we establish a test-set scaling-mechanism theorem which works directly at the train/test level and remains meaningful in degenerate regimes. Third, we clarify the theoretical roles of width and data in scaling: width improves finite-sample signal observability, while data governs the decay of the generalization penalty and determines whether model expansion remains beneficial. The framework naturally extends to post-normalized residual architectures, including ResNet-like and Transformer-like blocks, provided the corresponding regularity and first-order non-degeneracy conditions hold.

2 Related Work

Empirical studies show power-law gains with increased scale (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022; Rosenfeld et al., 2019; Hernandez et al., 2021). While recent theory explores global scaling trends (Bahri et al., 2024), we focus specifically on architecture-specific test-risk improvements. Prior work shows skip connections stabilize training (He et al., 2016a;b; Veit et al., 2016; Balduzzi et al., 2017) and depth increases expressivity (Telgarsky, 2016; Eldan & Shamir, 2016). Other analyses address trainability via mean-field limits (Yang & Schoenholz, 2017; Lu et al., 2020), optimization effects of overparameterized depth (Arora et al., 2018a), dynamical-systems and stability viewpoints (Haber & Ruthotto, 2018; Chang et al., 2018), or specific initializations (Zhang et al., 2019; De & Smith, 2020; Bachlechner et al., 2021). Conversely, we prove when residual blocks directly improve *test risk*. Classical (Bartlett et al., 2017; Golowich et al., 2018; Neyshabur et al., 2017; Arora et al., 2018b) and ResNet-specific Rademacher bounds (Ma et al., 2020; Kammonen et al., 2020) establish complexity upper bounds. The comparison between these works and our work is shown in Appendix A.

3 Notation and Assumptions

Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$. Let $S_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{M} \sim \mathcal{D}^M$ and $S_{\text{test}} = \{(\tilde{x}_j, \tilde{y}_j)\}_{j=1}^{K} \sim \mathcal{D}^K$ be independent training and test sets. The loss is $\ell : \mathbb{R}^C \times \mathcal{Y} \to [0, B_\ell]$. For any model f, define $\mathcal{L}_{\text{train}}(f) = \frac{1}{M}\sum_{i=1}^{M} \ell(f(x_i), y_i)$, $\mathcal{L}_{\text{test}}(f) = \frac{1}{K}\sum_{j=1}^{K} \ell(f(\tilde{x}_j), \tilde{y}_j)$, $R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)]$.

We decompose the network at layer l^* as $f = f_{\text{top}} \circ f_{\text{bot}}$, where $f_{\text{bot}} : \mathcal{X} \to \mathbb{R}^N$ and $f_{\text{top}} : \mathbb{R}^N \to \mathbb{R}^C$. The old hypothesis class is $\mathcal{H}_{\text{old}} = \{x \mapsto f_{\text{top}}(f_{\text{bot}}(x)) : f_{\text{top}} \in \mathcal{F}_{\text{top}},\ f_{\text{bot}} \in \mathcal{F}_{\text{bot}}\}$. and $f^*_{\text{old}} \in \mathcal{H}_{\text{old}}$ denotes a trained old model. After inserting a residual block h after layer l^*, the new hypothesis class is $\mathcal{H}_{\text{new}} = \{x \mapsto f_{\text{top}}\big(f_{\text{bot}}(x) + h(f_{\text{bot}}(x))\big) : f \in \mathcal{H}_{\text{old}},\ h \in \mathcal{F}_{\text{res}}\}$. We parameterize the inserted block by θ, $h_\theta(\cdot)$, and assume $h_0 = 0$, so zero initialization recovers the old model. The inserted block is the purely additive map $z \mapsto z + h_\theta(z)$ and carries no normalization of its own, so $h_0 = 0$ recovers f^*_{old} exactly under both pre-norm and post-norm placements; the placement affects only the constants of Lemma 1 (Appendix F).

We also define activation-gradient averages at layer l^* by $\bar{\mu} := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\nabla_{x^{(l^*)}} \ell\big(f_{\text{top}}(x^{(l^*)}), y\big)\right]$, $\mu := \frac{1}{M}\sum_{i=1}^{M} \nabla_{x_i^{(l^*)}} \ell\Big(f_{\text{top}}(x_i^{(l^*)}), y_i\Big)$, $g := \frac{1}{K}\sum_{j=1}^{K} \nabla_{\tilde{x}_j^{(l^*)}} \ell\Big(f_{\text{top}}(\tilde{x}_j^{(l^*)}), \tilde{y}_j\Big)$.

A *jumpboard model* is a near-identity model in $\mathcal{H}_{\text{new}}$ that improves over f^*_{old} along a local descent direction at zero initialization. The *training-set jumpboard model*, denoted by $\tilde{f}_S$, is constructed from an empirical first-order descent direction on S_{train}. Its empirical jumpboard margin is denoted by $\Delta_{\text{train},S} := \mathcal{L}_{\text{train}}(f^*_{\text{old}}) - \mathcal{L}_{\text{train}}(\tilde{f}_S) > 0$. The optimization gain relative to the comparison jumpboard model is denoted by $\Delta_{\text{ERM}} \geq 0$.

Let $\epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H}, n, \delta)$ denote the normalized-residual-network generalization bound. For brevity, we write $\epsilon_M = \epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H}_{\text{new}}, M, \delta)$, $\epsilon_K = \epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H}_{\text{new}}, K, \delta)$. In the main theorems, the total generalization cost is $2(\epsilon_M + \epsilon_K)$ for the direct train/test proof route. $\epsilon_{\text{gen}}^{\text{norm}}$ is a shorthand for the explicit norm-based generalization bound in Lemma 1, Corollary 2 in Appendix F, not an additional assumption.

Assumption 1 (pathwise differentiability under expectation). For any admissible parameter direction v and scalar step size η in a neighborhood of 0, write $h_{\eta v}(z) := h_\theta(z)\big|_{\theta=\eta v}$. For every admissible direction v, define $G_v(\eta) := \mathbb{E}_{(x,y)\sim D}\left[\ell\left(f_{\text{top}}\left(z + h_{\eta v}(z)\right), y\right)\right], z = f_{\text{bot}}(x)$. We assume that G_v is differentiable at $\eta = 0$, and that differentiation may be exchanged with expectation: $G_v'(0) = \mathbb{E}_{(x,y)\sim D}\left[\nabla_z \ell(f_{\text{top}}(z), y)^\top \frac{\partial h_\theta(z)}{\partial \theta}\Big|_{\theta=0} v\right]$.

Assumption 2 (Lipschitz boundedness). The loss ℓ is L_ℓ-Lipschitz and bounded in $[0, B_\ell]$. The top map f_{top} is L_{top}-Lipschitz and satisfies $\|f_{\text{top}}(0)\| \le B_0$.

Assumption 3 (gradient-covariance control). Let $\zeta = \nabla_{x^{(l^*)}} \ell$ and $\Sigma = \text{Cov}(\zeta)$. There exist a width-independent constant $C_\Sigma > 0$ and a scalar τ^2 such that $\lambda_{\max}(\Sigma) \le C_\Sigma \tau^2, \tau^2 \ge \max_{1\le i\le N} \text{Var}(\zeta_i)$.

This assumption is intended as a mild structural condition in practical settings. Our experiments suggest that different coordinates of activation gradients are nearly uncorrelated.

Assumption 4 (normalized residual network regularity). Consider an L-layer normalized residual network with normalized function $\mathcal{N}_l$: $T_l(z) = \mathcal{N}_l(z + h_l(z)), l = 0, \ldots, L-1$. Assume: (i) each residual branch $h_l(\cdot; W_l)$ is s_l-Lipschitz in the input; (ii) $\|h_l(z; W_l) - h_l(z; W_l')\| \le L_l^{(p)} \|W_l - W_l'\|_F \|z\|$; (iii) there exists $r_{\min} > 0$ such that $\inf_{x,\theta,l} \|z_l(x) + h_l(z_l(x); \theta)\| \ge r_{\min}$; (iv) $\|W_l\|_\sigma \le s_l$, $\|W_l\|_F \le b_l$, and $\|x\| \le B_x$.

Assumption 4ϵ (engineering-ϵ alternative). Assumption 4(iii) is an idealized non-degeneracy condition used to ensure that the normalization operators are Lipschitz on the reachable set. In practical normalized architectures, possible singularities of normalization operators are often removed by an engineering constant $\epsilon_{\text{eng}} > 0$. We assume that the resulting stabilized normalization operators are uniformly Lipschitz and bounded on the reachable set: for each layer l, there exist finite constants $c_{l,\epsilon} > 0$ and $B_{\text{norm},\epsilon} > 0$ such that, for all reachable normalization inputs a, a', $\|\mathcal{N}_{l,\epsilon}(a) - \mathcal{N}_{l,\epsilon}(a')\| \le c_{l,\epsilon}\|a - a'\|, \|\mathcal{N}_{l,\epsilon}(a)\| \le B_{\text{norm},\epsilon}$. Under this alternative, Assumption 4(iii) can be dropped, with the constants inside $\epsilon_{\text{gen}}^{\text{norm}}$ replaced by ϵ_{eng}-dependent constants. Concrete stabilized normalization forms and their constant-level effects are given in Appendix L.

Condition 1 (first-order non-degeneracy of the inserted residual block). The inserted residual block class $\mathcal{F}_{\text{res}}$ is parameterized by θ, satisfies $h_0 = 0$, and is differentiable at $\theta = 0$. We assume that there exists at least one admissible insertion layer l^*. For the empirical train/test route, we use the training-set analogue. Given $S_{\text{train}} = \{(x_i, y_i)\}_{i=1}^M$, write $x_i^{(l^*)} := f_{\text{bot}}(x_i)$. We say that v_S is an empirical jumpboard direction if $\frac{1}{M}\sum_{i=1}^M \nabla_{x_i^{(l^*)}} \ell\big(f_{\text{top}}(x_i^{(l^*)}), y_i\big)^\top \frac{\partial h_\theta(x_i^{(l^*)})}{\partial \theta}\Big|_{\theta=0} v_S < 0$. The empirical direction v_S is used in the direct train/test route.

This condition is a mild non-degeneracy constraint rather than a strong global assumption. Condition 1 is not intended to hold for every trained model or every possible insertion layer. Rather, it characterizes the improvable regime of depth expansion: the inserted residual block must introduce a first-order direction that is not orthogonal to the relevant activation-gradient signal. If no such direction exists, the present first-order mechanism is exhausted and our theorems make no strict-improvement claim. Moreover, the parameter stationarity of the old trained model only makes the loss gradient orthogonal to the original tangent directions, which neither implies the activation gradient is zero nor orthogonal to the new tangent directions introduced by residual blocks. Further discussion of Condition 1 is provided in Appendix C and Corollary 1. Corollary 1 further shows that this condition is compatible with standard zero-output residual blocks: it reduces to requiring that the new residual features are not orthogonal to the remaining activation-gradient signal.

Condition 2 (selection condition). We do not attempt to prove convergence or algorithm-specific properties of the optimizer. Instead, Condition 2 isolates the optimization issue from the statistical

risk-comparison argument. Let $\widetilde{f} \in \mathcal{H}_{\text{new}}$ denote the comparison jumpboard model, $\widetilde{f}_S$ for empirical train/test route. Let $f_{\text{alg}} \in \mathcal{H}_{\text{new}}$ be the model returned by the training algorithm on the training dataset. Here, f_{alg} may denote either a scratch-trained model or a growth-trained model, depending on the experimental/training setting. We define the final selected model by the fallback rule $f_{\text{new}} \in \arg\min_{f \in \{f_{\text{alg}}, \widetilde{f}\}} \mathcal{L}_{\text{train}}(f)$. Then, by construction, $\mathcal{L}_{\text{train}}(f_{\text{new}}) \le \mathcal{L}_{\text{train}}(\widetilde{f})$. Equivalently, there exists $\Delta_{\text{ERM}} := \mathcal{L}_{\text{train}}(\widetilde{f}) - \mathcal{L}_{\text{train}}(f_{\text{new}}) \ge 0$ such that $\mathcal{L}_{\text{train}}(f_{\text{new}}) = \mathcal{L}_{\text{train}}(\widetilde{f}) - \Delta_{\text{ERM}}$. When $\Delta_{\text{ERM}} = 0$, the selected model matches the jumpboard model on the training sample. When $\Delta_{\text{ERM}} > 0$, the training algorithm extracts an additional empirical gain from the expanded class.

This is a very weak no-loss condition. If the model returned by the training algorithm performs no better than the jumpboard model on the training set, we may simply select the jumpboard model itself as f_{new}. Therefore the optimization comparison does not require global convergence or a strong optimization guarantee; it only formalizes the fact that the final selected model is not worse, on the training sample, than the near-identity jumpboard comparison model. For the population-risk route, this condition is an analytical device.

4 Main Results

Throughout this section, assume Assumptions 1 to 3 and either Assumption 4 or Assumption 4ϵ, together with Conditions 1 and 2. For each route, f_{new} denotes the final selected model obtained from Condition 2 with the corresponding comparison jumpboard model $\widetilde{f}_S$.

Theorem 1 (Population-risk comparison with a random jumpboard margin)**.** *Let $\tilde{f}_S \in H_{\text{new}}$ be the training-set jumpboard model, allowing $\tilde{f}_S = f^*_{\text{old}}$ in the degenerate case, and let f_{new} be selected according to Condition 2 with comparison model $\tilde{f}_S$. Define the random population margin $\Delta_R^S := R(f^*_{\text{old}}) - R(\tilde{f}_S)$. Then, with probability at least $1-\delta$ over S_{train},*

$$R(f_{\text{new}}) \le R(f^*_{\text{old}}) - \Delta_R^S - \Delta_{\text{ERM}} + 2\epsilon_M$$

On the same event, if $\Delta_R^S + \Delta_{\text{ERM}} \ge 2\epsilon_M$, then $R(f_{\text{new}}) \le R(f^*_{\text{old}})$, with strict inequality whenever $\Delta_R^S + \Delta_{\text{ERM}} > 2\epsilon_M$.

Theorem 2 (Direct train/test comparison with a random test margin)**.** *Let $\widetilde{f}_S \in \mathcal{H}_{\text{new}}$ be the jumpboard model constructed from a first-order descent direction, and suppose $\mathcal{L}_{\text{train}}(f_{\text{new}}) \le \mathcal{L}_{\text{train}}(\widetilde{f}_S) - \Delta_{\text{ERM}}$, $\Delta_{\text{ERM}} \ge 0$. Define the finite-test-set random margin $\Delta_R^{\text{test}} := \mathcal{L}_{\text{test}}(f^*_{\text{old}}) - \mathcal{L}_{\text{test}}(\widetilde{f}_S)$. Then, with probability at least $1-3\delta$,*

$$\mathcal{L}_{\text{test}}(f_{\text{new}}) \le \mathcal{L}_{\text{test}}(f^*_{\text{old}}) - \Delta_R^{\text{test}} - \Delta_{\text{ERM}} + 2(\epsilon_M + \epsilon_K).$$

Consequently, on the event $\Delta_R^{\text{test}} + \Delta_{\text{ERM}} > 2(\epsilon_M + \epsilon_K)$, one has $\mathcal{L}_{\text{test}}(f_{\text{new}}) < \mathcal{L}_{\text{test}}(f^*_{\text{old}})$.

Sampling convention for the alignment analysis. For Theorem 4 and all uses of its alignment bound, including the probabilistic interpretation in Corollary 3, we assume that, conditional on the reference network, its decomposition, and the designated insertion layer, S_{train} and S_{test} are independent with distributions $\mathcal{D}^M$ and $\mathcal{D}^K$, respectively. The gradient moments, covariance conditions, and alignment probabilities in this analysis are interpreted under this conditioning. This independent-reference sampling requirement applies only to the alignment analysis, not to the comparison inequalities in Theorems 1 and 2, provided that their uniform generalization bounds hold over a fixed norm-controlled hypothesis class containing all compared models.

Theorem 2 does not assume that Δ_R^{test} is deterministically positive. For the training-set jumpboard model $\tilde{f}_S$, positivity of this finite-sample margin is supported probabilistically by the empirical train/test alignment mechanism, i.e., $\mathbb{P}(\mu^\top g \le 0)$. We show this in Appendix H, Corollary 3. The relevant failure event has two sources: the empirical direction μ may fail to align with the population direction $\bar{\mu}$, and the independent test direction g may fluctuate away from $\bar{\mu}$. Under the covariance-control condition in Appendix F, the empirical alignment bound gives $\mathbb{P}(\mu^\top g \le 0) \le \frac{4C_\Sigma \tau^2}{M\|\bar{\mu}\|^2} + \frac{4C_\Sigma \tau^2}{K\|\bar{\mu}\|^2} + \frac{4C_\Sigma \tau^2 \operatorname{tr}(\Sigma)}{KM\|\bar{\mu}\|^4}$. Therefore the probability that the training-selected direction remains positively aligned with the independent test direction increases with both the training size M, the

test size K, and the width-dependent signal $\|\bar{\mu}\|^2$. In the uniformly active regime $\|\bar{\mu}\|^2 = \Theta(N)$, this becomes $\mathbb{P}(\mu^\top g \le 0) = O\big(\frac{1}{MN} + \frac{1}{KN}\big)$, up to the mixed covariance term. Thus width does not make the finite-test margin deterministic; rather, it increases the probability that the empirical jumpboard gain is visible on the test set, while the training size controls the reliability of the training-selected direction. The theorems first reveal a qualitative scaling mechanism. On this basis, if one further introduces model-dependent assumptions, this qualitative mechanism is possible to be converted into the power-law form commonly observed in empirical scaling laws. Specifically, in Appendix M we present one possible assumption-based route.

5 Depth, Width, Data, and Scaling Behavior

Theorem 2 provides a direct criterion for test-set scaling. In particular, on the corresponding high-probability event, $L_{\text{test}}(f_{\text{new}}) \le L_{\text{test}}(f^*_{\text{old}}) - \Delta_{\text{R}}^{test} - \Delta_{\text{ERM}} + 2(\epsilon_M + \epsilon_K)$. Hence, a strict scaling improvement occurs whenever $\Delta_{\text{test}} + \Delta_{\text{ERM}} > 2(\epsilon_M + \epsilon_K)$. This inequality exposes the distinct roles of the scaling variables in our framework. Depth creates new residual directions and hence the first-order jumpboard gain Δ_{test}; subsequent optimization contributes the additional gain Δ_{ERM}, which becomes increasingly important as the first-order signal weakens near the deepest-model regime. Data controls the generalization cost through ϵ_M and ϵ_K. Finally, width and sample size jointly determine how reliably a training-side improving direction remains favorable on the test set through the train–test alignment probability. Based on this decomposition, we next examine how depth, width, and data must scale jointly for test-performance scaling to persist.

Depth. The role of depth in our framework is existential rather than merely parametric. Depth is the source of *new improving directions*: it enlarges the function class in a way that can be converted into a smaller population risk and, under the main theorems, into improved test performance. This mechanism need not persist with uniform strength. As the model becomes deeper and better adapted, the available first-order descent signal may shrink, and Condition 1 may eventually fail at the relevant insertion layers. This motivates a functional notion of the *deepest model*: a model is deepest if inserting additional residual blocks can no longer improve performance through the first-order mechanism. Equivalently, the model has already reached first-order saturation at the relevant layers, so that infinitesimal perturbations of the intermediate representations no longer decrease the loss to first order. In this regime, further depth may still enlarge the hypothesis class formally, but it no longer creates a useful first-order descent direction from zero initialization. Equivalently, the deepest-model regime is reached when Condition 1 fails at every admissible insertion layer: the activation-gradient signal is zero, or no zero-output residual block can generate a tangent direction non-orthogonal to the remaining activation gradient.

Data. Data determines whether the gain created by model expansion survives statistical uncertainty. Under the norm-controlled framework of Lemma 1, the generalization term decays at the standard $M^{-1/2}$ statistical rate, but with constants that depend on the architecture through norm, depth, width, and normalization factors. When these architecture-dependent factors are controlled or treated as fixed, the bound reduces to the familiar shorthand $\widetilde{O}(\sqrt{d/M})$, where d is the effective parameter dimension. When depth, width, or normalization constants are themselves scaling variables, the corresponding architecture-dependent factors must be kept explicit; see Appendix F.1 for the resulting data requirement. In other words, model expansion is beneficial only when the growth of data is fast enough to offset the statistical price of the enlarged hypothesis class.

Depth–width coupling. Near the deepest-model regime, the remaining first-order improvement signal becomes small. In Theorem 2, the main difficulty becomes whether the favorable direction is still visible at finite sample size. This is quantified by Theorem 4 in the Appendix E. As $\|\bar{\mu}\| \to 0$, train/test directional alignment becomes harder to observe reliably, even if a beneficial direction still exists. Under the uniformly active regime, $\|\bar{\mu}\|^2 = \Theta(N)$, so the bound improves to $\mathbb{P}(\mu^\top g \le 0) = O\big(\frac{1}{KN} + \frac{1}{MN}\big)$. Therefore, as depth approaches the deepest-model regime, width acts as a compensating factor by improving finite-sample observability of weak first-order gains. This is why our theory naturally leads to a depth–width coupling: deeper models are not guaranteed to improve at fixed width, but can continue to improve in a jointly scaled regime where width and data grow sufficiently fast.

6 EXPERIMENTS

This section empirically examines the main theoretical ingredients of our framework. Our purpose is not to claim state-of-the-art performance, but to test whether the empirical phenomena predicted by the theory are visible in a controlled setting: near-diagonal activation-gradient covariance, test-loss performance with scale, the validation of the depth-expansion Mechanism, decay of first-order signals near the deepest-model regime.

For architectures, in ResNet-based experiments, we use residual networks of depth $6n + 2$ on CIFAR-10 and CIFAR-100, and enlarge width by multiplying the number of channels in each residual stage by a width factor. Unless otherwise stated, the normalization layer is chosen to match the theoretical setting in the main text. In width-observability experiments, we evaluate this prediction using width-controlled 32-layer LLaMA-style Transformer models on C4. The model depth, training corpus, token budget, and optimization protocol are fixed, while the hidden width is varied. Activation gradients are measured at the final Transformer block. For each sequence, we use the sequence-level activation-gradient signal obtained by summing the token-level gradients at the insertion representation. For training details, all models are trained under the same optimization setup, including optimizer type, learning-rate schedule, batch size, data preprocessing, and augmentation pipeline, which are shown in Appendix N. Appendix O discusses more external Transformer evidences from prior work.

6.1 EMPIRICAL SUPPORT FOR ASSUMPTION 3

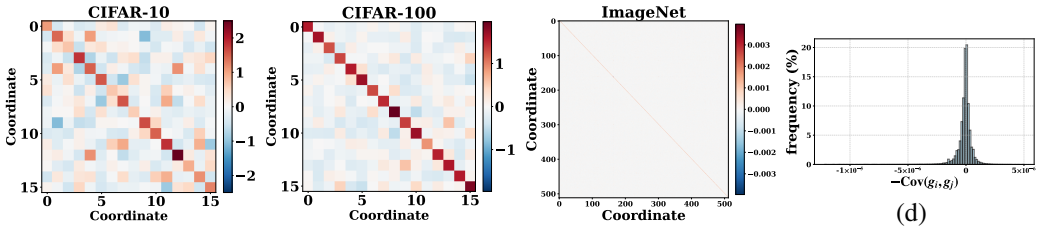


Figure 1: Empirical support for Assumption 3. **Figure (a),(b),(c)**: A representative covariance heatmap of activation gradients. The activation before the fully connected layer. ((a) : CIFAR-10, ResNet-8, base channel 4; (b) : CIFAR-100, ResNet-8, base channel 4; (c) :ImageNet: ResNet-18, channel 64). **Figure (d):** Distribution of sampled off-diagonal covariance entries for a higher-dimensional case (ResNet-56, channel 16, activation before the first ResNet block, sampling $2 * 10^7$ pairs). All figures suggest that activation-gradient covariance is close to diagonal in practice.

This experiment provides empirical evidence for the covariance-control behavior required by Assumption 3. Rather than directly verifying the full spectral condition, we test a stronger and more interpretable empirical property: whether different coordinates of the activation gradient are nearly uncorrelated across samples. If so, the covariance matrix is close to diagonal, and its spectral behavior is governed mainly by coordinate-wise variances rather than by large cross-coordinate correlations.

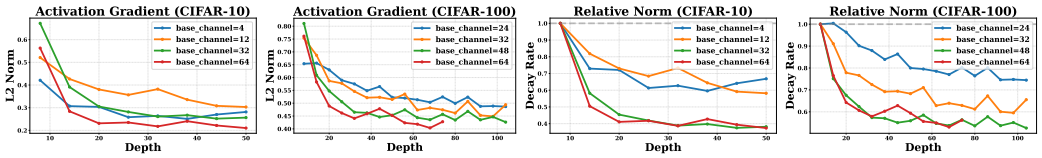
To test this, we fix a target layer l^*, compute sample-level activation gradients for a trained model, and examine their empirical covariance structure. Since storing the full covariance matrix is expensive at larger widths, we use two complementary visualizations: Figure 1 (a) (b) and (c) show a representative covariance heatmap for a low-dimensional case, while Figure 1 (d) shows the distribution of sampled off-diagonal covariance entries over a much larger set of coordinate pairs. Besides CIFAR datasets, we also show ImageNet results in this part.

The results in Figure 1 show a clear and consistent pattern: the off-diagonal covariance entries are sharply concentrated near zero, with only a very small tail away from the origin. Empirically, this indicates that the off-diagonal part of the activation-gradient covariance matrix is extremely small, so the covariance structure is close to diagonal in practice. The heatmap supports this interpretation visually, while the histogram shows that near-zero off-diagonal covariance is not an isolated artifact of a few coordinates but a dominant distributional phenomenon.

From the perspective of our theory, this observation provides empirical consistency with evidence for Assumption 3. If the covariance matrix is approximately diagonal, then its spectral behavior

is controlled primarily by coordinate-wise variances, making the spectral-control condition used in Theorem 4 in Appendix E substantially more plausible in trained normalized residual networks. We emphasize that the experiment is not intended to prove that Σ is exactly diagonal in a strict mathematical sense. Rather, it shows that the off-diagonal correlations are sufficiently small in practice for the "approximately diagonal, spectrally controlled" approximation underlying our width-alignment analysis to be credible.

6.2 Decay of first-order activation-gradient signals with depth



(a) (b) (c) (d)

Figure 2: Decay of first-order average activation-gradient signals with depth. **(a),(b)**: absolute ℓ_2-norm of the activation gradient immediately before the final fully connected layer. **(c),(d)**: the same quantity after width-wise normalization by the shallowest model. The absolute plot shows that the remaining first-order signal decreases as depth increases, while the normalized plot shows that this decay pattern is structurally consistent across widths.

This experiment is designed to probe the weakening of the first-order improvement signal as the model becomes deeper. According to the discussion in the main text, approaching the deepest-model regime should be accompanied by a shrinkage of the activation-gradient signal, which makes additional depth harder to exploit through the first-order mechanism.

To test this phenomenon, we consider ResNet models with different depths and varying the base channel width. For each trained model, we select the activation immediately before the final fully connected layer and compute the corresponding activation gradient with respect to the loss. We use the ℓ_2-norm of the average activation gradient as an empirical proxy for the remaining first-order signal. In addition to the absolute norm, we also report a width-wise normalized version, obtained by dividing each curve by its value at the shallowest depth, in order to isolate the relative decay pattern across widths.

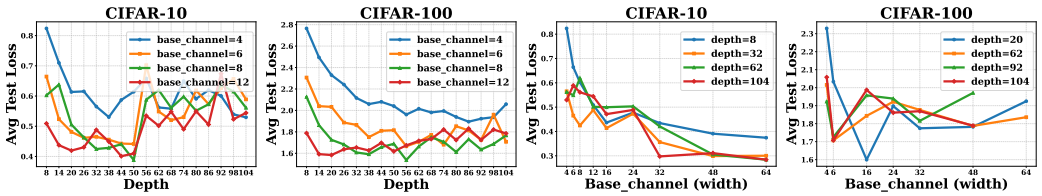
Figure 2 shows two complementary views of this phenomenon. The Figure 2(a), (b) report the absolute activation-gradient norm, while the Figure 2(c), (d) shows the corresponding relative decay after width-wise normalization. In the absolute plot, the overall trend is clear: as depth increases, the activation-gradient norm decreases substantially across all tested widths, indicating that deeper models carry a weaker remaining first-order improvement signal. This directly supports the deepest-model interpretation in the main text.

From the viewpoint of the theory, Figure 2 supports the claim that near the deepest-model regime, the population-level first-order signal becomes small, so the gain available from further depth expansion weakens. This helps explain why strict improvement becomes harder to certify in very deep models, and why width becomes increasingly important as a compensating factor in the weak-signal regime.

6.3 Test loss under joint depth and width scaling

This experiment is designed to test whether the empirical test-loss behavior is consistent with the qualitative predictions of the main theorems, namely that depth expansion can improve test performance in a jointly scaled regime, while width stabilizes the finite-sample observability of such gains. Figure 3 provides two complementary views: test loss as a function of depth at several widths, and test loss as a function of width at several depths. Curves report the mean over 5 random seeds.

The first observation is that the overall trend is favorable: increasing depth and width often reduces test loss, which is consistent with the theorem-level prediction that model expansion can improve test performance when the available gain dominates the statistical cost. At the same time, for CIFAR-10, this improvement is clearly non-monotonic. After an initial improvement, the test loss rebounds noticeably at larger depths. This suggests that increasing depth alone is not sufficient; when width



(a) (b) (c) (d)

Figure 3: Test loss under joint depth and width scaling. (a),(b) show test loss as a function of depth at fixed widths; (c), (d) show test loss as a function of width at fixed depths.

does not grow adequately with depth, the test-side benefit of further deepening becomes unstable and may partially disappear. For CIFAR-100, we do not observe the same non-monotonic rebound within the tested depth range, possibly because the maximum depth explored here is not yet large enough.

The second observation is that width acts as a stabilizing factor rather than as an unlimited source of gain. In the width-scaling plot, increasing width produces a substantial reduction in test loss in the narrow-to-moderate regime, especially for deeper models. However, once width becomes sufficiently large, the curves flatten and further widening yields only marginal improvement. This saturation behavior is consistent with our theoretical picture.

6.4 Direct Validation of the Depth-Expansion Mechanism

Table 1: Direct comparison of the old model, the constructed jumpboard model, the fully growth-trained model, and the same target architecture trained from scratch on CIFAR-10. For each growth step, the old model is the preceding growth-trained network. The jumpboard model is obtained by constructing a target hidden-space perturbation from the activation gradient and step size 10^{-3}.

	Train Loss				Test Loss			
Growth step	Old	Jumpboard	Growth	Scratch	Old	Jumpboard	Growth	Scratch
ResNet-10 → ResNet-11	0.0070	0.0069	0.0067	0.0082	0.2589	**0.2566**	0.2626	0.2874
ResNet-11 → ResNet-12	0.0067	0.0064	0.0062	**0.0059**	0.2626	0.2623	**0.2580**	0.2822
ResNet-12 → ResNet-13	0.0062	0.0059	**0.0057**	0.0059	0.2580	0.2557	**0.2510**	0.2948

We directly examine the sequence $f_{\text{old}} \to \widetilde{f}_S \to f_{\text{growth}}$, while using a scratch-trained model with the same target architecture as a control. For each expansion, we first determine a target hidden-space perturbation δ_S from the empirical activation gradient and a small step size. Letting $z_i = f_{\text{bot}}(x_i)$ denote the inputs to the inserted residual block, we construct the jumpboard by solving $\theta_{\text{JB}} = \arg\min_\theta \left\| \frac{1}{M} \sum_{i=1}^{M} h_\theta(z_i) - \delta_S \right\|_2^2$. This auxiliary optimization realizes the prescribed first-order perturbation using an actual residual block and is distinct from the subsequent end-to-end growth training.

Table 1 shows that the constructed jumpboard consistently improves over the corresponding old model, providing direct evidence that the activation-gradient-derived direction can be realized by a parameterized residual block and translated into a test-side gain. Subsequent growth training further reduces the training objective, while its test-side effect depends on whether the additional optimization gain survives finite-sample generalization effects. This supports the separation between representational gain, optimization gain, and generalization transfer in our framework.

Finally, the growth-trained models consistently achieve better test performance than models with the same target architectures trained from scratch. This suggests that the observed improvement is not explained by increased architectural depth alone, but also depends on the structured expansion path through the jumpboard.

6.5 Width–Data Scaling of Train–Test Alignment

Our theory predicts that width and sample size jointly control whether a training-identified first-order direction remains favorable on an independent test sample. In particular, Theorem 4 controls the

Table 2: Empirical probability of positive train–test activation-gradient alignment, $\Pr(\mu^\top g > 0)$, under different model widths and finite-sample budgets.

$M = K$	$N = 512$	1024	1536	2048	2560	3072
1	71.5%	72.0%	72.8%	74.4%	75.2%	77.3%
2	84.3%	84.9%	85.7%	87.4%	88.6%	90.3%
4	94.8%	95.2%	96.6%	97.3%	98.0%	98.6%
8	99.3%	99.6%	99.7%	99.8%	$> 99.9\%$	$> 99.9\%$

Table 3: Width dependence of the empirical first-order activation-gradient signal on C4. $\widehat{S} = \|\widehat{\mu}\|_2^2$ denotes the squared norm of the empirical mean activation gradient. Values are rounded representative central estimates across evaluation runs.

Width N	512	1024	1536	2048	2560	3072
$\widehat{S}$	0.15	0.18	0.20	0.22	0.24	0.26

misalignment probability $\Pr(\mu^\top g \le 0)$, suggesting that positive train–test alignment should become more reliable as both representation width and sample size increase.

For each model, we independently sample two non-overlapping subsets of sizes M and K to estimate the training- and test-side activation gradients, $\mu_M = \frac{1}{M}\sum_{i=1}^{M} q_i, \qquad g_K = \frac{1}{K}\sum_{j=1}^{K} \widetilde{q}_j$. We then estimate the probability of positive alignment $\Pr(\mu_M^\top g_K > 0)$ by repeated independent resampling. We use $M = K \in \{1, 2, 4, 8\}$, with each sampled sequence containing 2048 tokens. For every width and sampling-budget configuration, the alignment probability is estimated from $500{,}000$ Monte Carlo trials.

Table 2 shows that positive train–test alignment becomes more reliable with both width and sample size. At a fixed sampling budget, increasing width improves directional reliability, while at a fixed width, increasing the sample size reduces finite-sample estimation uncertainty. These observations support the joint width–data mechanism in our analysis: width improves the observability of favorable first-order directions, whereas additional data stabilizes their estimation on finite samples.

6.6 Width Scaling of the First-Order Signal

Our width-dependent analysis relies on the intuition that a wider representation can support a stronger aggregate first-order signal. We therefore examine how the empirical mean activation-gradient signal changes with hidden width. Using the width-controlled LLaMA-style models described above, we compute $\widehat{\mu} = \frac{1}{S}\sum_{i=1}^{S} q_i, \qquad \widehat{S} = \|\widehat{\mu}\|_2^2$, where q_i is the sequence-level activation gradient measured at the final Transformer block. All evaluations in this experiment are conducted on C4.

Table 3 shows a clear increase in the aggregate activation-gradient signal as the hidden width grows. The trend persists across the investigated width range, indicating that wider representations retain a stronger empirical first-order signal. This observation is consistent with the width-supported regime used in our analysis and provides a complementary explanation for the improved train–test directional reliability observed in Section 6.5: width not only improves the finite-sample stability of the direction, but is also associated with a stronger observable first-order signal.

7 Conclusion

In this paper, we develop a theorem-driven account of when depth expansion in normalized residual networks improves test risk. Our framework links three ingredients: a jumpboard model with lower population risk, optimization gain, and generalization transfer. This yields two complementary qualitative scaling-mechanism results: a tighter population-risk route and a more robust train/test route. The analysis also clarifies that depth creates improving directions, width preserves their finite-sample observability, and data pays for the statistical cost of model expansion.

AI USE STATEMENT

We used generative AI tools to assist with manuscript translation and language editing, plotting-code development, and checks of the manuscript and figures against experimental logs. These tools also provided feedback on the presentation of the methodology and experimental comparisons and helped summarize and interpret selected results. We take full responsibility for the final content of this paper.

REFERENCES

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International conference on machine learning*, pp. 244–253. PMLR, 2018a.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International conference on machine learning*, pp. 254–263. PMLR, 2018b.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. In *Uncertainty in artificial intelligence*, pp. 1352–1361. PMLR, 2021.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In *International conference on machine learning*, pp. 342–350. PMLR, 2017.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

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.

Bo Chang, Lili Meng, Eldad Haber, Lars Ruthotto, David Begert, and Elliot Holtham. Reversible architectures for arbitrarily deep residual neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Soham De and Sam Smith. Batch normalization biases residual blocks towards the identity function in deep networks. *Advances in Neural Information Processing Systems*, 33:19964–19975, 2020.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pp. 907–940. PMLR, 2016.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference on learning theory*, pp. 297–299. PMLR, 2018.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34 (1):014004, 2018.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016b.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, DDL Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 10, 2022.

Jinshu Huang, Mingfei Sun, and Chunlin Wu. On the generalization behavior of deep residual networks from a dynamical system perspective. *arXiv preprint arXiv:2602.20921*, 2026.

Aku Kammonen, Jonas Kiessling, Petr Plecháč, Mattias Sandberg, Anders Szepessy, and Raúl Tempone. Smaller generalization error derived for a deep residual neural network compared to shallow networks. *arXiv preprint arXiv:2010.01887*, 2020.

Aku Kammonen, Jonas Kiessling, Petr Plecháč, Mattias Sandberg, Anders Szepessy, and Raul Tempone. Smaller generalization error derived for a deep residual neural network compared with shallow networks. *IMA Journal of Numerical Analysis*, 43(5):2585–2632, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Yoav Levine, Noam Wies, Or Sharir, Hofit Bata, and Amnon Shashua. The depth-to-width interplay in self-attention. *arXiv preprint arXiv:2006.12467*, 2020.

Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, and Lexing Ying. A mean field analysis of deep resnet and beyond: Towards provably optimization via overparameterization from depth. In *International Conference on Machine Learning*, pp. 6426–6436. PMLR, 2020.

Chao Ma, Qingcan Wang, et al. Rademacher complexity and the generalization error of residual networks. *Communications in Mathematical Sciences*, 18(6):1755–1774, 2020.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*, 2019.

Matus Telgarsky. Benefits of depth in neural networks. In *Conference on learning theory*, pp. 1517–1539. PMLR, 2016.

Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. *Advances in neural information processing systems*, 29, 2016.

Peihao Wang, Rameswar Panda, Lucas Torroba Hennigen, Philip Greengard, Leonid Karlinsky, Rogerio Feris, David Daniel Cox, Zhangyang Wang, and Yoon Kim. Learning to grow pretrained models for efficient transformer training, 2023. URL `https://arxiv.org/abs/2303.00980`.

Ge Yang and Samuel Schoenholz. Mean field residual networks: On the edge of chaos. *Advances in neural information processing systems*, 30, 2017.

Boyang Zhang, Daning Cheng, Yunquan Zhang, Fangming Liu, and Wenguang Chen. Compression for better: A general and stable lossless compression framework. *arXiv preprint arXiv:2412.06868*, 2024.

Hongyi Zhang, Yann N Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. *arXiv preprint arXiv:1901.09321*, 2019.

A Related Work

Table 4 positions our contribution relative to the closest prior theory work. Existing results mainly fall into three categories: broad explanations of neural scaling laws, generalization bounds for residual architectures, and optimization or mean-field analyses of deep ResNets. Our paper studies a different target: an architecture-specific old-vs-new comparison problem under normalized residual depth expansion. The key distinction is the combination of three ingredients within one framework: a constructive representational-gain step, an optimization comparison step, and a generalization-transfer step, which together yield two complementary test-risk comparison routes.

Prior work	Main focus	Difference from our setting	Difference from our paper
Bahri et al. (2024)	A general theory of neural scaling laws and scaling regimes for trained deep networks.	Not specific to normalized residual architectures; does not study constructive *depth expansion by inserting a residual block* into a trained reference model; does not provide an old-vs-new finite-sample *test-risk comparison theorem*.	We study an *architecture-specific* question: when a normalized residual network is enlarged by inserting one residual block, under what conditions does the resulting model provably avoid worsening or strictly improve *test risk*?
Ma et al. (2020)	Rademacher-complexity and generalization-error bounds for residual networks.	Provides complexity control, but does not prove that depth expansion creates a strictly better comparison model, nor connect such existence to optimization output and then to a final old-vs-new test-risk inequality.	Our norm-based generalization ingredient is embedded into a larger framework that additionally includes a constructive representational-gain step and a final test-risk comparison step.
Kammonen et al. (2023)	Generalization-error analysis showing that a deep residual network can enjoy a smaller bound than a shallow alternative in a specialized random-feature setting.	The setting is specialized and is not formulated as residual-block insertion into a trained model; it does not provide our two-route comparison between old and expanded models on test risk.	Our contribution is not merely that deeper residual models may generalize better, but that *depth expansion from a trained reference model* can be translated into explicit old-vs-new test-risk inequalities.
Lu et al. (2020)	Mean-field analysis and favorable optimization/global-convergence properties for deep ResNets in an overparameterized regime.	Focuses on optimization and mean-field dynamics rather than finite-sample scaling-law comparisons after residual-block insertion; does not provide a representation–optimization–generalization decomposition ending in test-risk comparison.	Our focus is to bridge constructive representational gain, optimization gain, and generalization transfer, rather than to prove a mean-field global-optimization result.
Huang et al. (2026)	Generalization bounds for discrete- and continuous-time ResNets from a dynamical-systems perspective.	Still mainly a generalization-bound paper; it does not formulate the constructive block-insertion improvement question or provide our complementary population-risk and direct train/test comparison routes.	Our paper studies a narrower but different target: *architecture-specific test-risk improvement under normalized residual depth expansion*, including a direct train/test route that remains meaningful in degenerate regimes.

Table 4: Comparison with the closest prior theory work. Existing theory studies global neural scaling laws, residual-network generalization bounds, and optimization or mean-field analyses of deep ResNets. Our paper instead focuses on an architecture-specific question: when inserting a residual block into a trained normalized residual network can be converted into a provable non-worsening or strict improvement in test risk.

B NOTATION

For convenience, we collect here the notation used throughout the appendix. Unless otherwise stated, $\|\cdot\|$ denotes the Euclidean norm for vectors, $\|\cdot\|_\sigma$ denotes the spectral norm of a matrix, and $\|\cdot\|_F$ denotes the Frobenius norm.

Data, losses, and risks. Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$. The training set and test set are

$$S_{\text{train}} = \{(x_i, y_i)\}_{i=1}^M \sim \mathcal{D}^M, \qquad S_{\text{test}} = \{(\tilde{x}_j, \tilde{y}_j)\}_{j=1}^K \sim \mathcal{D}^K,$$

where the two samples are independent and i.i.d. The loss function is

$$\ell : \mathbb{R}^C \times \mathcal{Y} \to [0, B_\ell].$$

For a predictor f, we define

$$\mathcal{L}_{\text{train}}(f) \triangleq \frac{1}{M}\sum_{i=1}^M \ell(f(x_i), y_i), \qquad \mathcal{L}_{\text{test}}(f) \triangleq \frac{1}{K}\sum_{j=1}^K \ell(f(\tilde{x}_j), \tilde{y}_j),$$

and the population risk

$$R(f) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)].$$

Network decomposition. At a designated intermediate layer l^*, the model is decomposed as

$$f = f_{\text{top}} \circ f_{\text{bot}},$$

where

$$f_{\text{bot}} : \mathcal{X} \to \mathbb{R}^N, \qquad f_{\text{top}} : \mathbb{R}^N \to \mathbb{R}^C.$$

Here N is the hidden width at layer l^*, and C is the output dimension. For a training sample x_i, its hidden representation at layer l^* is

$$x_i^{(l^*)} \triangleq f_{\text{bot}}(x_i).$$

Similarly, for a test sample $\tilde{x}_j$, we write

$$\tilde{x}_j^{(l^*)} \triangleq f_{\text{bot}}(\tilde{x}_j).$$

Gradient quantities. The average training activation gradient at layer l^* is

$$\mu \triangleq \frac{1}{M}\sum_{i=1}^M \nabla_{x_i^{(l^*)}} \ell\Big(f_{\text{top}}(x_i^{(l^*)}), y_i\Big) \in \mathbb{R}^N.$$

Its population counterpart is

$$\bar{\mu} \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\nabla_{x^{(l^*)}} \ell\Big(f_{\text{top}}(x^{(l^*)}), y\Big)\Big] \in \mathbb{R}^N.$$

For the j-th test sample, define the single-sample test gradient

$$\zeta^{(j)} \triangleq \nabla_{\tilde{x}_j^{(l^*)}} \ell\Big(f_{\text{top}}(\tilde{x}_j^{(l^*)}), \tilde{y}_j\Big),$$

and the average test gradient

$$g \triangleq \frac{1}{K}\sum_{j=1}^K \zeta^{(j)}.$$

In Theorem 4, Σ denotes the covariance matrix of the single-sample activation gradient, namely

$$\Sigma \triangleq \text{Cov}(\zeta),$$

and τ^2, C_Σ are constants such that $\lambda_{\max}(\Sigma) \le C_\Sigma \tau^2$.

Hypothesis classes and model expansion. The old hypothesis class is

$$\mathcal{H}_{\text{old}} = \{x \mapsto f_{\text{top}}(f_{\text{bot}}(x)) : f_{\text{top}} \in \mathcal{F}_{\text{top}},\ f_{\text{bot}} \in \mathcal{F}_{\text{bot}}\}.$$

A trained model in the old class is denoted by

$$f^*_{\text{old}} \in \mathcal{H}_{\text{old}}.$$

After inserting a residual block h after layer l^*, the new hypothesis class is

$$\mathcal{H}_{\text{new}} = \left\{x \mapsto f_{\text{top}}\big(f_{\text{bot}}(x) + h(f_{\text{bot}}(x))\big) : f \in \mathcal{H}_{\text{old}},\ h \in \mathcal{F}_{\text{res}}\right\}.$$

Here $\mathcal{F}_{\text{res}}$ denotes the residual-block function class. The new trained model produced by the optimizer or fallback rule in Condition 2 is denoted by

$$f_{\text{new}} \in \mathcal{H}_{\text{new}}.$$

Given a comparison jumpboard model $\widetilde{f}$, the final selected model is

$$f_{\text{new}} \in \arg\min_{f \in \{f_{\text{alg}}, \widetilde{f}\}} \mathcal{L}_{\text{train}}(f).$$

Thus f_{new} is the final selected model used in the main theorems, not necessarily the raw optimizer output.

Residual-block parameterization and jumpboard models. We parameterize the inserted residual block by θ, writing it as h_θ, with zero initialization $h_0 = 0$. We distinguish two comparison jumpboard models.

For the population-risk route, let v_{pop} be a population first-order descent direction. Along the path

$$\theta(\eta) = \eta v_{\text{pop}},$$

the population-risk profile is

$$G_{\text{pop}}(\eta) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\Big(f_{\text{top}}\big(x^{(l^*)} + h_{\eta v_{\text{pop}}}(x^{(l^*)})\big), y\Big)\right].$$

For a sufficiently small step size $\eta_{\text{pop}} > 0$, the corresponding population jumpboard model is

$$\widetilde{f}_{\text{pop}}(x) := f_{\text{top}}\big(f_{\text{bot}}(x) + h_{\eta_{\text{pop}} v_{\text{pop}}}(f_{\text{bot}}(x))\big).$$

For the direct train/test route, let v_S be an empirical first-order descent direction selected on S_{train}. Along the path

$$\theta(\eta) = \eta v_S,$$

we define the training-set jumpboard model by

$$\widetilde{f}_S(x) := f_{\text{top}}(f_{\text{bot}}(x) + h_{\eta_S v_S}(f_{\text{bot}}(x))),$$

where $\eta_S > 0$ is chosen so that the empirical training loss decreases along this direction. The model $\widetilde{f}_S$ is random through its dependence on S_{train}.

Improvement margins. The population jumpboard margin used in the population-risk route is

$$\Delta_R := R(f^*_{\text{old}}) - R(\widetilde{f}_{\text{pop}}) > 0.$$

The empirical training-set jumpboard margin used in the direct train/test route is

$$\Delta_{\text{train},S} := L_{\text{train}}(f^*_{\text{old}}) - L_{\text{train}}(\widetilde{f}_S) > 0.$$

Given the comparison jumpboard model

$$\widetilde{f} = \begin{cases} \widetilde{f}_{\text{pop}}, & \text{in the population-risk route,} \\ \widetilde{f}_S, & \text{in the direct train/test route,} \end{cases}$$

the optimization gain is

$$\Delta_{\text{ERM}} := L_{\text{train}}(\widetilde{f}) - L_{\text{train}}(f_{\text{new}}) \geq 0.$$

The finite-test-set jumpboard margin used in Theorem 2 is

$$\Delta_R^{\text{test}} := L_{\text{test}}(f^*_{\text{old}}) - L_{\text{test}}(\widetilde{f}_S).$$

Rademacher complexity and generic generalization notation. For a function class $\mathcal{G}$ and a sample $S = \{z_i\}_{i=1}^m$, the empirical Rademacher complexity is

$$\hat{\mathfrak{R}}_S(\mathcal{G}) \triangleq \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i) \right],$$

where σ_i are i.i.d. Rademacher random variables. We write

$$\epsilon_{\text{gen}}(\mathcal{F}, m, \delta)$$

for a standard uniform generalization bound for a hypothesis class $\mathcal{F}$.

Norm-based generalization bound for normalized residual networks. The appendix uses the specialized bound from Lemma 1, abbreviated as

$$\epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H}, m, \delta).$$

When the class is $\mathcal{H}_{\text{new}}$, we use the shorthand

$$\epsilon_M \triangleq \epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H}_{\text{new}}, M, \delta), \qquad \epsilon_K \triangleq \epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H}_{\text{new}}, K, \delta),$$

and

$$\epsilon_{TT} \triangleq \epsilon_M + \epsilon_K.$$

In Theorem 5, the total generalization cost appears as $2\epsilon_M$; in Theorem 2, it appears as $2(\epsilon_M + \epsilon_K) = 2\epsilon_{TT}$. The population-level qualitative scaling conclusion is stated in Theorem 7.

Layerwise residual-network notation in Lemma 1. The l-th normalized residual layer is written as

$$T_l(z) = \mathcal{N}_l(z + h_l(z)), \qquad l = 0, \ldots, L-1,$$

where L is the network depth and $\mathcal{N}_l$ denotes the normalization operator (e.g., RMSNorm or LayerNorm). The layer input is denoted by z_l, and the residual branch parameters by W_l.

The structural constants used in Lemma 1 are:

- s_l: the input-Lipschitz constant of the residual branch $h_l(\cdot; W_l)$;
- $L_l^{(p)}$: the parameter-Lipschitz constant of h_l, namely

$$\|h_l(z; W_l) - h_l(z; W_l')\| \le L_l^{(p)} \|W_l - W_l'\|_F \, \|z\|;$$

- $r_{\min} > 0$: the non-degeneracy lower bound before normalization,

$$\inf_{x, \theta, l} \|z_l(x) + h_l(z_l(x); \theta)\| \ge r_{\min};$$

- b_l: the Frobenius-norm radius of the parameter constraint $\|W_l\|_F \le b_l$;
- $\bar{b} \triangleq \max_l b_l$;
- d_l: the parameter dimension of layer l;
- $d \triangleq \sum_{l=0}^{L-1} d_l$: the total parameter dimension;
- B_x: a uniform input norm bound such that $\|x\| \le B_x$;
- $\gamma_{\max}$: the maximum coordinate scale of the normalization layer;
- $B_{\text{norm}} \triangleq \gamma_{\max}\sqrt{N}$: the post-normalization representation norm bound;
- $B_* \triangleq \max\{B_x, B_{\text{norm}}\}$.

Top map and loss regularity constants. The loss ℓ is assumed to be L_ℓ-Lipschitz and bounded by B_ℓ. The top map f_{top} is assumed to be $\mathcal{L}_{\text{top}}$-Lipschitz, with reference-point bound

$$\|f_{\text{top}}(0)\| \le B_0.$$

Layerwise Lipschitz amplification constants. For the normalization operator in layer l, we denote its Lipschitz factor by c_l. The output sensitivity to a perturbation at layer l is summarized by

$$\Lambda_l \triangleq L_{\text{top}} \cdot L_l^{(p)} \cdot B_* \cdot c_l \prod_{k>l} c_k(1+s_k),$$

and

$$\Lambda_{\max} \triangleq \max_{0\le l\le L-1} \Lambda_l.$$

These constants enter the covering-number bound and hence the complexity estimate in Lemma 1.

Covering parameter. In the proof of Lemma 1, $\epsilon_0 > 0$ denotes the covering accuracy for the output/loss class. A common choice is

$$\epsilon_0 = M^{-1/2}.$$

Failure probabilities. We use $\delta, \delta_1, \delta_2, \delta_3$ for generic confidence parameters. In Theorem 5, the total failure probability is decomposed as

$$\delta_{\text{gen}} = \delta_1 + \delta_2, \qquad \delta_{\text{test}}$$

for the generalization and test-concentration parts, respectively.

Standing convention. Throughout the appendix, symbols with subscript "old" refer to the model or hypothesis class before depth expansion, symbols with subscript "new" refer to their counterparts after inserting the new residual block, and a tilde (as in $\tilde{f}$) denotes the auxiliary jumpboard model used only for comparison in the proof.

C Clarification of Condition 1: From Hidden-Space Descent to Residual-Block Realizability

This appendix clarifies the meaning of Condition 1 and explains how it should be interpreted in standard residual architectures. The key point is that Condition 1 combines two logically separate requirements: first, the existence of a descent direction in the hidden representation space; second, the ability of the inserted residual block to realize such a direction to first order near the zero residual function.

Throughout this section, let

$$z = x^{(l^*)} = f_{\text{bot}}(x) \in \mathbb{R}^N,$$

and define the activation-gradient signal

$$q(z,y) := \nabla_z \ell(f_{\text{top}}(z), y) \in \mathbb{R}^N.$$

The old model is

$$f^*_{\text{old}}(x) = f_{\text{top}}(z),$$

whereas the expanded model obtained by inserting a residual block is

$$f_\theta(x) = f_{\text{top}}\big(z + h_\theta(z)\big).$$

C.1 Hidden-space descent direction

The most direct way to understand the first-order mechanism is to first ignore the parameterization of the residual block and ask whether there exists a hidden-space perturbation direction

$$\phi : \mathbb{R}^N \to \mathbb{R}^N$$

such that moving the representation z along $\phi(z)$ decreases the population risk to first order.

For such a perturbation, define

$$G_\phi(\eta) := \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell\big(f_{\text{top}}(z + \eta\phi(z)), y\big)\right].$$

By the local C^1 regularity assumption,

$$G'_\phi(0) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z,y)^\top \phi(z)\right].$$

Therefore, if

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z,y)^\top \phi(z)\right] \neq 0,$$

then, by replacing ϕ with $-\phi$ if necessary, we may assume

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z,y)^\top \phi(z)\right] < 0.$$

It follows that there exists a sufficiently small $\eta_0 > 0$ such that

$$G_\phi(\eta_0) < G_\phi(0).$$

In other words,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\big(f_{\mathrm{top}}(z+\eta_0\phi(z)),y\big)\right] < \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\big(f_{\mathrm{top}}(z),y\big)\right].$$

This is the hidden-space descent mechanism underlying the jumpboard model.

Empirical hidden-space descent direction. The same construction has a training-set analogue. Given $S_{\mathrm{train}} = \{(x_i,y_i)\}_{i=1}^M$, write $z_i = x_i^{(l^*)} = f_{\mathrm{bot}}(x_i)$ and

$$q_i := \nabla_{z_i}\ell(f_{\mathrm{top}}(z_i),y_i).$$

For a hidden-space perturbation $\phi : \mathbb{R}^N \to \mathbb{R}^N$, define the empirical profile

$$G_{\phi,S}(\eta) := \frac{1}{M}\sum_{i=1}^M \ell(f_{\mathrm{top}}(z_i+\eta\phi(z_i)),y_i).$$

Then

$$G'_{\phi,S}(0) = \frac{1}{M}\sum_{i=1}^M q_i^\top \phi(z_i).$$

Hence, if

$$\frac{1}{M}\sum_{i=1}^M q_i^\top \phi(z_i) < 0,$$

then for sufficiently small $\eta_S > 0$,

$$G_{\phi,S}(\eta_S) < G_{\phi,S}(0).$$

This gives the empirical hidden-space descent mechanism used to construct the training-set jumpboard model $\widetilde{f}_S$ in the direct train/test route.

C.2 Residual-block realizability

The previous argument only shows that a hidden-space perturbation $\phi(z)$ would reduce the population risk. To use this direction inside the expanded network, we must require that the inserted residual block can realize such a perturbation, at least to first order near zero initialization.

Let the inserted residual block be parameterized by

$$\theta \in \mathbb{R}^p,$$

and assume that

$$h_0(z) = 0$$

for all relevant hidden states z. For each fixed z, define the parameter Jacobian at zero by

$$J_\theta h_0(z) := \left.\frac{\partial h_\theta(z)}{\partial\theta}\right|_{\theta=0} \in \mathbb{R}^{N\times p}.$$

Here $J_\theta h_0(z)$ is a Jacobian with respect to the residual-block parameters, not with respect to the hidden input z. Given a parameter direction

$$u \in \mathbb{R}^p,$$

the induced first-order hidden-space perturbation is

$$\phi_u(z) := J_\theta h_0(z)u \in \mathbb{R}^N.$$

Equivalently, along the parameter path

$$\theta(\eta) = \eta u,$$

we have the first-order expansion

$$h_{\eta u}(z) = h_0(z) + \eta J_\theta h_0(z)u + r_\eta(z),$$

where the remainder satisfies

$$\frac{\mathbb{E}_{(x,y)\sim\mathcal{D}}\|r_\eta(z)\|}{|\eta|} \to 0 \qquad \text{as } \eta \to 0.$$

Since $h_0(z) = 0$, this becomes

$$h_{\eta u}(z) = \eta\phi_u(z) + r_\eta(z).$$

Thus the inserted residual block realizes the hidden-space perturbation ϕ_u to first order. This is the precise meaning of the parameter-space form of Condition 1.

C.3 Equivalent Jacobian form of Condition 1

Using the notation above, Condition 1 can be equivalently written as follows.

Condition 1' (first-order residual non-degeneracy). There exists a parameter direction $u \in \mathbb{R}^p$ such that

$$a_u := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(x^{(l^*)}, y)^\top J_\theta h_0(x^{(l^*)})u\right] \neq 0.$$

If $a_u > 0$, we replace u by $-u$. Hence, without loss of generality, we may assume

$$a_u < 0.$$

This form makes the dimensions explicit:

$$q(x^{(l^*)}, y) \in \mathbb{R}^N, \qquad J_\theta h_0(x^{(l^*)}) \in \mathbb{R}^{N\times p}, \qquad u \in \mathbb{R}^p.$$

Therefore,

$$J_\theta h_0(x^{(l^*)})u \in \mathbb{R}^N,$$

and the inner product

$$q(x^{(l^*)}, y)^\top J_\theta h_0(x^{(l^*)})u$$

is a well-defined scalar.

Equivalently, Condition 1' states that the first-order tangent class

$$\mathcal{T}_h := \{z \mapsto J_\theta h_0(z)u : u \in \mathbb{R}^p\}$$

is not orthogonal to the population activation-gradient functional

$$\phi \mapsto \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z, y)^\top \phi(z)\right].$$

Importantly, this assumption does not require the residual block to span all possible hidden-space perturbations. It only requires that its first-order tangent class contains at least one direction that has a nonzero projection onto the population activation-gradient signal.

Population version. The population residual-block realizability condition is that there exists a parameter direction v_{pop} such that

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z, y)^\top \left.\frac{\partial h_\theta(z)}{\partial\theta}\right|_{\theta=0} v_{\text{pop}}\right] < 0.$$

This is the condition used to construct the population jumpboard model $\widetilde{f}_{\text{pop}}$.

Training-set version. The empirical residual-block realizability condition is that there exists a parameter direction v_S, possibly depending on S_{train}, such that

$$\frac{1}{M}\sum_{i=1}^{M} q_i^\top \left.\frac{\partial h_\theta(z_i)}{\partial \theta}\right|_{\theta=0} v_S < 0.$$

This is the condition used to construct the training-set jumpboard model $\widetilde{f}_S$. The direction v_S is random through its dependence on the training sample.

C.4 Construction of the jumpboard model

Under Condition 1', choose u such that

$$a_u = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z,y)^\top J_\theta h_0(z) u\right] < 0.$$

Define the parameter path

$$\theta(\eta) = \eta u,$$

and the corresponding expanded model

$$\widetilde{f}_{pop} := f_{\text{top}}\big(z + h_{\eta u}(z)\big).$$

The associated population-risk profile is

$$G(\eta) := R(\widetilde{f}_\eta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\big(f_{\text{top}}(z + h_{\eta u}(z)), y\big)\right].$$

Using

$$h_{\eta u}(z) = \eta J_\theta h_0(z) u + r_\eta(z),$$

and the local differentiability of the loss-composition map, we obtain

$$G'(0) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z,y)^\top J_\theta h_0(z) u\right] = a_u < 0.$$

Therefore, there exists a sufficiently small $\eta_0 > 0$ such that

$$G(\eta_0) < G(0).$$

Equivalently,

$$R(\widetilde{f}_{pop}) < R(f^*_{\text{old}}).$$

The model

$$\widetilde{f}_{pop}(x) = f_{\text{top}}\big(x^{(l^*)} + h_{\eta_0 u}(x^{(l^*)})\big)$$

is the jumpboard model used in the main proof.

By analogy with the procedure described above, we can construct an empirical (training-set) jumpboard model $\widetilde{f}_S$, $\mathcal{L}_{train}(\widetilde{f}_S) < \mathcal{L}_{train}(f^*_{\text{old}})$, for the training set following the same construction logic.

C.5 Relation to standard zero-output residual blocks

We now explain how Condition 1' maps to standard residual blocks. The key observation is that many residual branches can be written in the form

$$h_{U,V}(z) = V\psi_U(z),$$

where

$$z \in \mathbb{R}^N$$

is the input hidden representation,

$$\psi_U : \mathbb{R}^N \to \mathbb{R}^m$$

is the feature map produced by all operations in the residual branch before the last output projection, U denotes the parameters of these preceding operations, and

$$V \in \mathbb{R}^{N\times m}$$

is the final linear projection back to the residual-stream dimension. Thus

$$h_{U,V} : \mathbb{R}^N \to \mathbb{R}^N,$$

so that $h_{U,V}(z)$ has the same dimension as z and can be added residually.

This form is not a special architecture. It is an abstraction of the standard "feature extraction plus output projection" structure used in residual networks. The map ψ_U collects all preceding transformations, such as normalization, activation functions, MLP layers, attention operations, or convolutional layers, while V denotes the last projection that maps the residual-branch features back to the dimension of the residual stream.

Example 1: MLP or convolutional residual branch. A two-layer MLP residual branch has the form

$$h(z) = W_2\sigma(W_1 z).$$

This is a special case of $h_{U,V}(z) = V\psi_U(z)$ by taking

$$\psi_U(z) = \sigma(W_1 z), \qquad V = W_2,$$

where $U = W_1$. Similarly, a convolutional residual branch can be written as

$$h(z) = \mathrm{Conv}_{\mathrm{out}}(\psi_U(z)),$$

where $\psi_U(z)$ denotes all preceding convolutions, normalizations, and activation functions, and $\mathrm{Conv}_{\mathrm{out}}$ is the final convolutional projection. Since a convolution is a linear map with parameter sharing, it plays the same role as V in the abstract expression

$$h_{U,V}(z) = V\psi_U(z).$$

Thus standard ResNet-style residual branches fall into this form once the final linear or convolutional output projection is separated from the preceding feature extractor.

Example 2: Transformer/LLM residual branch. The same structure appears in Transformer blocks used in large language models. For an MLP sublayer, the residual branch often has the schematic form

$$h(z) = W_{\mathrm{out}}\,\sigma(W_{\mathrm{in}}\,\mathrm{Norm}(z)),$$

or, in a gated MLP,

$$h(z) = W_{\mathrm{out}}\left(\sigma(W_{\mathrm{gate}}\mathrm{Norm}(z)) \odot W_{\mathrm{up}}\mathrm{Norm}(z)\right).$$

Both are of the form $h_{U,V}(z) = V\psi_U(z)$. In the first case,

$$\psi_U(z) = \sigma(W_{\mathrm{in}}\mathrm{Norm}(z)), \qquad V = W_{\mathrm{out}}.$$

In the gated case,

$$\psi_U(z) = \sigma(W_{\mathrm{gate}}\mathrm{Norm}(z)) \odot W_{\mathrm{up}}\mathrm{Norm}(z), \qquad V = W_{\mathrm{out}}.$$

The attention sublayer has the same output-projection structure:

$$h(z) = W_O\,\mathrm{Attn}_U(\mathrm{Norm}(z)),$$

where Attn_U contains the query, key, and value projections and the attention operation, while W_O is the final output projection. Hence it also fits the abstract form with

$$\psi_U(z) = \mathrm{Attn}_U(\mathrm{Norm}(z)), \qquad V = W_O.$$

We now return to the first-order non-degeneracy condition. Suppose the final output projection is initialized at zero:

$$V = 0,$$

while the preceding feature extractor is fixed at some initialization U_0. Then

$$h_{U_0,0}(z) = 0.$$

Therefore, inserting this residual block initially recovers the old model exactly.

However, this zero-output initialization does not make the first-order tangent space vanish. For a perturbation direction

$$\Delta V \in \mathbb{R}^{N\times m},$$

consider the parameter path

$$V(\eta) = \eta\Delta V, \qquad U(\eta) = U_0.$$

Then

$$h_{U_0,\eta\Delta V}(z) = \eta\Delta V\psi_{U_0}(z).$$

Thus the first-order hidden-space perturbation induced by changing the final projection is

$$\phi_{\Delta V}(z) = \Delta V\psi_{U_0}(z).$$

Condition 1' holds if there exists ΔV such that

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z,y)^\top \Delta V\psi_{U_0}(z)\right] \neq 0.$$

A sufficient condition is

$$C := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z,y)\psi_{U_0}(z)^\top\right] \neq 0.$$

Indeed, choosing

$$\Delta V = -C$$

gives

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z,y)^\top \Delta V\psi_{U_0}(z)\right] = -\left\|\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z,y)\psi_{U_0}(z)^\top\right]\right\|_F^2 < 0.$$

The empirical analogue is obtained by replacing the population expectation with the training average:

$$C_S := \frac{1}{M}\sum_{i=1}^{M} q_i\psi_{U_0}(z_i)^\top.$$

If $C_S \neq 0$, choosing

$$\Delta V = -C_S$$

gives

$$\frac{1}{M}\sum_{i=1}^{M} q_i^\top \Delta V\psi_{U_0}(z_i) = -\left\|\frac{1}{M}\sum_{i=1}^{M} q_i\psi_{U_0}(z_i)^\top\right\|_F^2 < 0.$$

Therefore, the same zero-output residual block construction also yields an empirical jumpboard direction v_S whenever the residual features are not training-set-orthogonal to the empirical activation-gradient signal. Therefore, a standard residual branch with a zero-initialized final projection satisfies Condition 1' whenever its residual features are not population-orthogonal to the activation-gradient signal.

This also explains why zeroing the entire residual branch is not the intended condition. If all layers in the branch were initialized to zero, the first-order Jacobian with respect to the parameters could vanish. In contrast, zero-initializing only the final output projection ensures $h_{U_0,0}(z) = 0$ while preserving nontrivial first-order directions through perturbations of V. This is precisely the zero-output residual parameterization needed for the jumpboard construction.

Corollary: standard zero-output residual blocks satisfy Condition 1. Based on the above analysis, we obtain the following corollary.

Corollary 1 (First-order non-degeneracy for standard zero-output residual blocks)**.** *Consider an inserted residual branch of the form*

$$h_{U,V}(z) = V\psi_U(z),$$

where $z = x^{(l^*)} \in \mathbb{R}^N$, $\psi_U(z) \in \mathbb{R}^m$, and $V \in \mathbb{R}^{N\times m}$ is the final output projection. Suppose the block is initialized by

$$U = U_0, \qquad V = 0.$$

Then $h_{U_0,0}(z) = 0$, so inserting the block initially recovers the old model exactly.

For the population-risk route, define

$$C_{\mathrm{pop}} := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z,y)\psi_{U_0}(z)^\top\right], \qquad q(z,y) := \nabla_z \ell(f_{\mathrm{top}}(z), y).$$

If

$$C_{\mathrm{pop}} \neq 0,$$

then the inserted residual block satisfies the population first-order non-degeneracy condition. In particular, choosing

$$\Delta V = -C_{\mathrm{pop}}$$

gives

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z,y)^\top \Delta V \psi_{U_0}(z)\right] = -\|C_{\mathrm{pop}}\|_F^2 < 0.$$

Hence there exists a population descent direction and therefore a population jumpboard model $\widetilde{f}_{\mathrm{pop}}$.

For the empirical train/test route, given $S_{\mathrm{train}} = \{(x_i, y_i)\}_{i=1}^M$, *define*

$$z_i := f_{\mathrm{bot}}(x_i), \qquad q_i := \nabla_{z_i} \ell(f_{\mathrm{top}}(z_i), y_i),$$

and

$$C_S := \frac{1}{M}\sum_{i=1}^M q_i \psi_{U_0}(z_i)^\top.$$

If

$$C_S \neq 0,$$

then the inserted residual block satisfies the empirical first-order non-degeneracy condition. In particular, choosing

$$\Delta V = -C_S$$

gives

$$\frac{1}{M}\sum_{i=1}^M q_i^\top \Delta V \psi_{U_0}(z_i) = -\|C_S\|_F^2 < 0.$$

Hence there exists an empirical descent direction and therefore a training-set jumpboard model $\widetilde{f}_S$.

Proof. Since $V = 0$, we have $h_{U_0,0}(z) = 0$, so the expanded network initially represents the old model. Now perturb only the final projection by setting

$$V(\eta) = \eta \Delta V, \qquad U(\eta) = U_0.$$

Then

$$h_{U_0,\eta\Delta V}(z) = \eta \Delta V \psi_{U_0}(z),$$

so the induced first-order hidden-space perturbation is

$$\phi_{\Delta V}(z) = \Delta V \psi_{U_0}(z).$$

For the population route, the first-order variation of the risk along this direction is

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z,y)^\top \Delta V \psi_{U_0}(z)\right].$$

Taking $\Delta V = -C_{\mathrm{pop}}$ gives

$$-\|C_{\mathrm{pop}}\|_F^2 < 0.$$

The empirical statement follows identically by replacing the population expectation with the training average and choosing $\Delta V = -C_S$. □

This corollary shows that Condition 1 is not an artificial functional assumption. It is satisfied by standard zero-output residual branches whenever the residual features are not orthogonal to the relevant activation-gradient signal, either in population or on the training set.

C.6 WHY THE DERIVATIVE IS TAKEN WITH RESPECT TO PARAMETERS

The Jacobian in Condition 1' is the parameter Jacobian

$$J_\theta h_0(z) = \left.\frac{\partial h_\theta(z)}{\partial \theta}\right|_{\theta=0},$$

not the input Jacobian

$$\frac{\partial h_\theta(z)}{\partial z}.$$

This is because the jumpboard construction fixes the old representation

$$z = x^{(l^*)}$$

and moves only the parameters of the newly inserted residual block. The path used to construct the jumpboard model is

$$\theta(\eta) = \eta u,$$

so the first-order change in the model comes from changing θ, not from changing the input z to the residual branch.

For example, in the zero-last-layer block

$$h_V(z) = V\psi_U(z), \qquad V = 0,$$

the input Jacobian satisfies

$$\left.\frac{\partial h_V(z)}{\partial z}\right|_{V=0} = 0,$$

because the output layer V is zero. However, the parameter Jacobian with respect to V is nonzero:

$$\left.\frac{\partial h_V(z)}{\partial V}\right|_{V=0}[\Delta V] = \Delta V \psi_U(z).$$

Thus using the input Jacobian would incorrectly rule out standard zero-last-layer residual blocks, while the parameter Jacobian captures exactly the first-order hidden-space perturbations created by training the inserted block.

C.7 WHY CONDITION 1 IS MILD.

We further explain why Condition 1 is a generic non-orthogonality requirement. Suppose the old model is trained to stationarity with respect to its own parameters. This only gives

$$\nabla_{\theta_{\text{old}}} R(f^*_{\text{old}}) = 0,$$

or, at the empirical level,

$$\nabla_{\theta_{\text{old}}} \mathcal{L}_{\text{train}}(f^*_{\text{old}}) = 0.$$

By the chain rule, this means that the activation-gradient signal is orthogonal to perturbations of the hidden representation that can already be generated by the old parameters. It does not imply that the activation-gradient signal itself is zero.

Now consider a standard zero-output residual branch of the form

$$h_{U,V}(z) = V\psi_U(z),$$

where the output projection is initialized at $V = 0$. Then $h_{U,0}(z) = 0$, so the old model is exactly embedded in the expanded class. However, perturbing V gives nontrivial first-order residual directions:

$$h_{U,\eta\Delta V}(z) = \eta \Delta V \psi_U(z).$$

The first-order population variation is therefore

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z,y)^\top \Delta V \psi_U(z)\right], \qquad q(z,y) := \nabla_z \ell(f_{\text{top}}(z), y).$$

Equivalently, Condition 1 holds whenever

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z,y)\psi_U(z)^\top\right] \neq 0.$$

If this matrix is nonzero, choosing the sign of ΔV appropriately gives a strict first-order descent direction. Thus the condition fails only in the special case where the activation-gradient signal is population-orthogonal to all residual features realizable by the inserted block. This is a degeneracy condition, not the typical case.

D POPULATION JUMPBOARD EXISTENCE: PROOF OF THEOREM 3

Theorem 3 (Value of Depth)**.** *Under Assumptions 1 and 2(a), there exists a jumpboard model* $\tilde{f} \in \mathcal{H}_{\text{new}}$ *such that*

$$R(\tilde{f}_{pop}) < R(f^*_{\text{old}}).$$

Proof. We divide the proof into three steps.

Step 1: Construct a descent path in parameter space. By Condition 1, the residual block class $\mathcal{F}_{\text{res}}$ admits a parameterization h_θ with zero initialization $h_0 = 0$, and there exists a parameter direction v_{pop} such that

$$a \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\nabla_{x^{(l^*)}}\ell\Big(f_{\text{top}}(x^{(l^*)}), y\Big)^\top \cdot \frac{\partial h_\theta(x^{(l^*)})}{\partial\theta}\Big|_{\theta=0} \cdot v_{pop}\right] \neq 0.$$

If $a > 0$, we replace v by $-v$. Hence, without loss of generality, we may assume

$$a < 0.$$

Now consider the one-dimensional parameter path

$$\theta(t) = tv, \qquad t \geq 0,$$

and define the corresponding population-risk profile

$$G(t) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\ell\Big(f_{\text{top}}\big(x^{(l^*)} + h_{tv}(x^{(l^*)})\big), y\Big)\Big].$$

Since $h_0 = 0$, at $t = 0$ the inserted residual block vanishes, so the model reduces to the old model. Therefore,

$$G(0) = R(f^*_{\text{old}}).$$

Step 2: Compute the directional derivative at the origin. By Assumption 1, the map

$$v \mapsto \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\Big(f_{\text{top}}(x^{(l^*)} + v), y\Big)\right]$$

is continuously differentiable in a neighborhood of the origin. Hence, by the chain rule, G is differentiable at $t = 0$, and

$$G'(0) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\nabla_{x^{(l^*)}}\ell\Big(f_{\text{top}}(x^{(l^*)}), y\Big)^\top \cdot \frac{\partial h_{tv_{pop}}(x^{(l^*)})}{\partial t}\Big|_{t=0}\right].$$

Since $\theta = tv_{pop}$, we have

$$\frac{\partial h_{tv}(x^{(l^*)})}{\partial t}\Big|_{t=0} = \frac{\partial h_\theta(x^{(l^*)})}{\partial\theta}\Big|_{\theta=0} \cdot v_{pop}.$$

Substituting this identity into the previous display yields

$$G'(0) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\nabla_{x^{(l^*)}}\ell\Big(f_{\text{top}}(x^{(l^*)}), y\Big)^\top \cdot \frac{\partial h_\theta(x^{(l^*)})}{\partial\theta}\Big|_{\theta=0} \cdot v_{pop}\right] = a < 0.$$

Thus, the population risk decreases to first order along the path $\theta(t) = tv$ at the origin.

Step 3: Use negativity of the derivative to obtain a strictly better model. Since G is differentiable at 0 and $G'(0) < 0$, there exists $t_0 > 0$ such that

$$G(t_0) < G(0).$$

Let

$$\theta_0 = t_0 v,$$

and define the corresponding jumpboard model

$$\tilde{f}(x) \triangleq f_{\mathrm{top}}\Big(f_{\mathrm{bot}}(x) + h_{\theta_0}(f_{\mathrm{bot}}(x))\Big).$$

Because $h_{\theta_0} \in \mathcal{F}_{\mathrm{res}}$, we have $\tilde{f} \in \mathcal{H}_{\mathrm{new}}$. Moreover,

$$R(\tilde{f}) = G(t_0) < G(0) = R(f^*_{\mathrm{old}}).$$

Therefore, there exists a model in the expanded hypothesis class $\mathcal{H}_{\mathrm{new}}$ whose population risk is strictly smaller than that of the old model f^*_{old}.

This completes the proof. □

Remark. The proof is purely population-level. It does not require any concentration on the training set or the test set. Its only purpose is to establish the *existence* of a strictly better model in the enlarged hypothesis class. In particular, the jump model $\tilde{f}$ is an auxiliary comparison model; the actual optimizer is not required to find this model exactly.

E Proof of Theorem 4

For completeness, we state Theorem 4.

We formalize the independent-reference sampling convention stated in the main text. Let $\mathcal{F}_0$ be the sigma-field generated by the complete reference network $f^*_{\mathrm{old}} = f_{\mathrm{top}} \circ f_{\mathrm{bot}}$, its decomposition, and the designated insertion layer l^*. We assume

$$\mathcal{L}(S_{\mathrm{train}}, S_{\mathrm{test}} \mid \mathcal{F}_0) = \mathcal{D}^M \otimes \mathcal{D}^K.$$

Thus, conditional on the reference information, the training and test samples are independent i.i.d. samples from $\mathcal{D}$. All gradient moments and probability statements in this appendix are interpreted conditionally on $\mathcal{F}_0$. This sampling convention is used for the alignment analysis and is not an additional assumption for the risk-comparison inequalities in Theorems 1 and 2.

Theorem 4 (Train/test alignment for empirical jumpboards)**.** *Fix the trained reference model $f^*_{\mathrm{old}} = f_{\mathrm{top}} \circ f_{\mathrm{bot}}$ and an insertion layer l^*. Under the conditional sampling setup, let*

$$\zeta := \nabla_{x^{(l^*)}} \ell\big(f_{\mathrm{top}}(x^{(l^*)}), y\big) \in \mathbb{R}^N$$

be the activation-gradient random vector under $(x, y) \sim \mathcal{D}$. Define

$$\bar{\mu} := \mathbb{E}[\zeta], \qquad \Sigma := \mathrm{Cov}(\zeta),$$

and assume $\bar{\mu} \neq 0$. Let

$$\mu := \frac{1}{M}\sum_{i=1}^{M} \zeta_i, \qquad g := \frac{1}{K}\sum_{j=1}^{K} \widetilde{\zeta}_j,$$

where $\zeta_1, \ldots, \zeta_M$ are the training activation gradients and $\widetilde{\zeta}_1, \ldots, \widetilde{\zeta}_K$ are the independent test activation gradients. Suppose Assumption 3 holds, namely

$$\lambda_{\max}(\Sigma) \leq C_\Sigma \tau^2.$$

Then

$$\mathbb{P}(\mu^\top g \leq 0) \leq \frac{4C_\Sigma \tau^2}{M\|\bar{\mu}\|^2} + \frac{4C_\Sigma \tau^2}{K\|\bar{\mu}\|^2} + \frac{4C_\Sigma \tau^2 \,\mathrm{tr}(\Sigma)}{KM\|\bar{\mu}\|^4}.$$

In particular, if

$$\mathrm{tr}(\Sigma) \leq N\tau^2 \qquad \textit{and} \qquad \|\bar{\mu}\|^2 = \Theta(N),$$

then

$$\mathbb{P}(\mu^\top g \leq 0) = O\left(\frac{1}{MN} + \frac{1}{KN}\right).$$

Thus the probability that the training-selected empirical direction remains positively aligned with the independent test direction improves with width, training size, and test size.

Proof. We condition throughout on $\mathcal{F}_0$ and suppress this conditioning in intermediate expressions for readability. Under the sampling convention above, the reference network and insertion layer are fixed, and the training and test gradients form two independent collections of i.i.d. copies of the conditional gradient vector ζ.

The proof separates the failure event into two sources. First, the empirical training direction μ may fail to align with the population direction $\bar{\mu}$. Second, even when the training direction is reliable, the independent test direction g may fluctuate enough to make $\mu^\top g \le 0$.

Define the good training-alignment event

$$A := \left\{ \mu^\top \bar{\mu} \ge \frac{1}{2}\|\bar{\mu}\|^2 \right\}.$$

Then

$$\mathbb{P}(\mu^\top g \le 0) \le \mathbb{P}(A^c) + \mathbb{P}(\mu^\top g \le 0,\ A).$$

We bound the two terms separately.

First, since

$$\mu^\top \bar{\mu} = \|\bar{\mu}\|^2 + (\mu - \bar{\mu})^\top \bar{\mu},$$

the event A^c implies

$$(\mu - \bar{\mu})^\top \bar{\mu} < -\frac{1}{2}\|\bar{\mu}\|^2.$$

Therefore,

$$\mathbb{P}(A^c) \le \mathbb{P}\left(\left| (\mu - \bar{\mu})^\top \bar{\mu} \right| \ge \frac{1}{2}\|\bar{\mu}\|^2 \right).$$

Because μ is the average of M i.i.d. copies of ζ, we have

$$\mathbb{E}[\mu] = \bar{\mu}, \qquad \mathrm{Cov}(\mu) = \frac{\Sigma}{M}.$$

Hence

$$\mathrm{Var}\big((\mu - \bar{\mu})^\top \bar{\mu}\big) = \bar{\mu}^\top \mathrm{Cov}(\mu)\bar{\mu} = \frac{1}{M}\bar{\mu}^\top \Sigma \bar{\mu}.$$

Using $\lambda_{\max}(\Sigma) \le C_\Sigma \tau^2$, we obtain

$$\bar{\mu}^\top \Sigma \bar{\mu} \le \lambda_{\max}(\Sigma)\|\bar{\mu}\|^2 \le C_\Sigma \tau^2 \|\bar{\mu}\|^2.$$

Thus

$$\mathrm{Var}\big((\mu - \bar{\mu})^\top \bar{\mu}\big) \le \frac{C_\Sigma \tau^2}{M}\|\bar{\mu}\|^2.$$

By Chebyshev's inequality,

$$\mathbb{P}(A^c) \le \frac{\frac{C_\Sigma \tau^2}{M}\|\bar{\mu}\|^2}{\frac{1}{4}\|\bar{\mu}\|^4} = \frac{4C_\Sigma \tau^2}{M\|\bar{\mu}\|^2}.$$

This is the training-sample failure term.

We now bound the second term,

$$\mathbb{P}(\mu^\top g \le 0,\ A).$$

Condition on the training sample S_{train}. Then μ is fixed, while g remains random and independent of the training sample. Since

$$g = \frac{1}{K}\sum_{j=1}^{K} \widetilde{\zeta}_j,$$

we have

$$\mathbb{E}[g \mid S_{\mathrm{train}}] = \bar{\mu}, \qquad \mathrm{Cov}(g \mid S_{\mathrm{train}}) = \frac{\Sigma}{K}.$$

Therefore,

$$\mathbb{E}[\mu^\top g \mid S_{\mathrm{train}}] = \mu^\top \bar{\mu}$$

and

$$\mathrm{Var}(\mu^\top g \mid S_{\mathrm{train}}) = \mu^\top \mathrm{Cov}(g \mid S_{\mathrm{train}})\mu = \frac{1}{K}\mu^\top \Sigma \mu.$$

Again using $\lambda_{\max}(\Sigma) \le C_\Sigma \tau^2$, we get

$$\mathrm{Var}(\mu^\top g \mid S_{\mathrm{train}}) \le \frac{C_\Sigma \tau^2}{K} \|\mu\|^2.$$

On the event A,

$$\mu^\top \bar{\mu} \ge \frac{1}{2}\|\bar{\mu}\|^2.$$

If $\mu^\top g \le 0$, then

$$\mu^\top g - \mu^\top \bar{\mu} \le -\mu^\top \bar{\mu},$$

and therefore

$$\left|\mu^\top g - \mu^\top \bar{\mu}\right| \ge \mu^\top \bar{\mu}.$$

By Chebyshev's inequality, conditional on S_{train},

$$\mathbb{P}(\mu^\top g \le 0 \mid S_{\mathrm{train}}) \le \frac{\mathrm{Var}(\mu^\top g \mid S_{\mathrm{train}})}{(\mu^\top \bar{\mu})^2}.$$

On A, this gives

$$\mathbb{P}(\mu^\top g \le 0 \mid S_{\mathrm{train}}) \le \frac{\frac{C_\Sigma \tau^2}{K}\|\mu\|^2}{\frac{1}{4}\|\bar{\mu}\|^4} = \frac{4C_\Sigma \tau^2 \|\mu\|^2}{K\|\bar{\mu}\|^4}.$$

Taking expectation over the training sample and restricting to A, we obtain

$$\mathbb{P}(\mu^\top g \le 0,\ A) \le \mathbb{E}\left[\mathbf{1}_A \frac{4C_\Sigma \tau^2 \|\mu\|^2}{K\|\bar{\mu}\|^4}\right].$$

Since $\mathbf{1}_A \le 1$,

$$\mathbb{P}(\mu^\top g \le 0,\ A) \le \frac{4C_\Sigma \tau^2}{K\|\bar{\mu}\|^4}\mathbb{E}\|\mu\|^2.$$

It remains to compute $\mathbb{E}\|\mu\|^2$. Since

$$\mathbb{E}[\mu] = \bar{\mu}, \qquad \mathrm{Cov}(\mu) = \frac{\Sigma}{M},$$

we have

$$\mathbb{E}\|\mu\|^2 = \|\mathbb{E}\mu\|^2 + \mathrm{tr}(\mathrm{Cov}(\mu)) = \|\bar{\mu}\|^2 + \frac{\mathrm{tr}(\Sigma)}{M}.$$

Therefore,

$$\mathbb{P}(\mu^\top g \le 0,\ A) \le \frac{4C_\Sigma \tau^2}{K\|\bar{\mu}\|^4}\left(\|\bar{\mu}\|^2 + \frac{\mathrm{tr}(\Sigma)}{M}\right).$$

Equivalently,

$$\mathbb{P}(\mu^\top g \le 0,\ A) \le \frac{4C_\Sigma \tau^2}{K\|\bar{\mu}\|^2} + \frac{4C_\Sigma \tau^2\,\mathrm{tr}(\Sigma)}{KM\|\bar{\mu}\|^4}.$$

Combining this with the bound on $\mathbb{P}(A^c)$, we obtain

$$\mathbb{P}(\mu^\top g \le 0) \le \frac{4C_\Sigma \tau^2}{M\|\bar{\mu}\|^2} + \frac{4C_\Sigma \tau^2}{K\|\bar{\mu}\|^2} + \frac{4C_\Sigma \tau^2\,\mathrm{tr}(\Sigma)}{KM\|\bar{\mu}\|^4}.$$

This proves the first claim.

For the width-dependent form, assume additionally that

$$\mathrm{tr}(\Sigma) \le N\tau^2$$

and

$$\|\bar{\mu}\|^2 = \Theta(N).$$

Then

$$\frac{4C_\Sigma \tau^2}{M\|\bar{\mu}\|^2} = O\left(\frac{1}{MN}\right),$$

and

$$\frac{4C_\Sigma\tau^2}{K\|\bar{\mu}\|^2} = O\left(\frac{1}{KN}\right).$$

For the third term,

$$\frac{4C_\Sigma\tau^2\,\mathrm{tr}(\Sigma)}{KM\|\bar{\mu}\|^4} \le \frac{4C_\Sigma\tau^2\cdot N\tau^2}{KM\|\bar{\mu}\|^4} = O\left(\frac{1}{KMN}\right),$$

because $\|\bar{\mu}\|^4 = \Theta(N^2)$. This term is lower order relative to the first two terms. Hence

$$\mathbb{P}(\mu^\top g \le 0) = O\left(\frac{1}{MN} + \frac{1}{KN}\right).$$

This completes the proof. □

F Proof of Lemma 1

Remark (pre-norm and post-norm placements). Throughout, "inserting a residual block after layer l^*" means composing the purely additive map $z \mapsto z + h_\theta(z)$ with the old network. The inserted block carries no normalization of its own (if its branch normalizes its input, as in $h_\theta(z) = V\psi_U(\mathcal{N}(z))$, that normalization is part of h_θ), and any normalization following the insertion point belongs to f_{top}. With this convention, $h_0 = 0$ recovers f^*_{old} exactly, both for post-norm layers $T_l(z) = \mathcal{N}_l\big(z + h_l(z)\big)$ and for pre-norm layers $T_l(z) = z + h_l\big(\mathcal{N}_l(z)\big)$. Assumption is stated for the post-norm form. For the pre-norm form, condition (iii) is replaced by the non-degeneracy of the residual stream, $\inf_{x,\theta,l}\|z_l(x)\| \ge r_{\min}$, and the state bound B_* is needed only for the normalized branch inputs, $\|\mathcal{N}_l(z)\| \le B_{\mathrm{norm}}$, provided that f_{top} begins with a normalization layer. All results of Sections 4 and 5 are insensitive to the placement, namely

$$A_l = c_l(1+s_l),\quad P_l = c_l L_l^{(p)} B_* \text{ (post-norm)} \qquad \text{versus} \qquad A_l = 1+c_l s_l,\quad P_l = L_l^{(p)} B_{\mathrm{norm}} \text{ (pre-norm)},$$

with $\Lambda_l = L_{\mathrm{top}} P_l \prod_{k>l} A_k$ in both cases;

Lemma 1 (Norm-based Rademacher bound for normalized residual networks)**.** *Consider a fixed network architecture*

$$z_0 = x, \qquad z_{l+1} = T_l(z_l;\theta_l),\quad l = 0,\dots,L-1, \qquad f_\Theta(x) = f_{\mathrm{top}}(z_L;U),$$

where $\Theta = (\theta_0,\dots,\theta_{L-1},U)$ belongs to a nonempty, fixed admissible parameter set $\mathcal{K}$. Let

$$H = \{f_\Theta : \Theta \in \mathcal{K}\}.$$

Each θ_l collects all parameters of the complete layer map that are allowed to vary in H, including residual-branch weights and biases and any trainable normalization parameters. Likewise, U collects all variable parameters of the top map. Fixed parameters are not counted among the variable parameter blocks. Parameter tensors are identified with their vectorizations; for a collection of tensors, $\|\cdot\|_F$ denotes the Euclidean norm of their concatenated vectorizations.

For each $l = 0,\dots,L$, define the class-wide reachable set

$$\mathcal{Z}_l = \{z_l(x;\Theta) : x \in \mathcal{X},\ \Theta \in \mathcal{K}\}.$$

Assume the following uniform regularity and boundedness conditions.

(i) Layerwise regularity. *For every $l = 0,\dots,L-1$, there exist finite nonnegative constants A_l, P_l, and $B_l^{(z)}$ such that*

$$\|T_l(z;\theta_l) - T_l(z';\theta_l)\| \le A_l\|z - z'\|, \tag{1}$$

$$\|T_l(z;\theta_l) - T_l(z;\theta_l')\| \le P_l\|\theta_l - \theta_l'\|_F, \tag{2}$$

$$\|T_l(z;\theta_l)\| \le B_l^{(z)}. \tag{3}$$

These inequalities hold for every $z, z' \in \mathcal{Z}_l$ and every θ_l, θ_l' in the l-th coordinate projection of $\mathcal{K}$. In particular, they hold for the cross-pairings of states and parameters used when comparing two admissible networks. The quantity $B_l^{(z)}$ is a uniform constant, not a function of z.

(ii) Top-map regularity. *There exist finite nonnegative constants L_{top} and P_{top} such that*

$$\|f_{\mathrm{top}}(z;U) - f_{\mathrm{top}}(z';U)\| \le L_{\mathrm{top}}\|z - z'\|, \tag{4}$$

$$\|f_{\mathrm{top}}(z;U) - f_{\mathrm{top}}(z;U')\| \le P_{\mathrm{top}}\|U - U'\|_F, \tag{5}$$

for every $z, z' \in \mathcal{Z}_L$ and every admissible U, U'. When the top map is fixed throughout H, the second inequality and its parameter block are omitted.

(iii) Parameter bounds. *Every variable layer block satisfies*

$$\theta_l \in \mathbb{R}^{d_l}, \qquad \|\theta_l\|_F \le b_l < \infty.$$

If the top map is variable, assume also

$$U \in \mathbb{R}^{d_{\mathrm{top}}}, \qquad \|U\|_F \le b_{\mathrm{top}} < \infty.$$

The set $\mathcal{K}$ may have additional joint constraints; it need not be a Cartesian product of blockwise constraint sets.

(iv) Loss regularity. *The loss is uniformly L_ℓ-Lipschitz in its prediction argument and takes values in $[0, B_\ell]$ on the outputs generated by H, for finite $L_\ell, B_\ell \ge 0$.*

For every variable layer block, define

$$\Lambda_l = L_{\mathrm{top}} P_l \prod_{k=l+1}^{L-1} A_k,$$

where an empty product equals 1*. For a variable top map, define*

$$\Lambda_{\mathrm{top}} = P_{\mathrm{top}}.$$

Relabel all variable parameter blocks as ϑ_j, $j = 1, \dots, J$, with corresponding dimensions, norm radii, and sensitivities (d_j, b_j, Λ_j). Suppose $J \ge 1$.

For any sample $S = \{(x_i, y_i)\}_{i=1}^m$ with $m \ge 1$, define

$$\widehat{\mathfrak{R}}_S(\ell \circ H) = \mathbb{E}_\sigma\left[\sup_{\Theta \in \mathcal{K}} \frac{1}{m}\sum_{i=1}^m \sigma_i \ell(f_\Theta(x_i), y_i)\right],$$

where the σ_i are independent Rademacher signs. Then, for every $\varepsilon_0 > 0$,

$$\boxed{\widehat{\mathfrak{R}}_S(\ell \circ H) \le B_\ell \sqrt{\frac{2}{m}\sum_{j=1}^J d_j \log\left(1 + \frac{2b_j J L_\ell \Lambda_j}{\varepsilon_0}\right)} + \varepsilon_0.} \tag{6}$$

In particular, with

$$d_{\mathrm{tot}} = \sum_{j=1}^J d_j, \qquad \bar{b} = \max_{1 \le j \le J} b_j, \qquad \Lambda_{\max} = \max_{1 \le j \le J} \Lambda_j,$$

one has

$$\boxed{\widehat{\mathfrak{R}}_S(\ell \circ H) \le B_\ell \sqrt{\frac{2d_{\mathrm{tot}}}{m} \log\left(1 + \frac{2\bar{b} J L_\ell \Lambda_{\max}}{\varepsilon_0}\right)} + \varepsilon_0.} \tag{7}$$

If $J = 0$, the class consists of one fixed predictor and its empirical Rademacher complexity, under the definition above, equals zero.

Proof. The argument first bounds the sensitivity of the entire network to all variable parameter blocks, then constructs a finite loss-class cover with admissible centers, and finally controls its Rademacher complexity.

Step 1: Sensitivity of the hidden states. Fix an input x and two admissible parameter vectors Θ, Θ'. Denote their hidden states by z_l, z'_l and put $e_l = \|z_l - z'_l\|$. Since $z_0 = z'_0 = x$, $e_0 = 0$. By equation 1–equation 2,

$$\begin{aligned} e_{l+1} &= \|T_l(z_l;\theta_l) - T_l(z'_l;\theta'_l)\| \\ &\le \|T_l(z_l;\theta_l) - T_l(z'_l;\theta_l)\| + \|T_l(z'_l;\theta_l) - T_l(z'_l;\theta'_l)\| \\ &\le A_l e_l + P_l\|\theta_l - \theta'_l\|_F. \end{aligned}$$

Iterating this recursion gives

$$\|z_L - z'_L\| \le \sum_{l=0}^{L-1} P_l \left(\prod_{k=l+1}^{L-1} A_k\right) \|\theta_l - \theta'_l\|_F. \tag{8}$$

A fixed layer parameter block contributes zero to this sum.

Step 2: Sensitivity of the output and loss. Using equation 4–equation 5 and then equation 8,

$$\begin{aligned} \|f_\Theta(x) - f_{\Theta'}(x)\| &\le L_{\mathrm{top}}\|z_L - z'_L\| + P_{\mathrm{top}}\|U - U'\|_F \\ &\le \sum_{l=0}^{L-1} \Lambda_l\|\theta_l - \theta'_l\|_F + \Lambda_{\mathrm{top}}\|U - U'\|_F. \end{aligned}$$

The top-parameter term is omitted when the top map is fixed. Relabeling the variable blocks and taking the supremum over x yields

$$\sup_{x\in\mathcal{X}} \|f_\Theta(x) - f_{\Theta'}(x)\| \le \sum_{j=1}^{J} \Lambda_j\|\vartheta_j - \vartheta'_j\|_F. \tag{9}$$

Let $F_\Theta(x,y) = \ell(f_\Theta(x), y)$. Loss Lipschitzness implies

$$\|F_\Theta - F_{\Theta'}\|_\infty \le \sum_{j=1}^{J} a_j\|\vartheta_j - \vartheta'_j\|_F, \qquad a_j := L_\ell\Lambda_j, \tag{10}$$

where the supremum norm is over $\mathcal{X}\times\mathcal{Y}$.

Step 3: A finite loss-class cover with admissible centers. Let $I = \{j : a_j > 0\}$. If I is empty, equation 10 shows that every element of $\ell \circ H$ is the same function. Its empirical Rademacher complexity is zero, so both bounds follow. Assume henceforth that $I \neq \varnothing$.

For $j \in I$, set

$$r_j = \frac{\varepsilon_0}{Ja_j},$$

and equip the admissible parameter set with the pseudometric

$$d_r(\Theta,\Theta') = \max_{j\in I} \frac{\|\vartheta_j - \vartheta'_j\|_F}{r_j}.$$

Take a maximal set $\Theta^{(1)}, \dots, \Theta^{(Q)} \in \mathcal{K}$ whose distinct elements have d_r-distance greater than 1. To bound its cardinality, project to the active coordinates I. The open product balls

$$\prod_{j\in I} B^{d_j}\left(\vartheta_j^{(q)}, r_j/2\right), \qquad q = 1,\dots,Q,$$

are pairwise disjoint and are all contained in

$$\prod_{j\in I} B^{d_j}(0, b_j + r_j/2).$$

Comparing their Euclidean volumes gives

$$Q \le \prod_{j\in I}\left(1 + \frac{2b_j}{r_j}\right)^{d_j}.$$

This finite bound also ensures that a maximal separated set can be obtained by a finite greedy construction.

By maximality, every $\Theta \in \mathcal{K}$ has an index q with $d_r(\Theta, \Theta^{(q)}) \le 1$. Therefore, by equation 10,

$$\|F_\Theta - F_{\Theta^{(q)}}\|_\infty \le \sum_{j \in I} a_j r_j = \frac{|I|}{J}\varepsilon_0 \le \varepsilon_0.$$

Thus

$$\mathcal{F}_{\varepsilon_0} = \{F_{\Theta^{(1)}}, \dots, F_{\Theta^{(Q)}}\}$$

is an ε_0-cover of $\ell \circ H$ in the uniform norm, with every center belonging to $\ell \circ H$. Its size satisfies

$$\log Q \le \sum_{j \in I} d_j \log\left(1 + \frac{2b_j J a_j}{\varepsilon_0}\right) = \sum_{j=1}^{J} d_j \log\left(1 + \frac{2b_j J L_\ell \Lambda_j}{\varepsilon_0}\right). \tag{11}$$

The terms with $a_j = 0$ vanish. This construction does not require the admissible parameter set to factor across blocks.

Step 4: Finite-class Rademacher bound. For each cover element define

$$v_i^{(q)} = F_{\Theta^{(q)}}(x_i, y_i), \qquad Z_q = \frac{1}{m}\sum_{i=1}^{m} \sigma_i v_i^{(q)}.$$

Since $0 \le v_i^{(q)} \le B_\ell$, for any $t > 0$,

$$\begin{aligned}\mathbb{E}_\sigma e^{tZ_q} &= \prod_{i=1}^{m} \cosh\left(\frac{t v_i^{(q)}}{m}\right) \\ &\le \exp\left(\frac{t^2}{2m^2}\sum_{i=1}^{m}(v_i^{(q)})^2\right) \le \exp\left(\frac{t^2 B_\ell^2}{2m}\right).\end{aligned}$$

Consequently,

$$\begin{aligned}\mathbb{E}_\sigma \max_{1 \le q \le Q} Z_q &\le \frac{1}{t}\log\left(\sum_{q=1}^{Q} \mathbb{E}_\sigma e^{tZ_q}\right) \\ &\le \frac{\log Q}{t} + \frac{t B_\ell^2}{2m}.\end{aligned}$$

Optimizing over t gives the finite-class bound

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_{\varepsilon_0}) \le B_\ell \sqrt{\frac{2\log Q}{m}}. \tag{12}$$

If $Q = 1$ or $B_\ell = 0$, the same inequality follows directly.

Step 5: Transfer from the cover to the full loss class. For every $F_\Theta \in \ell \circ H$, choose a cover element $F_{\Theta^{(q)}}$ within uniform distance ε_0. For each realization of the Rademacher signs,

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\sigma_i F_\Theta(x_i, y_i) &\le \frac{1}{m}\sum_{i=1}^{m}\sigma_i F_{\Theta^{(q)}}(x_i, y_i) + \frac{1}{m}\sum_{i=1}^{m}|F_\Theta(x_i, y_i) - F_{\Theta^{(q)}}(x_i, y_i)| \\ &\le \max_{1 \le q \le Q}\frac{1}{m}\sum_{i=1}^{m}\sigma_i F_{\Theta^{(q)}}(x_i, y_i) + \varepsilon_0.\end{aligned}$$

Taking the supremum over Θ and then expectation gives

$$\widehat{\mathfrak{R}}_S(\ell \circ H) \le \widehat{\mathfrak{R}}_S(\mathcal{F}_{\varepsilon_0}) + \varepsilon_0.$$

Combining this with equation 12 and equation 11 proves equation 6.

Step 6: Simplification. Since $b_j \le \bar{b}$ and $\Lambda_j \le \Lambda_{\max}$,

$$\sum_{j=1}^{J} d_j \log\left(1 + \frac{2b_j J L_\ell \Lambda_j}{\varepsilon_0}\right) \le d_{\text{tot}} \log\left(1 + \frac{2\bar{b} J L_\ell \Lambda_{\max}}{\varepsilon_0}\right).$$

Substitution into equation 6 yields equation 7. □

Thus, we have the following corollary.

Corollary 2 (Uniform generalization bound)**.** *Under the conditions of Lemma 1, with probability at least $1-\delta$ over a sample S of size m, one has*

$$\sup_{f\in\mathcal{H}} |R(f) - L_S(f)| \le \epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H}, m, \delta),$$

where one may take

$$\epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H}, m, \delta) := 2B_\ell \sqrt{\frac{2d\log\big(2\bar{b}LL_\ell\Lambda_{\max}/\rho + 1\big)}{m}} + 2\rho + 3B_\ell\sqrt{\frac{\log(2/\delta)}{2m}}.$$

Taking $\rho = m^{-1/2}$ gives the dominant scale

$$\epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H}, m, \delta) = \widetilde{O}\left(B_\ell\sqrt{\frac{d}{m}}\right),$$

up to the explicit norm-, depth-, and normalization-dependent factors contained in $\Lambda_{\max}$.

Applying Corollary 2 with $m = M$ gives

$$\epsilon_M = \epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H}_{\text{new}}, M, \delta),$$

and applying it with $m = K$ gives

$$\epsilon_K = \epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H}_{\text{new}}, K, \delta).$$

Proof. Let

$$\mathcal{F} := \ell \circ \mathcal{H} = \{(x, y) \mapsto \ell(f(x), y) : f \in \mathcal{H}\}.$$

For any $f \in \mathcal{H}$, define

$$P(\ell \circ f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)] = R(f),$$

and

$$P_S(\ell \circ f) := \frac{1}{m}\sum_{i=1}^{m} \ell(f(x_i), y_i) = L_S(f).$$

Therefore,

$$\sup_{f\in\mathcal{H}} |R(f) - L_S(f)| = \sup_{g\in\mathcal{F}} |Pg - P_S g|.$$

By the standard empirical Rademacher generalization bound for bounded loss classes, since every $g \in \mathcal{F}$ takes values in $[0, B_\ell]$, with probability at least $1-\delta$,

$$\sup_{g\in\mathcal{F}} |Pg - P_S g| \le 2\widehat{\mathfrak{R}}_S(\mathcal{F}) + 3B_\ell\sqrt{\frac{\log(2/\delta)}{2m}}.$$

By Lemma 1,

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) = \widehat{\mathfrak{R}}_S(\ell \circ \mathcal{H}) \le B_\ell \sqrt{\frac{2d\log\big(2\bar{b}LL_\ell\Lambda_{\max}/\rho + 1\big)}{m}} + \rho.$$

Substituting this bound into the previous display gives

$$\sup_{f\in\mathcal{H}}|R(f)-L_S(f)|\le 2B_\ell\sqrt{\frac{2d\log\big(2\bar{b}LL_\ell\Lambda_{\max}/\rho+1\big)}{m}}+2\rho+3B_\ell\sqrt{\frac{\log(2/\delta)}{2m}}.$$

Thus we may define

$$\epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H},m,\delta):=2B_\ell\sqrt{\frac{2d\log\big(2\bar{b}LL_\ell\Lambda_{\max}/\rho+1\big)}{m}}+2\rho+3B_\ell\sqrt{\frac{\log(2/\delta)}{2m}}.$$

Finally, taking $\rho=m^{-1/2}$, we obtain

$$\epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H},m,\delta)=2B_\ell\sqrt{\frac{2d\log\big(2\bar{b}LL_\ell\Lambda_{\max}\sqrt{m}+1\big)}{m}}+\frac{2}{\sqrt{m}}+3B_\ell\sqrt{\frac{\log(2/\delta)}{2m}}.$$

Hence the dominant term is

$$\widetilde{O}\left(B_\ell\sqrt{\frac{d}{m}}\right),$$

up to the explicit norm-, depth-, and normalization-dependent factors contained in $\Lambda_{\max}$. This proves the corollary. □

These are the two quantities used in the main theorems.

Because the bound is uniform over all $f\in\mathcal{H}_{\text{new}}$, it also applies to data-dependent choices such as the empirical jumpboard model $\widetilde{f}_S$ and the final selected model f_{new}.

Lemma 1 does not model the training dynamics. It only controls the generalization gap uniformly over a norm-bounded hypothesis class. Thus it applies to the final selected model once that model is viewed as an element of $\mathcal{H}_{\text{new}}$. For BatchNorm, the relevant issue is not the training algorithm but whether the normalization layer defines a fixed single-sample map. Fixed-statistics BatchNorm can be absorbed into the constants c_l, while training-time BatchNorm with mini-batch-dependent statistics requires additional assumptions.

Constants under the two placements. Let $\mathcal{N}_l$ be c_l-Lipschitz on the reachable set and let $h_l(\cdot;W_l)$ satisfy Assumption (i)–(ii). For the post-norm layer $T_l(z;W_l)=\mathcal{N}_l\big(z+h_l(z;W_l)\big)$,

$$\|T_l(u;W_l)-T_l(v;W_l)\|\le c_l\big(1+s_l\big)\|u-v\|,\qquad \|T_l(z;W_l)-T_l(z;W_l')\|\le c_lL_l^{(p)}\|z\|\,\|W_l-W_l'\|_F,$$

so $A_l=c_l(1+s_l)$ and $P_l=c_lL_l^{(p)}B_*$ with $B_*=\max\{B_x,B_{\text{norm}}\}$. For the pre-norm layer $T_l(z;W_l)=z+h_l\big(\mathcal{N}_l(z);W_l\big)$,

$$\|T_l(u;W_l)-T_l(v;W_l)\|\le(1+c_ls_l)\|u-v\|,\qquad \|T_l(z;W_l)-T_l(z;W_l')\|\le L_l^{(p)}\|\mathcal{N}_l(z)\|\,\|W_l-W_l'\|_F\le L_l^{(p)}B_{\text{norm}}\|W_l-W_l'\|_F$$

so $A_l=1+c_ls_l$ and $P_l=L_l^{(p)}B_{\text{norm}}$. In both cases $\Lambda_l=L_{\text{top}}P_l\prod_{k>l}A_k$. The inserted block $z\mapsto z+h_\theta(\mathcal{N}(z))$ has $A_{\text{ins}}=1+c\,s_{\text{ins}}$ and $P_{\text{ins}}=L_{\text{ins}}^{(p)}B_{\text{norm}}$, and equals the identity at $\theta=0$. For RMSNorm, $B_{\text{norm}}=\gamma_{\max}\sqrt{N}$ and $c_l\le\gamma_{\max}/\sqrt{r_{\min}^2/N+\varepsilon_{\text{eng}}}$, which covers both the $r_{\min}$.

F.1 Data Requirement and Architecture-Dependent Factors

We now spell out the data requirement implied by Lemma 1. This discussion is used only to interpret the role of data in the main text; it is not an additional assumption.

From Corollary 2, after taking $\rho=m^{-1/2}$, the normalized-network generalization term has the form

$$\epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H},m,\delta)=2B_\ell\sqrt{\frac{2d\log\big(2\bar{b}LL_\ell\Lambda_{\max}\sqrt{m}+1\big)}{m}}+\frac{2}{\sqrt{m}}+3B_\ell\sqrt{\frac{\log(2/\delta)}{2m}}.$$

For compactness, define the architecture factor

$$\Gamma_{\mathrm{arch}}(L,N,m) := \log\bigl(2\bar{b}LL_\ell\Lambda_{\max}\sqrt{m}+1\bigr).$$

Then the leading term is

$$\epsilon_{\mathrm{gen}}^{\mathrm{norm}}(\mathcal{H},m,\delta) = O\left(B_\ell\sqrt{\frac{d\,\Gamma_{\mathrm{arch}}(L,N,m)}{m}}\right) + O\left(B_\ell\sqrt{\frac{\log(1/\delta)}{m}}\right).$$

Thus, to achieve a target generalization accuracy $\varepsilon > 0$, it is sufficient, up to universal constants, to require

$$m \gtrsim \frac{B_\ell^2 d\,\Gamma_{\mathrm{arch}}(L,N,m) + B_\ell^2\log(1/\delta)}{\varepsilon^2}.$$

This relation should be interpreted carefully. If the architecture-dependent quantities are fixed or uniformly controlled, then $\Gamma_{\mathrm{arch}}(L,N,m) = O(\log m)$, and the familiar shorthand

$$\epsilon_{\mathrm{gen}}^{\mathrm{norm}} = \widetilde{O}\left(B_\ell\sqrt{\frac{d}{m}}\right)$$

is appropriate. In this restricted sense, the required sample size is nearly linear in the effective parameter dimension d, up to logarithmic factors.

However, when depth L, width N, or the normalization constants also scale, the factor Γ_{arch} is not merely a logarithmic factor in the sample size. It also contains the architecture-dependent amplification term $\Lambda_{\max}$. In particular,

$$\Lambda_{\max} = \max_{0\le l<L} L_{\mathrm{top}}L_l^{(p)}B_* c_l\prod_{k>l}c_k(1+s_k),$$

so the generalization bound carries a genuine dependence on depth and normalization. For example, if $\Lambda_{\max}$ grows exponentially in L, then $\log\Lambda_{\max}$ contributes linearly in L to Γ_{arch}. In that case, the required data size scales like

$$m \gtrsim \frac{d\bigl(\log m + L + \text{lower-order architecture terms}\bigr)}{\varepsilon^2},$$

rather than merely $d\log m/\varepsilon^2$.

Therefore, the main-text statement $\widetilde{O}(\sqrt{d/m})$ should be read only as a shorthand for the controlled-architecture regime. In the joint scaling regime studied in this paper, the more precise interpretation is that data must compensate for both the effective dimension d and the architecture-dependent amplification factor Γ_{arch}. Equivalently, one may define an effective statistical dimension

$$d_{\mathrm{stat}}(L,N,m) := d\,\Gamma_{\mathrm{arch}}(L,N,m),$$

so that the leading generalization term behaves as

$$O\left(B_\ell\sqrt{\frac{d_{\mathrm{stat}}(L,N,m)}{m}}\right).$$

This is the sense in which data controls the statistical cost of depth and width expansion.

G Proof of Theorem 5: Population-Risk Route

The *population jumpboard model*, denoted by $\tilde{f}_{pop}$, is constructed from a population first-order descent direction and is used in the population-risk route. Its population margin is denoted by $\Delta_R := R(f^*_{\mathrm{old}}) - R(\tilde{f}_{pop}) > 0$.

Theorem 5 (Test-set qualitative scaling mechanism via population risk)**.** *Let $\tilde{f}_{pop} \in \mathcal{H}_{\mathrm{new}}$ be the jumpboard model selected along a local descent path, and suppose $\mathcal{L}_{\mathrm{train}}(f_{\mathrm{new}}) \le \mathcal{L}_{\mathrm{train}}(\tilde{f}_{pop}) - \Delta_{\mathrm{ERM}}, \Delta_{\mathrm{ERM}} \ge 0$.Then there exist deterministic quantities $\eta_0 > 0$ and $\Delta_R > 0$, depending only on $\mathcal{D}$ and the current old model, such that with probability at least $1 - 2\delta - 6\exp\bigl(-\frac{K\Delta_R^2}{8B_\ell^2}\bigr)$, one has*

$$\mathcal{L}_{\text{test}}(f_{\text{new}}) \le \mathcal{L}_{\text{test}}(f^*_{\text{old}}) - \frac{\Delta_R}{2} - \Delta_{\text{ERM}} + 2\epsilon_M.$$

Hence, if $\Delta_R/2 + \Delta_{\text{ERM}} > 2\epsilon_M$*, then* $\mathcal{L}_{\text{test}}(f_{\text{new}}) < \mathcal{L}_{\text{test}}(f^*_{\text{old}})$.

Proof. We divide the proof into five steps.

Step 1: Fix a jump path. By Condition 1, there exists a parameter direction v_{pop} such that

$$a \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\nabla_{x^{(l^*)}}\ell\Big(f_{\text{top}}(x^{(l^*)}), y\Big)^\top \cdot \frac{\partial h_\theta(x^{(l^*)})}{\partial\theta}\Big|_{\theta=0} \cdot v_{pop}\right] \neq 0.$$

If $a > 0$, replace v by $-v$. Hence we may assume

$$a < 0.$$

Define the one-dimensional path

$$\theta(\eta) = \eta v, \qquad \eta \ge 0,$$

and the corresponding jumpboard model

$$\tilde{f}_{pop}(x) \triangleq f_{\text{top}}(f_{\text{bot}}(x) + h_{\eta v}(f_{\text{bot}}(x)))\,.$$

Step 2: Construct a deterministic positive margin at the population level. Let

$$G(\eta) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\ell\Big(f_{\text{top}}(x^{(l^*)} + h_{\eta v}(x^{(l^*)})), y\Big)\Big].$$

By Assumption 1 and the chain rule,

$$G'(0) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\nabla_{x^{(l^*)}}\ell\Big(f_{\text{top}}(x^{(l^*)}), y\Big)^\top \cdot \frac{\partial h_\theta(x^{(l^*)})}{\partial\theta}\Big|_{\theta=0} \cdot v_{pop}\right] = a < 0.$$

Hence there exists $\eta_0 > 0$ such that

$$R(\tilde{f}_{pop}) < R(f^*_{\text{old}}).$$

Define

$$\Delta_R \triangleq R(f^*_{\text{old}}) - R(\tilde{f}_{pop}) > 0.$$

Thus

$$R(\tilde{f}_{pop}) = R(f^*_{\text{old}}) - \Delta_R.$$

Step 3: Transfer the population margin to the test set by Hoeffding. Both f^*_{old} and $\tilde{f}_{\eta_0}$ are fixed functions. Since $\ell \in [0, B_\ell]$, Hoeffding's inequality gives, for any fixed f,

$$\mathbb{P}\big(\big|\mathcal{L}_{\text{test}}(f) - R(f)\big| > t\big) \le 2\exp\left(-\frac{2Kt^2}{B_\ell^2}\right).$$

Apply this to f^*_{old} and $\tilde{f}_{pop}$ with

$$t = \frac{\Delta_R}{4}.$$

By a union bound, with probability at least

$$1 - 4\exp\left(-\frac{K\Delta_R^2}{8B_\ell^2}\right),$$

both inequalities

$$\big|\mathcal{L}_{\text{test}}(f^*_{\text{old}}) - R(f^*_{\text{old}})\big| \le \frac{\Delta_R}{4}, \qquad \big|\mathcal{L}_{\text{test}}(\tilde{f}_{pop}) - R(\tilde{f}_{pop})\big| \le \frac{\Delta_R}{4}$$

hold simultaneously. On this event,

$$\mathcal{L}_{\text{test}}(\tilde{f}_{pop}) - \mathcal{L}_{\text{test}}(f^*_{\text{old}}) \le \big[R(\tilde{f}_{pop}) - R(f^*_{\text{old}})\big] + 2\cdot\frac{\Delta_R}{4} = -\Delta_R + \frac{\Delta_R}{2} = -\frac{\Delta_R}{2}.$$

Therefore,

$$\mathcal{L}_{\text{test}}(\tilde{f}_{pop}) \le \mathcal{L}_{\text{test}}(f^*_{\text{old}}) - \frac{\Delta_R}{2}. \tag{3}$$

Step 4: Move from the jumpboard model to the trained new model. Since $\tilde{f}_{pop}, f_{\text{new}} \in \mathcal{H}_{\text{new}}$, Lemma 1 yields the following.

With probability at least $1 - \delta_1$,

$$\mathcal{L}_{\text{train}}(\tilde{f}_{\eta_0}) \le R(\tilde{f}_{\eta_0}) + \epsilon_1, \qquad \epsilon_1 \triangleq \epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H}_{\text{new}}, M, \delta_1). \tag{4}$$

By the optimization assumption,

$$\mathcal{L}_{\text{train}}(f_{\text{new}}) \le \mathcal{L}_{\text{train}}(\tilde{f}_{\eta_0}) - \Delta_{\text{ERM}}, \qquad \Delta_{\text{ERM}} \ge 0. \tag{5}$$

With probability at least $1 - \delta_2$,

$$R(f_{\text{new}}) \le \mathcal{L}_{\text{train}}(f_{\text{new}}) + \epsilon_2, \qquad \epsilon_2 \triangleq \epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H}_{\text{new}}, M, \delta_2). \tag{6}$$

Finally, conditioning on the training set, f_{new} is fixed, so another application of Hoeffding with $t = \Delta_R/4$ gives

$$\mathcal{L}_{\text{test}}(f_{\text{new}}) \le R(f_{\text{new}}) + \frac{\Delta_R}{4} \tag{7}$$

with probability at least

$$1 - 2\exp\left(-\frac{K\Delta_R^2}{8B_\ell^2}\right).$$

Let

$$\epsilon_{\text{gen}}^{\text{norm}} \triangleq \max\{\epsilon_1, \epsilon_2\}.$$

Then

$$\epsilon_1 + \epsilon_2 \le 2\epsilon_{\text{gen}}^{\text{norm}}. \tag{8}$$

Step 5: Close the whole chain. On the intersection of the events in Steps 3 and 4, we have

$$\begin{aligned}
\mathcal{L}_{\text{test}}(f_{\text{new}}) &\le R(f_{\text{new}}) + \frac{\Delta_R}{4} \\
&\le \mathcal{L}_{\text{train}}(f_{\text{new}}) + \epsilon_2 + \frac{\Delta_R}{4} \\
&\le \mathcal{L}_{\text{train}}(\tilde{f}_{pop}) - \Delta_{\text{ERM}} + \epsilon_2 + \frac{\Delta_R}{4} \\
&\le R(\tilde{f}_{pop}) + \epsilon_1 - \Delta_{\text{ERM}} + \epsilon_2 + \frac{\Delta_R}{4} \\
&= R(f_{\text{old}}^*) - \Delta_R - \Delta_{\text{ERM}} + \epsilon_1 + \epsilon_2 + \frac{\Delta_R}{4}.
\end{aligned}$$

Moreover, on the Hoeffding event for f_{old}^*,

$$R(f_{\text{old}}^*) \le \mathcal{L}_{\text{test}}(f_{\text{old}}^*) + \frac{\Delta_R}{4}.$$

Substituting above equation into the previous bound gives

$$\mathcal{L}_{\text{test}}(f_{\text{new}}) \le \mathcal{L}_{\text{test}}(f_{\text{old}}^*) - \frac{\Delta_R}{2} - \Delta_{\text{ERM}} + \epsilon_1 + \epsilon_2.$$

Then, we arrive at

$$\boxed{\mathcal{L}_{\text{test}}(f_{\text{new}}) \le \mathcal{L}_{\text{test}}(f_{\text{old}}^*) - \frac{\Delta_R}{2} - \Delta_{\text{ERM}} + 2\epsilon_{\text{gen}}^{\text{norm}}}.$$

The total failure probability is bounded by

$$\delta_{\text{gen}} + \delta_{\text{test}}, \qquad \delta_{\text{gen}} = \delta_1 + \delta_2, \qquad \delta_{\text{test}} = 6\exp\left(-\frac{K\Delta_R^2}{8B_\ell^2}\right),$$

where the factor 6 comes from the three Hoeffding applications: f_{old}^*, $\tilde{f}_{\eta_0}$, and f_{new}.

Finally, the two stated consequences follow immediately:

- If
$$2\epsilon_{\text{gen}}^{\text{norm}} \le \frac{\Delta_R}{2},$$
then, since $\Delta_{\text{ERM}} \ge 0$,
$$\mathcal{L}_{\text{test}}(f_{\text{new}}) \le \mathcal{L}_{\text{test}}(f_{\text{old}}^*).$$
- If
$$\frac{\Delta_R}{2} + \Delta_{\text{ERM}} > 2\epsilon_{\text{gen}}^{\text{norm}},$$
then
$$\mathcal{L}_{\text{test}}(f_{\text{new}}) < \mathcal{L}_{\text{test}}(f_{\text{old}}^*).$$

This completes the proof. □

H Proof of Theorem 2: Directly on train/test risks.

Proof. Define
$$\epsilon_{\text{TT}} \coloneqq \epsilon_M + \epsilon_K.$$
Recall that
$$\epsilon_M = \epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H}_{\text{new}}, M, \delta), \qquad \epsilon_K = \epsilon_{\text{gen}}^{\text{norm}}(\mathcal{H}_{\text{new}}, K, \delta).$$
Let $\tilde{f}_S \in \mathcal{H}_{\text{new}}$ be the jumpboard model used in Theorem 2. We allow the degenerate case $\tilde{f}_S = f_{\text{old}}^*$, which is valid because the zero residual block $h_0 = 0$ recovers the old model inside $\mathcal{H}_{\text{new}}$. By definition,
$$\Delta_R^{\text{test}} = \mathcal{L}_{\text{test}}(f_{\text{old}}^*) - \mathcal{L}_{\text{test}}(\tilde{f}_S).$$

Step 1: uniform train–test comparison. By Lemma 1, applied to the training sample S_{train}, with probability at least $1 - \delta$,
$$\forall f \in \mathcal{H}_{\text{new}}, \qquad |\mathcal{L}_{\text{train}}(f) - R(f)| \le \epsilon_M.$$
Applying the same lemma to the independent test sample S_{test}, with probability at least $1 - \delta$,
$$\forall f \in \mathcal{H}_{\text{new}}, \qquad |\mathcal{L}_{\text{test}}(f) - R(f)| \le \epsilon_K.$$
By the union bound, both events hold simultaneously with probability at least $1 - 2\delta$. On this joint event, for every $f \in \mathcal{H}_{\text{new}}$,
$$\begin{aligned} |\mathcal{L}_{\text{train}}(f) - \mathcal{L}_{\text{test}}(f)| &\le |\mathcal{L}_{\text{train}}(f) - R(f)| + |R(f) - \mathcal{L}_{\text{test}}(f)| \\ &\le \epsilon_M + \epsilon_K = \epsilon_{\text{TT}}. \end{aligned}$$
Equivalently, for every $f \in \mathcal{H}_{\text{new}}$,
$$\mathcal{L}_{\text{train}}(f) \le \mathcal{L}_{\text{test}}(f) + \epsilon_{\text{TT}}, \qquad \mathcal{L}_{\text{test}}(f) \le \mathcal{L}_{\text{train}}(f) + \epsilon_{\text{TT}}. \tag{TT}$$

Step 2: compare the jumpboard model on train and test. Applying equation TT to $f = \tilde{f}_S$, we obtain
$$\mathcal{L}_{\text{train}}(\tilde{f}_S) \le \mathcal{L}_{\text{test}}(\tilde{f}_S) + \epsilon_{\text{TT}}.$$

Step 3: use the optimization gain. By the optimization assumption in Theorem 2,
$$\mathcal{L}_{\text{train}}(f_{\text{new}}) \le \mathcal{L}_{\text{train}}(\tilde{f}_S) - \Delta_{\text{ERM}}, \qquad \Delta_{\text{ERM}} \ge 0.$$

Step 4: move the optimized model back to test loss. Applying equation TT to $f = f_{\text{new}}$, we have
$$\mathcal{L}_{\text{test}}(f_{\text{new}}) \le \mathcal{L}_{\text{train}}(f_{\text{new}}) + \epsilon_{\text{TT}}.$$

Step 5: combine the inequalities. Combining above inequations, we get

$$\begin{aligned}\mathcal{L}_{\text{test}}(f_{\text{new}}) &\le \mathcal{L}_{\text{train}}(f_{\text{new}}) + \epsilon_{\text{TT}} \\ &\le \mathcal{L}_{\text{train}}(\tilde{f}_S) - \Delta_{\text{ERM}} + \epsilon_{\text{TT}} \\ &\le \mathcal{L}_{\text{test}}(\tilde{f}_S) - \Delta_{\text{ERM}} + 2\epsilon_{\text{TT}}.\end{aligned}$$

Using the definition of Δ_R^{test},

$$\mathcal{L}_{\text{test}}(\tilde{f}_S) = \mathcal{L}_{\text{test}}(f^*_{\text{old}}) - \Delta_R^{\text{test}},$$

we obtain

$$\mathcal{L}_{\text{test}}(f_{\text{new}}) \le \mathcal{L}_{\text{test}}(f^*_{\text{old}}) - \Delta_R^{\text{test}} - \Delta_{\text{ERM}} + 2\epsilon_{\text{TT}}.$$

Since $\epsilon_{\text{TT}} = \epsilon_M + \epsilon_K$, this is exactly

$$\boxed{\mathcal{L}_{\text{test}}(f_{\text{new}}) \le \mathcal{L}_{\text{test}}(f^*_{\text{old}}) - \Delta_R^{\text{test}} - \Delta_{\text{ERM}} + 2(\epsilon_M + \epsilon_K)}.$$

The above inequality holds with probability at least $1 - 2\delta$, and hence also with probability at least $1 - 3\delta$, as stated in Theorem 2. The extra δ is reserved for the optional width-alignment event used to justify the nondegenerate interpretation $\Delta_R^{\text{test}} > 0$.

Degenerate case. If $\tilde{f}_S = f^*_{\text{old}}$, then

$$\Delta_R^{\text{test}} = \mathcal{L}_{\text{test}}(f^*_{\text{old}}) - \mathcal{L}_{\text{test}}(f^*_{\text{old}}) = 0.$$

Therefore the preceding bound becomes

$$\mathcal{L}_{\text{test}}(f_{\text{new}}) \le \mathcal{L}_{\text{test}}(f^*_{\text{old}}) - \Delta_{\text{ERM}} + 2(\epsilon_M + \epsilon_K).$$

Thus, if

$$\Delta_{\text{ERM}} \ge 2(\epsilon_M + \epsilon_K),$$

then

$$\mathcal{L}_{\text{test}}(f_{\text{new}}) \le \mathcal{L}_{\text{test}}(f^*_{\text{old}}).$$

Hence the theorem remains meaningful even in the degenerate regime where the test-side jump margin vanishes.

Strict improvement. If

$$\Delta_R^{\text{test}} + \Delta_{\text{ERM}} > 2(\epsilon_M + \epsilon_K),$$

then the main inequality gives

$$\mathcal{L}_{\text{test}}(f_{\text{new}}) < \mathcal{L}_{\text{test}}(f^*_{\text{old}}),$$

which proves the strict-improvement claim.

Alignment and positivity of the finite-test margin. The finite-test margin in Theorem 2 can be related to the train/test alignment quantity by a local first-order expansion. Consider the empirical jumpboard path

$$f_\eta(x) = f_{\text{top}}\big(f_{\text{bot}}(x) + h_{\eta v_S}(f_{\text{bot}}(x))\big), \qquad f_0 = f^*_{\text{old}}.$$

Assume that, along this path, the test loss admits the local expansion

$$L_{\text{test}}(f_\eta) = L_{\text{test}}(f^*_{\text{old}}) - \eta\, \mu^\top g + O(\eta^2).$$

Then

$$\Delta_R^{\text{test}} = L_{\text{test}}(f^*_{\text{old}}) - L_{\text{test}}(\widetilde{f}_{\text{S}}) = \eta\, \mu^\top g + O(\eta^2).$$

Consequently, whenever $\mu^\top g > 0$, choosing $\eta > 0$ sufficiently small ensures $\Delta_R^{\text{test}} > 0$. In practice, the jumpboard step is intended to be infinitesimal. Empirical implementations often use very small perturbation scales, and in this local regime the first-order term is expected to dominate the higher-order remainderZhang et al. (2024). We use the above expansion only to interpret when the random margin Δ_R^{test} is likely to be positive; Theorem 2 itself conditions on the realized value of this margin.

Thus, we have the following corollary.

Corollary 3 (Alignment implies a positive finite-test margin)**.** *Consider the empirical jumpboard model*

$$\widetilde{f}_{S,\eta}(x) = f_{\mathrm{top}}\big(f_{\mathrm{bot}}(x) + \eta u_S(f_{\mathrm{bot}}(x))\big),$$

where u_S is the residual perturbation selected from the training activation-gradient direction. Suppose the jumpboard step size $\eta > 0$ lies in the small-step first-order regime, so that on the finite test set

$$\mathcal{L}_{\mathrm{test}}(\widetilde{f}_{S,\eta}) = \mathcal{L}_{\mathrm{test}}(f^*_{\mathrm{old}}) + \eta \frac{1}{K} \sum_{j=1}^{K} \nabla_{\widetilde{x}_j^{(l^*)}} \ell\big(f_{\mathrm{top}}(\widetilde{x}_j^{(l^*)}), \widetilde{y}_j\big)^\top u_S(\widetilde{x}_j^{(l^*)}) + o(\eta).$$

Assume further that the selected residual perturbation realizes the negative training activation-gradient direction on the test set, namely

$$\frac{1}{K} \sum_{j=1}^{K} \nabla_{\widetilde{x}_j^{(l^*)}} \ell\big(f_{\mathrm{top}}(\widetilde{x}_j^{(l^*)}), \widetilde{y}_j\big)^\top u_S(\widetilde{x}_j^{(l^*)}) = -\mu^\top g.$$

This assumption is easy to satisfy for residual architectures that admit a constant residual direction. In particular, the choice $u_S(z) \equiv -\mu$, which is determined entirely from the training sample, gives

$$\frac{1}{K} \sum_{j=1}^{K} q_j^\top u_S(\tilde{z}_j) = -\mu^\top g$$

for every test sample. Such a constant perturbation is directly realizable by residual blocks with an additive bias parameter: one may set the residual bias to $-\mu$ and the remaining residual contribution to zero. Hence, for bias-augmented residual blocks, the additional realizability condition used here is immediate. Then

$$\Delta_R^{\mathrm{test}} := \mathcal{L}_{\mathrm{test}}(f^*_{\mathrm{old}}) - \mathcal{L}_{\mathrm{test}}(\widetilde{f}_{S,\eta}) = \eta \mu^\top g + o(\eta).$$

.

Consequently, under the small-step first-order approximation,

$$\mu^\top g > 0 \quad \Longrightarrow \quad \Delta_R^{\mathrm{test}} > 0$$

for sufficiently small $\eta > 0$. In particular, when $\eta < 10^{-3}$ is in the empirically validated first-order regime, the sign of the finite-test jumpboard margin is determined by the alignment term $\mu^\top g$.

Width-dependent nondegenerate interpretation. The algebraic inequality above does not require $\Delta_R^{\mathrm{test}} > 0$. The final statement about the probability of $\Delta_R^{\mathrm{test}} > 0$ comes from the width-alignment mechanism. In particular, the width-alignment theorem in the appendix gives

$$\mathbb{P}(\mu^\top g \le 0) \le \frac{4C_\Sigma \tau^2}{M\|\bar{\mu}\|^2} + \frac{4C_\Sigma \tau^2}{K\|\bar{\mu}\|^2} + \frac{4C_\Sigma \tau^2 \,\mathrm{tr}(\Sigma)}{KM\|\bar{\mu}\|^4}.$$

When the train-side and test-side first-order directions are aligned, the jumpboard direction that decreases the train-side loss also has a favorable test-side first-order derivative. Therefore, for a sufficiently small jump size,

$$\mathcal{L}_{\mathrm{test}}(\tilde{f}) < \mathcal{L}_{\mathrm{test}}(f^*_{\mathrm{old}}),$$

which is equivalent to

$$\Delta_R^{\mathrm{test}} > 0.$$

In the uniformly active regime, where

$$\|\bar{\mu}\|^2 = \Theta(N),$$

the failure probability satisfies

$$\mathbb{P}(\mu^\top g \le 0) = O\left(\frac{1}{KN} + \frac{1}{MN}\right).$$

Hence, as the width N increases, the probability of observing a positive test-side jump $\Delta_R^{\mathrm{test}} > 0$ increases. If this alignment event is required simultaneously with the two uniform generalization events above, the union bound gives the stated probability level $1 - 3\delta$.

□

I PROOF OF THEOREM 7

Theorem 6 (Qualitative scaling mechanism in population risk)**.** *Let $\tilde{f}_{pop} \in \mathcal{H}_{\mathrm{new}}$ be the jumpboard model selected along a local descent path, and suppose $\mathcal{L}_{\mathrm{train}}(f_{\mathrm{new}}) \leq \mathcal{L}_{\mathrm{train}}(\tilde{f}_{pop}) - \Delta_{\mathrm{ERM}}, \Delta_{\mathrm{ERM}} \geq 0$. Then there exists a deterministic $\Delta_R > 0$, depending only on $\mathcal{D}$ and the current old model, such that with probability at least $1 - 2\delta$,*

$$R(f_{\mathrm{new}}) \leq R(f^*_{\mathrm{old}}) - \Delta_R - \Delta_{\mathrm{ERM}} + 2\epsilon_M.$$

In particular, if $\Delta_R + \Delta_{\mathrm{ERM}} > 2\epsilon_M$, then $R(f_{\mathrm{new}}) < R(f^*_{\mathrm{old}})$.

Theorem 7 (Qualitative scaling mechanism in population risk)**.** *Let $\tilde{f}_{pop} \in \mathcal{H}_{\mathrm{new}}$ be the jumpboard model selected along a local descent path, and suppose $\mathcal{L}_{\mathrm{train}}(f_{\mathrm{new}}) \leq \mathcal{L}_{\mathrm{train}}(\tilde{f}_{pop}) - \Delta_{\mathrm{ERM}}, \Delta_{\mathrm{ERM}} \geq 0$. Then there exists a deterministic $\Delta_R > 0$, depending only on $\mathcal{D}$ and the current old model, such that with probability at least $1 - 2\delta$,*

$$R(f_{\mathrm{new}}) \leq R(f^*_{\mathrm{old}}) - \Delta_R - \Delta_{\mathrm{ERM}} + 2\epsilon_M.$$

In particular, if $\Delta_R + \Delta_{\mathrm{ERM}} > 2\epsilon_M$, then $R(f_{\mathrm{new}}) < R(f^*_{\mathrm{old}})$.

Proof. By Appendix E, under the population version of Condition 1, there exists a population jumpboard model

$$\widetilde{f}_{\mathrm{pop}} \in \mathcal{H}_{\mathrm{new}}$$

such that

$$R(\widetilde{f}_{\mathrm{pop}}) < R(f^*_{\mathrm{old}}).$$

Define the deterministic population margin

$$\Delta_R := R(f^*_{\mathrm{old}}) - R(\widetilde{f}_{\mathrm{pop}}) > 0.$$

Thus

$$R(\widetilde{f}_{\mathrm{pop}}) = R(f^*_{\mathrm{old}}) - \Delta_R.$$

By Condition 2, with the comparison jumpboard model $\widetilde{f} = \widetilde{f}_{\mathrm{pop}}$, the final selected model satisfies

$$L_{\mathrm{train}}(f_{\mathrm{new}}) \leq L_{\mathrm{train}}(\widetilde{f}_{\mathrm{pop}}) - \Delta_{\mathrm{ERM}}, \qquad \Delta_{\mathrm{ERM}} \geq 0.$$

Let

$$\epsilon_M := \epsilon^{\mathrm{norm}}_{\mathrm{gen}}(\mathcal{H}_{\mathrm{new}}, M, \delta).$$

By the uniform generalization bound in Corollary 2, with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{H}_{\mathrm{new}}} |R(f) - L_{\mathrm{train}}(f)| \leq \epsilon_M.$$

This event is uniform over $\mathcal{H}_{\mathrm{new}}$, so it applies to both the comparison jumpboard model $\widetilde{f}_{\mathrm{pop}}$ and the data-dependent final selected model f_{new}.

On this event,

$$R(f_{\mathrm{new}}) \leq L_{\mathrm{train}}(f_{\mathrm{new}}) + \epsilon_M$$

and

$$L_{\mathrm{train}}(\widetilde{f}_{\mathrm{pop}}) \leq R(\widetilde{f}_{\mathrm{pop}}) + \epsilon_M.$$

Therefore,

$$\begin{aligned} R(f_{\mathrm{new}}) &\leq L_{\mathrm{train}}(f_{\mathrm{new}}) + \epsilon_M \\ &\leq L_{\mathrm{train}}(\widetilde{f}_{\mathrm{pop}}) - \Delta_{\mathrm{ERM}} + \epsilon_M \\ &\leq R(\widetilde{f}_{\mathrm{pop}}) - \Delta_{\mathrm{ERM}} + 2\epsilon_M \\ &= R(f^*_{\mathrm{old}}) - \Delta_R - \Delta_{\mathrm{ERM}} + 2\epsilon_M. \end{aligned}$$

Thus,

$$R(f_{\mathrm{new}}) \leq R(f^*_{\mathrm{old}}) - \Delta_R - \Delta_{\mathrm{ERM}} + 2\epsilon_M.$$

The theorem states the slightly more conservative probability $1-2\delta$; the above uniform event already gives the inequality with probability at least $1-\delta$. The weaker $1-2\delta$ statement follows immediately.

Finally, if

$$\Delta_R + \Delta_{\mathrm{ERM}} > 2\epsilon_M,$$

then

$$R(f_{\mathrm{new}}) < R(f^*_{\mathrm{old}}).$$

This proves the strict-improvement claim. □

J Discussion: relation between Theorems 5 and 2

Theorems 5 and 2 provide two parallel rigorous routes to the same scaling conclusion, but they operate through different intermediate objects and are effective in different regimes.

Theorem 5 proceeds through the population risk. Its proof first constructs a jumpboard model $\tilde{f}_{pop}$ whose population risk is lower than that of f^*_{old} by a deterministic margin $\Delta_R > 0$, then transfers this margin to the finite test set via Hoeffding's inequality, and finally compares $\tilde{f}_{pop}$ and f_{new} through the optimization gain Δ_{ERM} and the norm-based generalization bound in Lemma 1. As a result, Theorem 5 yields the bound

$$\mathcal{L}_{test}(f_{new}) \le \mathcal{L}_{test}(f^*_{old}) - \frac{\Delta_R}{2} - \Delta_{ERM} + 2\epsilon_M.$$

Its main advantage is that the generalization penalty is tighter: only the training-side complexity term $2\epsilon_M$ appears.

By contrast, Theorem 2 works entirely at the train/test level. It does not introduce the population risk as an intermediate quantity, and it does not use Hoeffding's inequality. Instead, it compares $\mathcal{L}_{train}$ and $\mathcal{L}_{test}$ directly on both $\tilde{f}$ and f_{new} through two applications of Lemma 1, which gives

$$\mathcal{L}_{test}(f_{new}) \le \mathcal{L}_{test}(f^*_{old}) - \Delta_R^{test} - \Delta_{ERM} + 2(\epsilon_M + \epsilon_K).$$

The price is that the generalization penalty is looser, since both the training-side and test-side complexity terms appear. The gain is that this route remains meaningful even when the population-level jump margin becomes degenerate.

This difference is especially important near the "deepest-model" regime. In Theorem 5, the useful margin is

$$\frac{\Delta_R}{2} + \Delta_{ERM} - 2\epsilon_M.$$

Here Δ_R comes from the first-order descent of the population-risk curve along the jump path. Under the present C^1 assumptions, it is only guaranteed to be positive, but it may be arbitrarily small. Thus Δ_R should be understood primarily as a directional certificate: it proves that the enlarged hypothesis class is strictly better, but in practice the dominant improvement is often carried by Δ_{ERM}, namely the extra empirical optimization gain available in $\mathcal{H}_{new}$. In other words, Δ_R certifies that the new class contains a better point, whereas Δ_{ERM} measures how much of this enlarged search space is actually exploited by optimization.

This also explains the complementary strengths of the two theorems. In the regular regime, where Δ_R is not too small and the sample sizes are large, Theorem 5 is sharper because it pays only $2\epsilon_M$. However, when $\Delta_R \to 0$, the Hoeffding transfer in Theorem 5 becomes ineffective, since its test-side success probability depends on

$$\delta_{test} = 6\exp\left(-\frac{K\Delta_R^2}{8B_\ell^2}\right),$$

which no longer provides useful control when the margin collapses. In this regime, Theorem 2 is more robust: it still yields a meaningful statement as long as the optimization gain Δ_{ERM} can dominate the combined train/test generalization cost $2(\epsilon_M + \epsilon_K)$. Thus, Theorem 5 is tighter in the nondegenerate regime, whereas Theorem 2 is safer in the degenerate one.

The two margins have different meanings. In Theorem 5, Δ_R is a deterministic population-risk margin associated with $\widetilde{f}_{\mathrm{pop}}$. In Theorem 2, Δ_R^{test} is a random finite-test margin associated with the

training-set jumpboard model $\widetilde{f}_S$. Thus Theorem 2 does not require a non-vanishing deterministic population margin; its strict improvement condition is instead

$$\Delta_R^{\mathrm{test}} + \Delta_{\mathrm{ERM}} > 2(\epsilon_M + \epsilon_K).$$

The role of width is also slightly different in the two routes. Through Theorem 4, width can strengthen the nondegenerate interpretation by making the alignment between train and test gradients more reliable, thereby supporting the sign of Δ_R^{test} in the nondegenerate version of Theorem 2. In the bias-type special case, width may also enlarge the effective first-order signal through $\|\bar{\mu}\|^2$, which in turn helps the detectability of the jump margin. At the same time, for both routes, increasing the data size reduces the complexity terms provided by Lemma 1, so the generalization cost can eventually be dominated by the available improvement margin.

From this perspective, Theorems 5 and 2 should not be viewed as competing statements, but as two independent closures of the same scaling mechanism. When both apply, they provide two separate proofs of the same qualitative conclusion; when the population-level margin is too small for Theorem 5 to be effective, Theorem 2 still survives as a fallback route. Their coexistence therefore gives a more robust justification than either theorem alone.

Finally, Theorem 7 is the common limiting form of this two-route picture. As $K \to \infty$, the test loss converges to the population risk and the extra test-side term disappears, so both Theorem 5 and Theorem 2 reduce to the same population-level qualitative scaling statement. In this sense, Theorem 7 is the infinite-test-set limit, while Theorems 5 and 2 are its two finite-sample realizations.

K Properties Near the Deepest-Model Regime

In realistic training environments, the models we care most about are usually not those at arbitrary intermediate depths, but those that are close to the deepest model. The reason is simple: when sufficient compute is available, we naturally expect the best-performing model to lie near the deepest feasible depth, since such a model has already exploited most of the representational benefit brought by depth expansion. Accordingly, the meaningful question is no longer merely whether further depth can still help, but rather what determines the remaining gain once the model is already close to the deepest-model regime. This subsection summarizes several direct conclusions for that regime.

Performance equivalence between the springboard model and the trained model. Under the above equivalence, the test risk of f_{new} can be rewritten as

$$\mathcal{L}_{test}(f_{new}) < \mathcal{L}_{test}(f^*_{old}) - \eta\, \mu^\top g + \Gamma_{\mathrm{deep}}(M, K, \delta),$$

where

$$\Gamma_{\mathrm{deep}}(M, K, \delta) := \epsilon_{gen}(\{f^*_{old}\}, M, \delta) + \epsilon_{gen}(\{f^*_{old}\}, K, \delta) + 2L_\ell L_{top}\big(\psi(M) + \psi(K)\big).$$

and $-\eta\, \mu^\top g$ is the first-order improvement term that remains visible in the near-deepest regime, while $\Gamma_{\mathrm{deep}}(M, K, \delta)$ represents the finite-sample generalization cost.

The most important aspect of the above formula is not merely that it still preserves a first-order improvement term, but that it shifts the generalization cost from the *whole-network complexity* to the *springboard-subspace complexity*. Here, the old model components (f_{top}, f_{bot}) are effectively fixed, and the only varying part is the newly introduced residual block $h \in \mathcal{F}_{res}$. As a result, the sample size needed to control generalization no longer has to be estimated in terms of the absolute scale of the full network parameter count; instead, it is governed by the decay rate of the complexity envelope ψ. In other words, near the deepest-model regime, the dominant factor in data demand is no longer how large the full model is, but whether the effective complexity of the newly added degrees of freedom remains sufficiently small.

Since the additional degrees of freedom are compressed into the springboard subspace in the near-deepest regime, the generalization cost enters the main formula only through $2L_\ell L_{top}\big(\psi(M) + \psi(K)\big)$, rather than being dominated directly by a full-network parameter-scale term. Therefore, under the same data scale, the effective generalization burden near the deepest-model regime may actually be lower than that in a full-network expansion regime farther away from the deepest model.

That is, being close to the deepest model does not necessarily imply a heavier generalization penalty; on the contrary, because the newly activated directions are already localized, the generalization control can be more refined and less costly.

As the model approaches the deepest-model regime, the deterministic improvement brought by further depth expansion becomes increasingly small. Consequently, the dominant term becomes the sign and stability of $-\eta\,\mu^\top g$. At this stage, the key question is no longer whether further depth can still be added, but whether this already weakened first-order signal can still be observed reliably under a finite test set. It is precisely here that width becomes decisive: the larger the width, the more stable the gradient alignment, and the more likely $\mu^\top g$ remains positive, so that the test-set improvement is less likely to be washed out by sampling noise. Hence, in the near-deepest regime, the marginal role of depth gradually weakens, while width becomes the dominant factor controlling whether the remaining gain can still be realized.

L Removing Assumption 4 via Engineering ε

In Lemma 1, Assumption 4 requires that the input to each normalization layer stay uniformly away from zero:

$$\inf_{x,\theta,l} \|z_l(x) + h_l(z_l(x);\theta)\| \geq r_{\min} > 0.$$

This condition is used only to guarantee a uniform Lipschitz bound for the normalization map. In practical deep-learning implementations, however, the denominator of the normalization layer always contains a small engineering constant $\varepsilon_{\text{eng}} > 0$. This removes the singularity at the origin and makes Assumption 4 unnecessary.

RMSNorm with engineering ε. Consider the stabilized RMSNorm map

$$\mathrm{RMSNorm}_\varepsilon(x) = \mathrm{diag}(\gamma)\,\frac{x}{\sqrt{\|x\|^2/N + \varepsilon_{\text{eng}}}},$$

where $\gamma = (\gamma_1, \ldots, \gamma_N)$ and

$$\gamma_{\max} \triangleq \max_i |\gamma_i|.$$

Let

$$s(x) \triangleq \sqrt{\|x\|^2/N + \varepsilon_{\text{eng}}}.$$

Then

$$\mathrm{RMSNorm}_\varepsilon(x) = \mathrm{diag}(\gamma)\,\frac{x}{s(x)}.$$

Differentiating gives

$$J_\varepsilon(x) = \frac{\mathrm{diag}(\gamma)}{s(x)} - \frac{\mathrm{diag}(\gamma)\,xx^\top}{N\,s(x)^3}.$$

Hence

$$\|J_\varepsilon(x)\|_\sigma \leq \frac{\gamma_{\max}}{s(x)} + \frac{\gamma_{\max}\|x\|^2}{N\,s(x)^3}.$$

Since

$$s(x)^2 = \frac{\|x\|^2}{N} + \varepsilon_{\text{eng}} \geq \varepsilon_{\text{eng}},$$

we have

$$\frac{1}{s(x)} \leq \frac{1}{\sqrt{\varepsilon_{\text{eng}}}}, \qquad \frac{\|x\|^2/N}{s(x)^3} = \frac{\|x\|^2/N}{(\|x\|^2/N + \varepsilon_{\text{eng}})^{3/2}} \leq \frac{1}{\sqrt{\varepsilon_{\text{eng}}}}.$$

Therefore

$$\|J_\varepsilon(x)\|_\sigma \leq \frac{2\gamma_{\max}}{\sqrt{\varepsilon_{\text{eng}}}}, \qquad \forall x \in \mathbb{R}^N.$$

So $\mathrm{RMSNorm}_\varepsilon$ is globally Lipschitz, with a uniform constant

$$c_\varepsilon \leq \frac{2\gamma_{\max}}{\sqrt{\varepsilon_{\text{eng}}}}.$$

LayerNorm with engineering ε. The same conclusion holds for LayerNorm. Indeed, after subtracting the mean, LayerNorm has the same structural form

$$\mathrm{LayerNorm}_{\varepsilon}(x) = \mathrm{diag}(\gamma)\,\frac{Px}{\sqrt{\|Px\|^2/N + \varepsilon_{\mathrm{eng}}}},$$

where

$$P = I - \frac{1}{N}\mathbf{1}\mathbf{1}^{\top}$$

is the centering projection. Since $\|P\|_{\sigma} \le 1$, the same calculation yields

$$\|\nabla \mathrm{LayerNorm}_{\varepsilon}(x)\|_{\sigma} \le \frac{2\gamma_{\max}}{\sqrt{\varepsilon_{\mathrm{eng}}}}, \qquad \forall x \in \mathbb{R}^N.$$

Thus stabilized LayerNorm is also globally Lipschitz.

Fixed-statistics BatchNorm. For BatchNorm, our theory applies to the evaluation-mode network after the running mean and variance are frozen. In this regime, BatchNorm is a deterministic affine single-sample map and can be absorbed into the Lipschitz constants in Assumption 4ϵ. We do not analyze mini-batch-dependent training-time BatchNorm dynamics; the optimization process may use BatchNorm during training, but our risk-comparison theorem is applied to the resulting evaluation-mode function. Training-time BatchNorm dynamics are outside the proof, but the trained evaluation-mode BatchNorm network is covered as a fixed-statistics normalized residual network.

BatchNorm requires a separate distinction between fixed-statistics and training-time forms. In the fixed-statistics regime, such as evaluation mode with frozen running mean and variance, BatchNorm is a deterministic affine single-sample map:

$$\mathrm{BN}_{\epsilon}(x) = \gamma \odot \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon_{\mathrm{BN}}}} + \beta,$$

where μ, σ^2, γ, and β are fixed. Therefore,

$$\|\mathrm{BN}_{\epsilon}(x) - \mathrm{BN}_{\epsilon}(x')\| \le c_{\mathrm{BN}}\|x - x'\|, \qquad c_{\mathrm{BN}} := \max_i \frac{|\gamma_i|}{\sqrt{\sigma_i^2 + \epsilon_{\mathrm{BN}}}}.$$

Thus fixed-statistics BatchNorm satisfies the Lipschitz requirement in Assumption 4ϵ, with its constant absorbed into $c_{l,\epsilon}$. If the reachable normalization-input set is bounded, then its output is also bounded because the map is affine.

Training-time BatchNorm is not covered by the present single-sample function-class argument without additional assumptions, because its output depends on mini-batch statistics and therefore couples different samples in the batch. Throughout this paper, BatchNorm is covered only in the fixed-statistics sense.

Consequence for Lemma 1. With engineering $\varepsilon_{\mathrm{eng}}$, Step 2 in the proof of Lemma 1 no longer needs Assumption 4(N). The layer map

$$T_l(z) = \mathcal{N}_l(z + h_l(z))$$

satisfies

$$\|T_l(u) - T_l(v)\| \le c_{\varepsilon}(1 + s_l)\|u - v\|, \qquad c_{\varepsilon} \le \frac{2\gamma_{\max}}{\sqrt{\varepsilon_{\mathrm{eng}}}}.$$

Accordingly, one may replace the old bound

$$c_l \le \frac{2\gamma_{\max}\sqrt{N}}{r_{\min}}$$

by the uniform engineering bound

$$c_l \le \frac{2\gamma_{\max}}{\sqrt{\varepsilon_{\mathrm{eng}}}}.$$

Therefore Assumption 4(N) can be removed from Lemma 1, at the price of letting the constants depend on $\varepsilon_{\mathrm{eng}}$.

More precisely, define

$$\Lambda_l^{(\varepsilon)} \triangleq L_{\mathrm{top}}\, L_l^{(p)}\, B_*\, c_\varepsilon \prod_{k>l} c_\varepsilon(1+s_k), \qquad \Lambda_{\max}^{(\varepsilon)} \triangleq \max_l \Lambda_l^{(\varepsilon)}.$$

Then the same proof as in Lemma 1 yields

$$\hat{\mathfrak{R}}_S(\ell \circ \mathcal{H}) \le B_\ell \sqrt{\frac{2d \log\!\big(2\bar{b} L L_\ell \Lambda_{\max}^{(\varepsilon)}/\epsilon_0 + 1\big)}{M}} + \epsilon_0,$$

and in particular, taking $\epsilon_0 = M^{-1/2}$,

$$\hat{\mathfrak{R}}_S(\ell \circ \mathcal{H}) = \mathcal{O}\left(B_\ell \sqrt{\frac{d \log M}{M}} \right).$$

Interpretation. This route is purely engineering, but it is fully compatible with practical implementations: the singularity at zero is removed directly at the level of the normalization operator, so the abstract nondegeneracy requirement

$$\inf_{x,\theta,l} \|z_l + h_l(z_l;\theta)\| > 0$$

is no longer needed. In this sense, the proof of Lemma 1 becomes closer to the actual normalized residual networks used in practice.

M From Qualitative Scaling to a Power-Law Form

The main results of this paper are qualitative test-risk comparison theorems: Theorem 5 gives a sharper route through population risk, while Theorem 2 gives a more robust route directly at the train/test level. This appendix explains how these qualitative statements can be converted into a power-law form under additional model-dependent assumptions.

The purpose of this section is not to claim that the power law follows unconditionally from the main theorems. Rather, the main theorems provide the mechanism of improvement, while the additional assumptions below specify how the remaining first-order improvement signal decays as the model approaches the irreducible-risk regime.

M.1 Depth–width coupling and parameter count

Near the deepest-model regime, the remaining first-order improvement signal may become weak. Therefore, increasing depth alone is generally insufficient: the width must also grow so that weak descent signals remain observable at finite sample size. We encode this requirement through a depth–width coupling function.

Depth–width coupling. Let

$$u : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$$

be a coupling function such that

$$L = u(N), \qquad \lim_{N\to\infty} u(N) = \infty.$$

Here N denotes the hidden width and L denotes the effective depth. We do not assume a specific form of u. Instead, u should be interpreted as the depth that can be reliably supported at width N.

Parameter count. For matrix-like residual architectures, the number of parameters per layer is typically proportional to N^2. Hence, along the coupled scaling path $L = u(N)$, the total parameter count can be written as

$$P(N) = a_{\mathrm{arch}}\, u(N)\, N^2,$$

where $a_{\mathrm{arch}} > 0$ is an architecture-dependent constant. Since N^2 is increasing and $u(N) \to \infty$, the map $P(N)$ is eventually strictly increasing. We may therefore define its inverse $N(P)$, and set

$$L(P) := u(N(P)).$$

Thus $L(P)$ is the depth supported by parameter budget P along the chosen depth–width scaling path.

M.2 PARAMETER-COUNT FORM OF THE QUALITATIVE THEOREM

Theorems 5 and 2 can be rewritten along the parameterized path $P \mapsto (L(P), N(P))$.

Proposition 1 (Parameter-count qualitative scaling)**.** *Fix a parameter budget P, and let $N(P)$ and $L(P) = u(N(P))$ be defined as above. Suppose the model with width $N(P)$ and depth $L(P)$ satisfies the assumptions of either Theorem 5 or Theorem 2.*

For the population-risk route, if

$$\frac{\Delta_R(P)}{2} + \Delta_{\mathrm{ERM}}(P) > 2\epsilon_M(P),$$

then

$$\mathcal{L}_{\mathrm{test}}(f_{\mathrm{new}}(P)) < \mathcal{L}_{\mathrm{test}}(f^*_{\mathrm{old}}(P)).$$

For the direct train/test route, if

$$\Delta_R^{\mathrm{test}}(P) + \Delta_{\mathrm{ERM}}(P) > 2(\epsilon_M(P) + \epsilon_K(P)),$$

then

$$\mathcal{L}_{\mathrm{test}}(f_{\mathrm{new}}(P)) < \mathcal{L}_{\mathrm{test}}(f^*_{\mathrm{old}}(P)).$$

In the nondegenerate train/test route, the probability that the favorable test-side direction is observable is controlled by the alignment estimate

$$\mathbb{P}(\mu^\top g \le 0) \le \frac{4C_\Sigma\tau^2}{M\|\bar{\mu}\|^2} + \frac{4C_\Sigma\tau^2}{K\|\bar{\mu}\|^2} + \frac{4C_\Sigma\tau^2\,\mathrm{tr}(\Sigma)}{KM\|\bar{\mu}\|^4}.$$

Consequently,

$$\mathbb{P}(\mu^\top g > 0) \ge 1 - \frac{4C_\Sigma\tau^2}{M\|\bar{\mu}\|^2} - \frac{4C_\Sigma\tau^2}{K\|\bar{\mu}\|^2} - \frac{4C_\Sigma\tau^2\,\mathrm{tr}(\Sigma)}{KM\|\bar{\mu}\|^4}.$$

If additionally $\mathrm{tr}(\Sigma) \le N\tau^2$ and $\|\bar{\mu}\|^2 = \Theta(N)$, then along the parameterized path,

$$\mathbb{P}(\mu^\top g \le 0) = O\!\left(\frac{1}{MN(P)} + \frac{1}{KN(P)}\right).$$

Proof. The first two claims are direct restatements of Theorem 5 and Theorem 2 after substituting the coupled width and depth

$$N = N(P), \qquad L = L(P) = u(N(P)).$$

The final probability bound follows from the alignment estimate:

$$\mathbb{P}(\mu^\top g \le 0) \le \frac{4C_\Sigma\tau^2}{M\|\bar{\mu}\|^2} + \frac{4C_\Sigma\tau^2}{K\|\bar{\mu}\|^2} + \frac{4C_\Sigma\tau^2\,\mathrm{tr}(\Sigma)}{KM\|\bar{\mu}\|^4}.$$

Taking complements gives

$$\mathbb{P}(\mu^\top g > 0) \ge 1 - \frac{4C_\Sigma\tau^2}{M\|\bar{\mu}\|^2} - \frac{4C_\Sigma\tau^2}{K\|\bar{\mu}\|^2} - \frac{4C_\Sigma\tau^2\,\mathrm{tr}(\Sigma)}{KM\|\bar{\mu}\|^4}.$$

If $\mathrm{tr}(\Sigma) \le N\tau^2$ and $\|\bar{\mu}\|^2 = \Theta(N)$, then along the parameterized path,

$$\mathbb{P}(\mu^\top g \le 0) = O\!\left(\frac{1}{MN(P)} + \frac{1}{KN(P)}\right),$$

up to the mixed covariance term. For the special linear coupling $u(N) = \kappa N$, we have

$$P = a_{\mathrm{arch}}\kappa N^3, \qquad N(P) = \left(\frac{P}{a_{\mathrm{arch}}\kappa}\right)^{1/3}.$$

Hence

$$\mathbb{P}(\mu^\top g \le 0) = O\!\left(\frac{1}{MP^{1/3}} + \frac{1}{KP^{1/3}}\right),$$

up to architecture-dependent constants.

□

In particular, if the uniformly active regime satisfies

$$\|\bar{\mu}\|^2 = \Theta(N),$$

then along the parameterized path,

$$\mathbb{P}(\mu^\top g > 0) \geq 1 - O\left(\frac{1}{MN(P)} + \frac{1}{KN(P)}\right).$$

For the special linear coupling $u(N) = \kappa N$, we have

$$P = a_{\text{arch}}\kappa N^3, \qquad N(P) = \left(\frac{P}{a_{\text{arch}}\kappa}\right)^{1/3}.$$

Hence,

$$\mathbb{P}(\mu^\top g > 0) \geq 1 - O\left(\frac{1}{MP^{1/3}} + \frac{1}{KP^{1/3}}\right),$$

up to architecture-dependent constants.

M.3 Additional assumptions for a power-law form

We now introduce two additional model-dependent assumptions. These assumptions are not used in the main theorems. They are only used to convert the qualitative scaling mechanism into an explicit power-law shape.

Consider an iterative depth-expansion process. Let R_k be the population risk after the k-th successful depth expansion, and define the remaining excess risk

$$\Delta_k := R_k - R^\star > 0,$$

where $R^\star$ denotes the irreducible limiting risk.

Assumption G: gradient–risk coupling. There exist constants $\gamma > 0$ and $\beta \geq 0$ such that

$$\|\bar{\mu}_k\|^2 = c_G\, \Delta_k^{1+\beta}.$$

This assumption states that the population-level first-order signal becomes weaker as the remaining excess risk decreases.

Assumption G': single-step gain quantification. There exists a finite curvature constant $q > 0$ such that the population-risk improvement at step k satisfies

$$\Delta_R^{(k)} \geq \frac{\|\bar{\mu}_k\|^2}{2q}.$$

Equivalently, the local one-step improvement is at least quadratic-gradient scale up to the curvature constant q.

Assumption G characterizes how the remaining excess risk controls the population first-order signal. Assumption G' converts this first-order signal into a lower bound on the risk decrease achieved by one successful depth expansion.

M.4 Recursion and its power-law upper envelope

Combining Assumptions G and G' gives

$$\Delta_R^{(k)} \geq \frac{c_G}{2q}\Delta_k^{1+\beta}.$$

Let

$$c := \frac{c_G}{2q}.$$

In the population/asymptotic closure described above,

$$\Delta_{k+1} \leq \Delta_k - c\Delta_k^{1+\beta}.$$

Case $\beta = 0$. If $\beta = 0$, the recursion becomes

$$\Delta_{k+1} \le (1 - c)\Delta_k.$$

When $0 < c < 1$, this gives exponential decay:

$$\Delta_k \le (1 - c)^k \Delta_0.$$

Thus $\beta = 0$ corresponds to an exponential regime rather than a power-law regime.

Case $\beta > 0$. When $\beta > 0$, the recursion implies the standard power-law upper envelope

$$\Delta_k \le \left(\Delta_0^{-\beta} + \beta c k\right)^{-1/\beta}.$$

Hence

$$\Delta_k = O\left(k^{-1/\beta}\right).$$

This is the sense in which the qualitative scaling mechanism becomes a power-law form under Assumptions G and G'. A matching asymptotic equivalence $\Delta_k \sim Ck^{-1/\beta}$ would require an additional lower-bound or approximate-equality assumption on the one-step gain.

Theorem 8 (Power-law upper envelope under a general depth–width coupling)**.** *Assume Assumptions G and G' with $\beta > 0$. Suppose that the number of successful depth-expansion steps is asymptotically proportional to the supported depth $L(P) = u(N(P))$. Then*

$$R(f_P) - R^\star = O\left(u(N(P))^{-1/\beta}\right),$$

where $N(P)$ is implicitly determined by

$$P = a_{\mathrm{arch}} u(N(P)) N(P)^2.$$

Proof. By the recursion derived above,

$$\Delta_k = R_k - R^\star = O(k^{-1/\beta}).$$

Along the coupled scaling path, the number of successful depth-expansion steps satisfies

$$k(P) \asymp L(P) = u(N(P)).$$

Substituting $k(P)$ into the preceding upper envelope gives

$$R(f_P) - R^\star = O\left(L(P)^{-1/\beta}\right) = O\left(u(N(P))^{-1/\beta}\right).$$

□

Polynomial sublinear coupling. More generally, suppose

$$u(N) = \kappa N^{1-\nu}, \qquad 0 \le \nu < 1.$$

Then

$$P = a_{\mathrm{arch}} \kappa N^{3-\nu},$$

so

$$N(P) = \left(\frac{P}{a_{\mathrm{arch}}\kappa}\right)^{1/(3-\nu)}.$$

Thus

$$L(P) = u(N(P)) = \kappa \left(\frac{P}{a_{\mathrm{arch}}\kappa}\right)^{(1-\nu)/(3-\nu)}.$$

Substituting this into Theorem 8 gives

$$R(f_P) - R^\star \sim C_\nu \, P^{-\frac{1-\nu}{(3-\nu)\beta}}.$$

Therefore the parameter-count scaling exponent is

$$a_{\mathrm{PL}}(\nu) = \frac{1 - \nu}{(3 - \nu)\beta}.$$

ν	Coupling type	Scaling exponent a_{PL}
0	Linear coupling $L = \kappa N$	$\frac{1}{3\beta}$
$0 < \nu < 1$	Sublinear depth growth	$\frac{1-\nu}{(3-\nu)\beta}$
$\nu \to 1$	Width-dominant growth	0

Table 5: Power-law exponents induced by different depth–width coupling functions.

M.5 Reliability constraint on the coupling function

The coupling function u cannot be chosen arbitrarily. It must be compatible with finite-sample observability. The alignment estimate gives

$$\mathbb{P}(\mu^\top g \le 0) \le \frac{4C_\Sigma \tau^2}{M\|\bar{\mu}_{k,P}\|^2} + \frac{4C_\Sigma \tau^2}{K\|\bar{\mu}_{k,P}\|^2} + \frac{4C_\Sigma \tau^2 \operatorname{tr}(\Sigma)}{KM\|\bar{\mu}_{k,P}\|^4}.$$

Thus reliable observation of the k-th depth-expansion gain requires the right-hand side to be small. Under Assumption G,

$$\|\bar{\mu}_{k,P}\|^2 = c_G \Delta_k^{1+\beta}.$$

Using the power-law upper envelope $\Delta_k = O(k^{-1/\beta})$, the signal may decay at the scale

$$\|\bar{\mu}_{k,P}\|^2 = O\Big(k^{-(1+\beta)/\beta}\Big).$$

Therefore, as k grows, larger training size, larger test size, or stronger width-supported signal is needed to keep the alignment failure probability small. For the finite-sample reliability discussion, we may use a width-supported variant of Assumption G. In this variant, the effective gradient–risk coupling coefficient is allowed to grow with the number of active coordinates. Concretely, suppose that

$$\|\bar{\mu}_{k,P}\|^2 \asymp N(P)\Delta_k^{1+\beta}.$$

This should be interpreted as a finite-sample observability condition rather than as the constant-c_G population coupling used in the power-law recursion.

Under this width-supported signal condition, the leading terms become

$$\mathbb{P}(\mu^\top g \le 0) = O\left(\frac{k^{(1+\beta)/\beta}}{MN(P)} + \frac{k^{(1+\beta)/\beta}}{KN(P)}\right),$$

up to the mixed covariance term.

Hence a sufficient reliability condition is

$$u(N(P))^{(1+\beta)/\beta} \lesssim \delta\, N(P) \min\{M, K\}.$$

This expresses the intended depth–width–data coupling: as the supported depth $u(N(P))$ grows, width and data must grow fast enough to keep the remaining first-order signal observable.

M.6 Interpretation

This appendix separates two levels of the theory.

First, Theorems 5 and 2 establish a qualitative scaling mechanism: residual depth expansion can create a jumpboard model, optimization can convert this into training gain, and the generalization terms determine whether the gain survives on the test set.

Second, Assumptions G and G' impose an additional asymptotic model for how the remaining first-order signal decays as the population risk approaches $R^\star$. Under these assumptions, and in the population/asymptotic-data regime where statistical errors are lower order, the remaining excess risk satisfies the recursion

$$\Delta_{k+1} \le \Delta_k - \frac{c_G}{2q}\Delta_k^{1+\beta}.$$

For $\beta > 0$, this yields the power-law upper envelope

$$\Delta_k = O(k^{-1/\beta}).$$

After substituting the depth–width coupling $L = u(N)$ and the parameter-count relation

$$P = a_{\text{arch}} u(N) N^2,$$

we obtain

$$R(f_P) - R^\star = O\Big(u(N(P))^{-1/\beta}\Big).$$

Thus the power-law exponent is not universal in this framework. It depends on both the gradient–risk coupling exponent β and the depth–width coupling function u. The main theorems provide the qualitative mechanism; Appendix N shows one additional assumption-based route by which that mechanism can take a power-law form.

N Experiments Setting

N.1 Model Architecture

We adopt the ResNet architecture tailored for CIFAR-scale images. The network depth is determined by $d = 6n + 2$, where n denotes the number of BasicBlocks per stage. The network width is controlled by the base channel number C, with the three stages having C, $2C$, and $4C$ channels respectively. We systematically vary both depth and width:

- **Depths:** $d \in \{8, 14, 20, 26, 32, 38, 44, 50, 56, 62, 68, 74, 80, 86, 92, 98, 104\}$.
- **Base channels:** $C \in \{4, 6, 8, 12, 16, 24, 32, 48, 64\}$.

This yields a total of 765 trained models on CIFAR-10 and trained models on CIFAR-100. Convolutional weights are initialized from $\mathcal{N}(0, \sqrt{2/\text{fan_out}})$.

N.2 Training Settings

All models are trained using SGD with momentum 0.9, weight decay 10^{-5}, initial learning rate 0.1, and batch size 32. The maximum number of training epochs is 1,000. Early stopping is applied when the training loss improvement between consecutive epochs falls below 10^{-5} (after epoch 200 for CIFAR-10 and after epoch 300 for CIFAR-100). For data augmentation, we apply random cropping with 4-pixel padding and random horizontal flipping. All images are normalized with channel-wise mean $(0.5, 0.5, 0.5)$ and standard deviation $(0.5, 0.5, 0.5)$. Other settings are default setting via Pytorch.

N.3 Gradient Covariance Experiment Settings

The models are set to evaluation mode with all computation made fully deterministic (deterministic algorithms enabled, cuDNN deterministic mode, TF32 disabled). All random seeds are fixed to 42. We use the full CIFAR-10 training set without data augmentation, with batch size 64. Backward hooks are registered on all BasicBlock modules to capture activation gradients, and the cross-entropy loss with `reduction=sum` is used for backpropagation.

N.4 Remark

For softmax cross-entropy, under the bounded-logit consequence of Assumption 2 and normalized representation control, the loss is effectively bounded on the reachable set. Thus the bounded-loss assumption is applied to the restricted reachable function class, not to arbitrary logits in $\mathbb{R}^C$.

O TRANSFORMER EXPERIMENTS BY OTHER WORKS

O.1 DEPTH–WIDTH BEHAVIOR IN SELF-ATTENTION

Existing studies on Transformer scaling provide external evidence that is consistent with the depth–width coupling predicted by our framework. A particularly relevant example is the work of Levine et al. (Levine et al., 2020), which studies the depth-to-width interplay in self-attention networks.

In their setting, depth refers to the number of self-attention layers, while width refers to the hidden representation dimension of the Transformer. This width is not identical to the width parameter N used in our theorems, where N denotes the dimension of the intermediate representation at the insertion layer. However, it is the closest Transformer analogue: both quantities control the dimension of the residual-stream representation on which residual updates and activation gradients are defined.

Levine et al. show that self-attention exhibits two qualitatively different regimes. Below a width-dependent depth threshold, adding layers can be highly effective. Beyond this regime, however, the benefit of additional depth becomes limited unless width also grows. Their theory predicts a logarithmic dependence of the depth threshold on representation width, and their empirical study examines depths from 6 to 48. The resulting fit is then used to give quantitative depth–width allocation suggestions for large Transformer models.

Table 6: Depth–width allocation data from Levine et al. (Levine et al., 2020). The trained GPT-3 configurations are taken from Brown et al. (Brown et al., 2020), while the projected optimal depth and width are obtained from the fit in Levine et al. These values are used here only as external consistency evidence for depth–width coupling.

Model	Params	Trained depth	Trained width	Projected depth	Projected width
GPT-3 Small	125M	12	768	23	555
GPT-3 Medium	350M	24	1024	32	886
GPT-3 Large	760M	24	1536	38	1220
GPT-3 XL	1.3B	24	2048	42	1550
GPT-3 2.7B	2.7B	32	2560	47	2110
GPT-3 6.7B	6.7B	32	4096	54	3150
GPT-3 13B	13.0B	40	5140	60	4200
GPT-3 175B	175.0B	96	12288	80	13500
Projected 1T	1T	–	–	95	30100

The table illustrates two points that are useful for our discussion. First, the preferred allocation is not obtained by increasing depth alone. As the model scale grows, the projected optimal width also grows substantially. Second, at very large scale, Levine et al. predict that width becomes especially important: for a projected 1-trillion-parameter self-attention model, their fit suggests a width of about 30K together with depth around 95.

This conclusion is aligned with the qualitative message of our theory. In our framework, depth creates new possible first-order improvement directions, but near the deepest-model regime the remaining first-order signal becomes weak. Width then becomes important because it improves the finite-sample observability of these weak directions. The Transformer evidence of Levine et al. supports the same high-level principle: depth and width should not be scaled independently, and the usefulness of additional depth depends on whether the representation dimension is large enough to support the new layers.

We emphasize that this evidence is indirect. Levine et al. study self-attention expressivity and architecture allocation, whereas our theorems study residual block insertion and old-vs-new test-risk comparison in normalized residual networks. Therefore, their results should not be interpreted as a direct verification of our probability bounds. Rather, they provide external empirical and theoretical support for the broader scaling picture suggested by our work: as models approach regimes where depth alone becomes less effective, width becomes an increasingly important compensating factor.

O.2 Data–parameter coupling in Transformer language models

Existing Transformer scaling results provide external evidence consistent with the data–complexity coupling predicted by our framework. A particularly relevant example is the compute-optimal scaling analysis of Hoffmann et al. (Hoffmann et al., 2022), often referred to as the Chinchilla scaling law.

Hoffmann et al. study the following allocation question: given a fixed training compute budget, how should one trade off the number of model parameters and the number of training tokens? They train more than 400 Transformer language models, ranging from below 70 million to over 16 billion parameters, and from 5 billion to over 400 billion training tokens. Their main empirical finding is that compute-optimal scaling requires increasing model size and training tokens in approximately equal proportions. In other words, when the parameter count doubles, the compute-optimal number of training tokens should also approximately double.

This observation is directly aligned with the qualitative role of data in our theory. In our framework, model expansion can create a potential test-risk gain, but this gain is useful only when the statistical cost of the enlarged hypothesis class is controlled. Under the norm-based generalization bound, this cost depends on both the sample size and architecture-dependent complexity factors. Therefore, as the effective model complexity grows, the amount of data must also grow in order to keep the generalization penalty from dominating the improvement margin.

Table 7: Parameter–token allocation in large Transformer language models, adapted from Hoffmann et al. (Hoffmann et al., 2022). Most pre-Chinchilla large dense language models were trained on roughly 300B tokens, despite having very different parameter counts. Chinchilla instead uses fewer parameters and substantially more training tokens under a comparable compute budget.

Model	Parameters	Training tokens
LaMDA	137B	168B
GPT-3	175B	300B
Jurassic-1	178B	300B
Gopher	280B	300B
MT-NLG	530B	270B
Chinchilla	70B	1.4T

Table 7 illustrates the empirical motivation. Many large dense Transformer models increased parameter count while keeping the number of training tokens roughly fixed. Hoffmann et al. argue that this leads to undertrained models. For the compute budget used to train Gopher, their analysis predicts that a smaller model trained on more tokens should be better. They test this by training Chinchilla, a 70B-parameter model on 1.4T tokens, using approximately the same training compute as Gopher. Despite being much smaller than Gopher, Chinchilla outperforms Gopher and several larger language models on a broad set of evaluation tasks.

The same conclusion is also reflected in the fitted scaling exponents. Hoffmann et al. estimate power-law relations

$$N_{\mathrm{opt}}(C) \propto C^{a}, \qquad D_{\mathrm{opt}}(C) \propto C^{b},$$

where C is the training compute budget, N_{opt} is the compute-optimal parameter count, and D_{opt} is the compute-optimal number of training tokens.

Table 8: Estimated compute-optimal scaling exponents from Hoffmann et al. (Hoffmann et al., 2022). Across three estimation methods, the optimal parameter count and the optimal number of training tokens scale at nearly the same rate with compute.

Method	a in $N_{\mathrm{opt}} \propto C^{a}$	b in $D_{\mathrm{opt}} \propto C^{b}$
Minimum over training curves	0.50	0.50
IsoFLOP profiles	0.49	0.51
Parametric loss model	0.46	0.54
Kaplan et al. (Kaplan et al., 2020)	0.73	0.27

Table 8 supports a joint-scaling interpretation. Unlike the earlier Kaplan-style allocation, where parameter count grows much faster than data, the Chinchilla estimates suggest that the compute-optimal frontier is closer to balanced parameter–data growth. This is consistent with our qualitative prediction that data is not a passive resource: it is required to pay for the statistical cost introduced by larger models.

We emphasize that these Transformer results are not direct validations of our residual-block insertion theorem. Hoffmann et al. study autoregressive Transformer language models, whereas our theorems analyze test-risk comparison after depth expansion in normalized residual networks. Moreover, their parameter–token ratio should not be interpreted as a theorem-level constant in our setting. Rather, the Chinchilla results provide external consistency evidence for the structural message of our framework: model expansion is beneficial only when the improvement margin, optimization effect, and amount of data jointly dominate the increased statistical complexity.

O.3 Model-growth evidence from LiGO

Another line of external evidence comes from model-growth methods for Transformer training. Wang et al. (Wang et al., 2023) propose LiGO (Linear Growth Operator), a method that initializes a larger Transformer from a smaller pretrained Transformer by learning a structured linear map between their parameters. To make this mapping tractable, LiGO factorizes the transformation into width-growth and depth-growth operators, with additional structure imposed through Kronecker factorization.

LiGO is not a direct implementation of our jumpboard construction. Our theory studies the insertion of a zero-output residual block into a trained normalized residual network and asks when this expansion yields an old-vs-new test-risk improvement. LiGO instead studies efficient Transformer training: it uses a learned parameter-space growth operator to transfer information from a smaller model to a larger one. Nevertheless, the empirical message is consistent with the qualitative mechanism in our framework. Both viewpoints suggest that model expansion is more effective when the expanded model is initialized or selected in a structured way, rather than being treated as an unrelated larger model trained from scratch.

Table 9: Relation between LiGO and our depth-expansion framework. LiGO is used as external consistency evidence, not as a direct validation of our theorem.

Aspect	LiGO	Relation to our framework
Starting point	A smaller pretrained Transformer.	A trained reference model f^*_{old}.
Expansion target	A larger Transformer obtained by increasing depth and/or width.	A normalized residual network enlarged by inserting a residual block.
Bridge mechanism	A learned linear growth operator maps small-model parameters to large-model parameters.	A jumpboard model provides a structured comparison point inside the expanded hypothesis class.
Role of depth and width	The growth operator explicitly includes depth-growth and width-growth components.	Depth creates new residual directions, while width improves the finite-sample observability of weak improvement signals.
Empirical implication	Structured growth can reduce the cost of training a larger Transformer while preserving performance.	Structured expansion can be beneficial when representational gain, optimization, and generalization terms jointly dominate the cost of expansion.

This comparison supports two qualitative points. First, useful scaling is not only a matter of increasing parameter count. The way in which a model is expanded matters. LiGO shows that a larger Transformer can benefit from a structured growth initialization that reuses information from a smaller pretrained model. This is aligned with our use of a jumpboard model: the expanded class is not used

merely because it is larger; it is useful because it contains a structured comparison model connected to the old model.

Second, LiGO explicitly treats depth and width growth as separate but coupled operators. This is consistent with our depth–width interpretation. In our theory, depth expansion creates new possible first-order improvement directions, but near the deepest-model regime these signals may become weak. Width then plays a compensating role by making weak directions more observable at finite sample size. LiGO's empirical success with combined depth and width growth provides external evidence that Transformer scaling benefits from structured architectural expansion rather than depth or width growth in isolation.

We emphasize the limitation of this comparison. LiGO is an empirical training efficiency method for Transformers, whereas our result is a conditional test-risk comparison theorem for normalized residual networks. Therefore, LiGO does not verify our probability bounds or our residual-block insertion theorem. It should instead be read as external consistency evidence for the broader principle underlying our framework: model expansion is most effective when the new model is connected to the old one through a structured bridge, and when depth, width, optimization, and data are scaled jointly.

P Condition 1 as an If-and-Only-If Criterion for First-Order Jumpboard Improvement

This appendix clarifies the precise mathematical content of Condition 1. In the main text, Condition 1 is used as a first-order non-degeneracy condition: after inserting a zero-output residual block at some layer l^*, there must exist a parameter direction of the inserted block that decreases the population risk to first order.

We show here that, under the differentiability assumptions already used in the paper, Condition 1 is not merely a sufficient condition. It is necessary and sufficient for the existence of a strictly negative first-order direction inside the tangent class generated by the inserted residual block at zero initialization.

This equivalence should be interpreted carefully. It is an if-and-only-if statement for the first-order jumpboard mechanism. It is not a necessary condition for every possible improvement in the enlarged hypothesis class: higher-order or non-local improvements may still exist even if the first-order condition fails.

P.1 Fixed-layer setup

Fix an admissible insertion layer l^*. Write the hidden representation at that layer as

$$z = x^{(l^*)} = f_{\mathrm{bot}}(x) \in \mathbb{R}^N.$$

The old trained model is written as

$$f^*_{\mathrm{old}}(x) = f_{\mathrm{top}}(z).$$

After inserting a residual block at layer l^*, the expanded model is parameterized by

$$f_\theta(x) = f_{\mathrm{top}}\big(z + h_\theta(z)\big),$$

where $h_\theta : \mathbb{R}^N \to \mathbb{R}^N$ is the inserted residual block.

The zero-output initialization condition is

$$h_0(z) = 0.$$

Therefore

$$f_0 = f^*_{\mathrm{old}}.$$

Define the population risk profile over the newly inserted block parameters by

$$\mathcal{R}_h(\theta) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\big(f_{\mathrm{top}}(z + h_\theta(z)), y\big)\right].$$

In particular,

$$\mathcal{R}_h(0) = R(f^*_{\text{old}}).$$

The activation-gradient signal at the insertion layer is

$$q(z, y) := \nabla_z \ell(f_{\text{top}}(z), y) \in \mathbb{R}^N.$$

Assume that h_θ is differentiable at $\theta = 0$. Define the parameter Jacobian of the inserted residual block at zero initialization by

$$J_\theta h_0(z) := \left.\frac{\partial h_\theta(z)}{\partial \theta}\right|_{\theta=0}.$$

If $\theta \in \mathbb{R}^p$, then

$$J_\theta h_0(z) \in \mathbb{R}^{N\times p}.$$

For a parameter direction $u \in \mathbb{R}^p$, define the corresponding first-order hidden-space perturbation by

$$\phi_u(z) := J_\theta h_0(z) u.$$

The first-order tangent class generated by the inserted residual block is

$$\mathcal{T}_h := \{z \mapsto J_\theta h_0(z) u : u \in \mathbb{R}^p\}.$$

Finally, define the first-order residual signal in direction u by

$$a_u := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z, y)^\top J_\theta h_0(z) u\right].$$

Equivalently,

$$a_u = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z, y)^\top \phi_u(z)\right].$$

P.2 Condition 1

For the fixed insertion layer l^*, the population version of Condition 1 can be written as follows:

$$\exists u \in \mathbb{R}^p \quad \text{such that} \quad a_u < 0. \tag{C1}$$

Thus Condition 1 says that the activation-gradient signal is not orthogonal to the tangent directions made available by the newly inserted residual block, and that one of these directions can be oriented so as to decrease risk to first order.

P.3 A directional derivative identity

We first record the basic derivative identity behind Condition 1.

Lemma 2 (Directional derivative identity)**.** *Under the pathwise differentiability and differentiation-under-expectation assumptions used in the main text, for every $u \in \mathbb{R}^p$,*

$$D\mathcal{R}_h(0)[u] = a_u.$$

Equivalently,

$$D\mathcal{R}_h(0)[u] = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z, y)^\top J_\theta h_0(z) u\right].$$

Proof. Consider the one-dimensional path

$$\theta(\eta) = \eta u.$$

Along this path,

$$\mathcal{R}_h(\eta u) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\big(f_{\text{top}}(z + h_{\eta u}(z)), y\big)\right].$$

Since h_θ is differentiable at $\theta = 0$,

$$h_{\eta u}(z) = h_0(z) + \eta J_\theta h_0(z) u + r_\eta(z),$$

where $r_\eta(z)$ is a first-order negligible remainder. Because $h_0(z) = 0$,

$$h_{\eta u}(z) = \eta J_\theta h_0(z) u + r_\eta(z).$$

By differentiability of the loss with respect to the hidden representation,

$$\ell\big(f_{\text{top}}(z + h_{\eta u}(z)), y\big) = \ell(f_{\text{top}}(z), y) + \eta q(z, y)^\top J_\theta h_0(z) u + o(\eta),$$

where the $o(\eta)$ term is controlled in the sense required by the paper's pathwise differentiability assumption.

Taking expectation and using differentiation under the expectation gives

$$\frac{d}{d\eta}\mathcal{R}_h(\eta u)\bigg|_{\eta=0} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z, y)^\top J_\theta h_0(z) u\right].$$

Thus

$$D\mathcal{R}_h(0)[u] = a_u.$$

□

P.4 Condition 1 is necessary and sufficient

Proposition 2 (Condition 1 is an if-and-only-if criterion)**.** *For the fixed insertion layer l^*, under the assumptions above, the following statements are equivalent.*

1. *The fixed-layer version of Condition 1 holds:*
$$\exists u \in \mathbb{R}^p \quad \textit{such that} \quad a_u < 0.$$
2. *The inserted residual block admits a strictly negative first-order population-risk direction at zero initialization:*
$$\exists u \in \mathbb{R}^p \quad \textit{such that} \quad D\mathcal{R}_h(0)[u] < 0.$$
3. *The population-risk gradient with respect to the newly inserted residual-block parameters is nonzero:*
$$\nabla_\theta \mathcal{R}_h(0) \neq 0.$$
4. *The activation-gradient functional is not orthogonal to the inserted block's tangent class:*
$$\exists \phi \in \mathcal{T}_h \quad \textit{such that} \quad \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z, y)^\top \phi(z)\right] \neq 0.$$

Moreover,

$$\nabla_\theta \mathcal{R}_h(0) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[J_\theta h_0(z)^\top q(z, y)\right].$$

Proof. By Lemma 2, for every $u \in \mathbb{R}^p$,

$$D\mathcal{R}_h(0)[u] = a_u.$$

Therefore

$$\exists u:\ a_u < 0 \quad \Longleftrightarrow \quad \exists u:\ D\mathcal{R}_h(0)[u] < 0.$$

This proves the equivalence between statements 1 and 2.

Next, the map

$$u \mapsto a_u$$

is a linear functional on $\mathbb{R}^p$. Thus the existence of a negative direction is equivalent to this linear functional being nonzero.

Indeed, if there exists u such that $a_u < 0$, then clearly $a_\cdot \not\equiv 0$. Conversely, suppose $a_\cdot \not\equiv 0$. Then there exists $u \in \mathbb{R}^p$ such that $a_u \neq 0$. If $a_u < 0$, we are done. If $a_u > 0$, choose $-u$. Since

$$a_{-u} = -a_u < 0,$$

a negative direction exists.

Hence

$$\exists u:\ a_u < 0 \quad \Longleftrightarrow \quad a. \not\equiv 0.$$

By Lemma 2,

$$a. \not\equiv 0 \quad \Longleftrightarrow \quad D\mathcal{R}_h(0) \not\equiv 0.$$

Since the inserted-block parameter space is $\mathbb{R}^p$, this is equivalent to

$$\nabla_\theta \mathcal{R}_h(0) \neq 0.$$

This proves the equivalence between statements 1, 2, and 3.

The expression for the gradient follows from the identity

$$D\mathcal{R}_h(0)[u] = \mathbb{E}\left[q(z,y)^\top J_\theta h_0(z) u\right] = \left\langle \mathbb{E}\left[J_\theta h_0(z)^\top q(z,y)\right], u \right\rangle.$$

Since this holds for every $u \in \mathbb{R}^p$, we obtain

$$\nabla_\theta \mathcal{R}_h(0) = \mathbb{E}\left[J_\theta h_0(z)^\top q(z,y)\right].$$

It remains to connect this to the tangent-class orthogonality statement. By definition,

$$\mathcal{T}_h = \{z \mapsto J_\theta h_0(z) u:\ u \in \mathbb{R}^p\}.$$

For every $\phi_u \in \mathcal{T}_h$,

$$\mathbb{E}\left[q(z,y)^\top \phi_u(z)\right] = \mathbb{E}\left[q(z,y)^\top J_\theta h_0(z) u\right] = a_u.$$

Therefore,

$$a. \equiv 0$$

if and only if

$$\mathbb{E}\left[q(z,y)^\top \phi(z)\right] = 0 \qquad \text{for every } \phi \in \mathcal{T}_h.$$

Equivalently,

$$a. \not\equiv 0$$

if and only if there exists $\phi \in \mathcal{T}_h$ such that

$$\mathbb{E}\left[q(z,y)^\top \phi(z)\right] \neq 0.$$

This proves statement 4 and completes the proof. □

P.5 Population jumpboard consequence

The previous proposition immediately recovers the population jumpboard existence argument.

Corollary 4 (Population jumpboard from Condition 1)**.** *If the equivalent conditions in Proposition 2 hold, then there exist a direction $v_{\mathrm{pop}} \in \mathbb{R}^p$ and a sufficiently small step size $\eta_{\mathrm{pop}} > 0$ such that the population jumpboard model*

$$\tilde{f}_{\mathrm{pop}}(x) := f_{\mathrm{top}}\left(f_{\mathrm{bot}}(x) + h_{\eta_{\mathrm{pop}} v_{\mathrm{pop}}}(f_{\mathrm{bot}}(x))\right)$$

satisfies

$$R(\tilde{f}_{\mathrm{pop}}) < R(f^*_{\mathrm{old}}).$$

Proof. By Proposition 2, there exists $v_{\mathrm{pop}} \in \mathbb{R}^p$ such that

$$D\mathcal{R}_h(0)[v_{\mathrm{pop}}] < 0.$$

Define the one-dimensional population risk profile

$$G_{\mathrm{pop}}(\eta) := \mathcal{R}_h(\eta v_{\mathrm{pop}}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\big(f_{\mathrm{top}}(z + h_{\eta v_{\mathrm{pop}}}(z)), y\big)\right].$$

Then

$$G'_{\mathrm{pop}}(0) = D\mathcal{R}_h(0)[v_{\mathrm{pop}}] < 0.$$

By differentiability at the origin, there exists a sufficiently small $\eta_{\mathrm{pop}} > 0$ such that

$$G_{\mathrm{pop}}(\eta_{\mathrm{pop}}) < G_{\mathrm{pop}}(0).$$

Since

$$G_{\mathrm{pop}}(0) = R(f^*_{\mathrm{old}}),$$

the corresponding model $\tilde{f}_{\mathrm{pop}}$ satisfies

$$R(\tilde{f}_{\mathrm{pop}}) < R(f^*_{\mathrm{old}}).$$

□

P.6 EMPIRICAL VERSION

The population condition above is an oracle-level statement, since the population expectation is generally unavailable. The corresponding empirical version is the one used to construct a training-set jumpboard.

Let

$$S_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{M}$$

be the training set. Write

$$z_i = x_i^{(l^*)} = f_{\text{bot}}(x_i),$$

and define

$$q_i := \nabla_{z_i} \ell(f_{\text{top}}(z_i), y_i).$$

Define the empirical risk profile over the inserted-block parameters by

$$\mathcal{R}_{h,S}(\theta) := \frac{1}{M} \sum_{i=1}^{M} \ell\big(f_{\text{top}}(z_i + h_\theta(z_i)), y_i\big).$$

Since $h_0 = 0$,

$$\mathcal{R}_{h,S}(0) = \mathcal{L}_{\text{train}}(f^*_{\text{old}}).$$

For $u \in \mathbb{R}^p$, define the empirical first-order residual signal

$$\widehat{a}_u := \frac{1}{M} \sum_{i=1}^{M} q_i^\top J_\theta h_0(z_i) u.$$

Proposition 3 (Empirical Condition 1 is an if-and-only-if criterion)**.** *For the fixed insertion layer l^*, the following statements are equivalent.*

1. *The empirical version of Condition 1 holds:*
$$\exists u \in \mathbb{R}^p \quad \textit{such that} \quad \widehat{a}_u < 0.$$
2. *The inserted residual block admits a strictly negative first-order empirical-risk direction at zero initialization:*
$$\exists u \in \mathbb{R}^p \quad \textit{such that} \quad D\mathcal{R}_{h,S}(0)[u] < 0.$$
3. *The empirical-risk gradient with respect to the newly inserted residual-block parameters is nonzero:*
$$\nabla_\theta \mathcal{R}_{h,S}(0) \neq 0.$$

Moreover,

$$\nabla_\theta \mathcal{R}_{h,S}(0) = \frac{1}{M} \sum_{i=1}^{M} J_\theta h_0(z_i)^\top q_i.$$

Proof. The proof is the finite-sample analogue of Proposition 2. For every $u \in \mathbb{R}^p$, the chain rule gives

$$D\mathcal{R}_{h,S}(0)[u] = \frac{1}{M} \sum_{i=1}^{M} q_i^\top J_\theta h_0(z_i) u = \widehat{a}_u.$$

Thus

$$\exists u:\ \widehat{a}_u < 0 \quad \Longleftrightarrow \quad \exists u:\ D\mathcal{R}_{h,S}(0)[u] < 0.$$

Since $u \mapsto \widehat{a}_u$ is linear and the parameter space is closed under sign reversal, the existence of a negative direction is equivalent to $\widehat{a}_\cdot \not\equiv 0$. This is equivalent to

$$\nabla_\theta \mathcal{R}_{h,S}(0) \neq 0.$$

The displayed gradient formula follows by writing the directional derivative as an inner product with u:

$$D\mathcal{R}_{h,S}(0)[u] = \left\langle \frac{1}{M} \sum_{i=1}^{M} J_\theta h_0(z_i)^\top q_i, u \right\rangle.$$

□

Corollary 5 (Training-set jumpboard from empirical Condition 1)**.** *If the equivalent conditions in Proposition 3 hold, then there exist a direction $v_S \in \mathbb{R}^p$ and a sufficiently small step size $\eta_S > 0$ such that the training-set jumpboard model*

$$\tilde{f}_S(x) := f_{\text{top}}\left(f_{\text{bot}}(x) + h_{\eta_S v_S}(f_{\text{bot}}(x))\right)$$

satisfies

$$\mathcal{L}_{\text{train}}(\tilde{f}_S) < \mathcal{L}_{\text{train}}(f^*_{\text{old}}).$$

Proof. By Proposition 3, there exists $v_S \in \mathbb{R}^p$ such that

$$D\mathcal{R}_{h,S}(0)[v_S] < 0.$$

Define

$$G_S(\eta) := \mathcal{R}_{h,S}(\eta v_S).$$

Then

$$G'_S(0) < 0.$$

Hence, by differentiability at the origin, there exists a sufficiently small $\eta_S > 0$ such that

$$G_S(\eta_S) < G_S(0).$$

Since

$$G_S(0) = \mathcal{L}_{\text{train}}(f^*_{\text{old}}),$$

the corresponding $\tilde{f}_S$ satisfies

$$\mathcal{L}_{\text{train}}(\tilde{f}_S) < \mathcal{L}_{\text{train}}(f^*_{\text{old}}).$$

□

P.7 A STANDARD ZERO-OUTPUT RESIDUAL BLOCK

We now specialize the criterion to a common residual-block parameterization. Suppose the inserted residual block has the form

$$h_{A,\xi}(z) = A\psi_\xi(z),$$

where $\psi_\xi(z) \in \mathbb{R}^r$ is the residual feature map and $A \in \mathbb{R}^{N\times r}$ is the final output projection.

Zero-output initialization is obtained by setting

$$A = 0.$$

The internal feature parameters ξ may be fixed or initialized independently, as long as ψ_ξ is well defined.

For a perturbation direction $B \in \mathbb{R}^{N\times r}$ of the output projection A,

$$J_A h_0(z)[B] = B\psi_\xi(z).$$

Therefore the population first-order signal in direction B is

$$a_B = \mathbb{E}\left[q(z,y)^\top B\psi_\xi(z)\right].$$

Define

$$M_{\text{pop}} := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(z,y)\psi_\xi(z)^\top\right] \in \mathbb{R}^{N\times r}.$$

Using the Frobenius inner product,

$$a_B = \langle B, M_{\text{pop}}\rangle_F.$$

Corollary 6 (Condition 1 for a zero-output projection block)**.** *For the block*

$$h_{A,\xi}(z) = A\psi_\xi(z)$$

with zero-output initialization $A = 0$, the fixed-layer population version of Condition 1 holds if and only if

$$M_{\text{pop}} = \mathbb{E}\left[q(z,y)\psi_\xi(z)^\top\right] \neq 0.$$

Moreover, when $M_{\text{pop}} \neq 0$, choosing

$$B = -M_{\text{pop}}$$

gives

$$a_B = -\|M_{\text{pop}}\|_F^2 < 0.$$

Proof. If $M_{\mathrm{pop}} = 0$, then for every B,

$$a_B = \langle B, M_{\mathrm{pop}} \rangle_F = 0.$$

Hence no first-order descent direction exists in the output-projection tangent space.

Conversely, suppose $M_{\mathrm{pop}} \neq 0$. Choose

$$B = -M_{\mathrm{pop}}.$$

Then

$$a_B = \langle -M_{\mathrm{pop}}, M_{\mathrm{pop}} \rangle_F = -\|M_{\mathrm{pop}}\|_F^2 < 0.$$

Therefore Condition 1 holds. □

The empirical analogue is obtained by replacing M_{pop} with

$$\widehat{M}_S := \frac{1}{M} \sum_{i=1}^{M} q_i \psi_\xi(z_i)^\top.$$

Thus, for the zero-output projection block, empirical Condition 1 holds if and only if

$$\widehat{M}_S \neq 0.$$

When $\widehat{M}_S \neq 0$, choosing $B = -\widehat{M}_S$ gives a strictly negative empirical first-order direction.

P.8 Relation to stationarity of the old model

The equivalence above also explains why a trained old model may be stationary with respect to its original parameters while still admitting a descent direction after depth expansion.

Let $\mathcal{T}_{\mathrm{old}}$ denote the tangent class generated by the original parameters of the old model at f^*_{old}. First-order stationarity of the old model with respect to its original parameters means

$$\mathbb{E}\left[q(z,y)^\top \phi_{\mathrm{old}}(z)\right] = 0 \qquad \text{for every } \phi_{\mathrm{old}} \in \mathcal{T}_{\mathrm{old}}.$$

Equivalently,

$$q \perp \mathcal{T}_{\mathrm{old}}$$

in the population inner product induced by $\mathcal{D}$.

However, after inserting a residual block, the available tangent directions include the new tangent class

$$\mathcal{T}_h = \{z \mapsto J_\theta h_0(z) u : \ u \in \mathbb{R}^p\}.$$

Stationarity with respect to $\mathcal{T}_{\mathrm{old}}$ does not imply stationarity with respect to $\mathcal{T}_h$. In general, it is possible that

$$q \perp \mathcal{T}_{\mathrm{old}}$$

while

$$q \not\perp \mathcal{T}_h.$$

In this case, the old model is stationary in the old parameterization, but the expanded model is not stationary at the embedded point $\theta = 0$. This is precisely the regime captured by Condition 1.

In parameter-gradient form, this says

$$\nabla_{\theta_{\mathrm{old}}} R(f^*_{\mathrm{old}}) = 0$$

does not imply

$$\nabla_{\theta_{\mathrm{new}}} \mathcal{R}_h(0) = 0.$$

The latter gradient is

$$\nabla_{\theta_{\mathrm{new}}} \mathcal{R}_h(0) = \mathbb{E}\left[J_\theta h_0(z)^\top q(z,y)\right],$$

and Condition 1 is equivalent to this vector being nonzero.

P.9 LAYERWISE AND GLOBAL VERSIONS

The results above were stated for a fixed insertion layer l^*. The global version of Condition 1 used in the main text is obtained by allowing the insertion layer to vary.

For each admissible layer l, let

$$\mathcal{T}_h^{(l)}$$

be the tangent class generated by inserting the residual block at layer l, and let

$$q^{(l)}(z^{(l)}, y) := \nabla_{z^{(l)}} \ell(f_{\text{top}}^{(l)}(z^{(l)}), y)$$

be the corresponding activation-gradient signal.

Then the global population version of Condition 1 is equivalent to

$$\exists l \quad \text{such that} \quad q^{(l)} \not\perp \mathcal{T}_h^{(l)}.$$

Equivalently, there exists an admissible layer l and an inserted-block parameter direction u such that

$$\mathbb{E}\left[q^{(l)}(z^{(l)}, y)^\top J_\theta h_0^{(l)}(z^{(l)})u\right] < 0.$$

Thus the global condition is simply the existence of at least one layer at which the newly inserted residual block opens a nonzero first-order risk direction.

P.10 BOUNDARY OF THE EQUIVALENCE

The equivalence in Proposition 2 is a first-order equivalence. It says that Condition 1 is necessary and sufficient for the inserted residual block to produce a strictly negative first-order direction at zero-output initialization.

It does not say that Condition 1 is necessary for arbitrary improvements in the enlarged hypothesis class. If Condition 1 fails, then

$$D\mathcal{R}_h(0)[u] = 0 \qquad \text{for every } u \in \mathbb{R}^p.$$

Thus the first-order jumpboard mechanism is exhausted at that insertion point. However, higher-order or non-local improvements may still exist.

For example, a one-dimensional risk profile of the form

$$R(\theta) = R(0) - \theta^4$$

has

$$R'(0) = 0,$$

so no first-order descent direction exists at $\theta = 0$. Nevertheless,

$$R(\theta) < R(0)$$

for every sufficiently small nonzero θ.

Therefore the precise interpretation is

Condition 1 is necessary and sufficient for first-order jumpboard improvement, not for all possible higher-order or global impr