Generalization of Dynamic Neural Architectures Scales with Intrinsic Dimension: A PAC-Bayes Theory

Anonymous Authors[1]

Abstract

We study the generalization of neural networks whose architectures are modified at training time via pruning, quantization, or width expansion. Extending the PAC-Bayes framework of MacAulay et al. (2023) to this dynamic setting, we prove that the generalization error scales with the *intrinsic dimension* of the Fisher information at the final iterate—not the ambient parameter count. Our central result (Theorem 3.4) provides a $\widetilde{O}\big(\sqrt{d_{\text{int}}/n}\big)$ upper bound, which is near-tight for a linear model class; the gap for deep neural networks remains open. A matching minimax lower bound (Theorem 3.5) confirms that this dependence is unavoidable. We further prove that Fisher-score pruning provably reduces the intrinsic dimension under spectral-gap conditions (Theorem 4.1), yielding a conditional end-to-end improvement that requires Assumptions (C1)–(C3) and a spectral gap. We provide a theoretical analysis identifying the conditions under which the theory predicts (or fails to predict) generalization improvement.

1. Introduction

Modern neural networks routinely undergo structural changes during training: pruning removes redundant weights (Han et al., 2015; Frankle & Carbin, 2019), width expansion adapts capacity (Evci et al., 2020; 2022), and quantization compresses representations. These *dynamic architectures* challenge standard generalization theory because the model class evolves over the course of optimization, and classical complexity measures (VC dimension, Rademacher complexity) can grow or shrink in ways that are difficult to track.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

The PAC-Bayes framework (McAllester, 1998; Catoni, 2007) has emerged as a powerful tool for non-vacuous generalization bounds in deep learning (Dziugaite & Roy, 2017; Neyshabur et al., 2018). In particular, MacAulay et al. (2023) established that for *static* neural networks, the generalization error depends on an *intrinsic dimension* defined via the eigenspectrum of the empirical Fisher information, rather than on the full parameter count p. This resolved a long-standing puzzle: the empirical success of overparameterized networks with millions of parameters trained on far fewer examples.

Gap and contribution. The result of MacAulay et al. (2023) applies only to architectures that remain fixed throughout training. Yet in practice, most deployed networks are pruned, expanded, or otherwise modified during the optimization process. This paper extends their framework to *dynamic architectures*, providing a formal proof connecting pruning to intrinsic-dimension reduction under spectral conditions.

Our main contributions are:

1. **Dynamic PAC-Bayes bounds.** We derive three progressively sharper PAC-Bayes inequalities for dynamic architectures (Theorems 3.1–3.4), culminating in a bound that scales as $\widetilde{O}\big(\sqrt{d_{\text{int}}/n}\big)$ where d_{int} is the intrinsic dimension of the Fisher information at the *final* iterate.
2. **Matching lower bound.** We prove via the Assouad lemma that the $\sqrt{d_{\text{int}}/n}$ dependence is minimax optimal (Theorem 3.5).
3. **Pruning reduces intrinsic dimension.** Under a spectral-gap condition and stability assumptions (C1)–(C3), we prove that Fisher-score pruning provably reduces d_{int} (Theorem 4.1).
4. **Theoretical failure analysis.** We analyze settings where the theory makes no prediction, including expansion–contraction cycles that may not reduce d_{int} due to parameter-space geometry, and spectra with no clear spectral gap.

Scope of this work. This is a purely theoretical paper: all

results are formal mathematical theorems and proofs. No experimental claims are made. The results are conditional on the stated assumptions (Section 3); they do not establish a general "link" between pruning and generalization for arbitrary architectures.

Related work. PAC-Bayes bounds for neural networks have been studied extensively (Neyshabur et al., 2018; Dziugaite & Roy, 2017; Pensia et al., 2020; Lotfi et al., 2022; Rivasplata et al., 2024). The intrinsic-dimension perspective originates in the compression-based framework of MacAulay et al. (2023) and connects to the spectral complexity literature (Bartlett et al., 2017; Majewski et al., 2018). Pruning theory includes the lottery ticket hypothesis (Frankle & Carbin, 2019; Malach et al., 2020; Zhou et al., 2019) and optimal brain surgeon/pruning (LeCun et al., 1990; Hassibi et al., 1993; Singh et al., 2020). Function-preserving growth (Chen et al., 2016) and neural architecture descent (Wu et al., 2020) provide alternative perspectives on dynamic architectures. Our work bridges these strands by showing that *pruning reduces intrinsic dimension*, which in turn tightens PAC-Bayes generalization bounds—conditional on spectral assumptions.

2. Preliminaries

2.1. Setting and Notation

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be the input–output space, $\mathcal{D}$ an unknown distribution over $\mathcal{Z}$, and $S = \{z_i\}_{i=1}^n$ an i.i.d. sample from $\mathcal{D}$. A neural network with parameters $\theta \in \mathbb{R}^p$ defines a predictor $h_\theta : \mathcal{X} \to \mathcal{Y}$. The loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, B]$ is bounded by $B > 0$.

A *dynamic architecture* procedure is a sequence

$$\theta_0 \xrightarrow{\text{train}} \theta_1 \xrightarrow{\text{prune/expand}} \theta_1' \xrightarrow{\text{train}} \theta_2 \to \cdots \to \theta_T, \qquad (1)$$

where each step may change the parameter dimension and the mapping h_θ. We denote the final architecture by $\mathcal{A}_T$ with parameters in $\mathbb{R}^{p_T}$.

2.2. Key Definitions

Definition 2.1 (Empirical Fisher Information)**.** Given a trained network with parameters $\theta \in \mathbb{R}^p$ and data S, the *empirical Fisher information matrix* is

$$\hat{F}(\theta) \;=\; \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell(h_\theta(x_i), y_i)\, \nabla_\theta \ell(h_\theta(x_i), y_i)^\top. \quad (2)$$

Definition 2.2 (Intrinsic Dimension)**.** For $\varepsilon > 0$, the *ε-intrinsic dimension* is

$$d_{\text{int}}(\varepsilon) \;=\; \big|\{j \in [p] : \lambda_j(\hat{F}(\theta)) > \varepsilon\}\big|, \qquad (3)$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are the eigenvalues of $\hat{F}(\theta)$.

Definition 2.3 (Fisher-Score Pruning)**.** Given a target sparsity $s \in (0, 1)$, *Fisher-score pruning* removes the $(1 - s)p$ parameters with smallest diagonal entries of $\hat{F}(\theta)$:

$$\mathcal{S} \;=\; \text{top}{-}s\big\{j : [\hat{F}(\theta)]_{jj}\big\}. \qquad (4)$$

Definition 2.4 (Compression Operator)**.** A mapping $\mathcal{C} : \mathbb{R}^p \to \mathbb{R}^k$ with $k < p$ is a *compression operator* if it is deterministic and computable from the data S and training history. The *compression size* is k.

2.3. Assumptions

We state all assumptions explicitly to make the conditional nature of our results transparent.

Assumption 2.5 (Bounded Loss)**.** $\ell(h_\theta(x), y) \in [0, B]$ for all θ, x, y, with $B \geq 1$ known.

Assumption 2.6 (Lipschitz Gradients)**.** For the final architecture $\mathcal{A}_T$, the loss gradients are L_∇-Lipschitz: $\|\nabla_\theta \ell(h_\theta(x), y) - \nabla_{\theta'} \ell(h_{\theta'}(x), y)\| \leq L_\nabla \|\theta - \theta'\|$.

Assumption 2.7 (Perturbation Stability (C1))**.** For a pruning step producing architecture $\mathcal{A}'$ from $\mathcal{A}$ with parameter mapping $\theta \mapsto \theta' = R\theta + \delta$, the perturbation satisfies $\|\delta\|_2 \leq \eta_p$ for known $\eta_p > 0$.

Assumption 2.8 (Submatrix Stability (C2))**.** The projection matrix R corresponding to the pruning step satisfies $\|R\|_{\text{op}} \leq 1$ (no amplification).

Assumption 2.9 (Fisher Stability (C3))**.** There exists $L_F > 0$ such that for adjacent architectures $\mathcal{A}, \mathcal{A}'$ with parameter mappings as in (C1): $\|\hat{F}_{\mathcal{A}'}(\theta') - R\,\hat{F}_{\mathcal{A}}(\theta)\,R^\top\|_{\text{op}} \leq L_F \eta_p$.

Assumption 2.10 (Spectral Gap)**.** The eigenvalues of $\hat{F}(\theta)$ at the final iterate satisfy a gap $\gamma > 0$: $\lambda_{d_{\text{int}}} > \varepsilon + \gamma$ and $\lambda_{d_{\text{int}}+1} < \varepsilon - \gamma$ for the chosen threshold ε.

Remark 2.11 (Discussion of Assumptions). Assumptions (C1)–(C3) formalize the intuition that pruning should not catastrophically alter the loss landscape. They hold for single-step pruning with small perturbation η_p. The spectral-gap assumption (Assumption 2.10) is strong: it requires a clear separation between "important" and "unimportant" Fisher directions. When this gap is small, our pruning result (Theorem 4.1) may not guarantee reduction in d_{int}.

3. PAC-Bayes Bounds for Dynamic Architectures

This section develops our main PAC-Bayes bounds. We build from a basic dynamic inequality (Theorem 3.1) to a compressed KL version (Theorem 3.3) and finally to the intrinsic-dimension bound (Theorem 3.4).

3.1. Dynamic PAC-Bayes Inequality

Theorem 3.1 (Dynamic PAC-Bayes Inequality)**.** *Let Assumption 2.5 hold. For any prior π over $\mathbb{R}^{p_T}$ and any*

data-dependent posterior ρ, with probability at least $1-\delta$ over S:

$$\mathbb{E}_{\theta\sim\rho}\big[\hat{R}(\theta)\big] \;\le\; \mathbb{E}_{\theta\sim\rho}\big[\hat{R}_S(\theta)\big] + \sqrt{\frac{B^2}{2n}\Big(\mathrm{KL}(\rho\|\pi) + \log\frac{n+1}{\delta}\Big)}. \tag{5}$$

Proof sketch. The proof adapts the standard PAC-Bayes argument of McAllester (1998) to the dynamic setting. The key observation is that the PAC-Bayes theorem holds for *any* data-dependent posterior, regardless of how the model class was constructed—the architecture history does not enter the PAC-Bayes inequality directly. One applies the change-of-measure inequality with ρ, π on the *final* parameter space $\mathbb{R}^{p_T}$, using that the loss is bounded by B. The Donsker–Varadhan variational formula yields the KL term, and a union bound over n ghost-sample terms produces the $\log(n+1)$ factor. □

Remark 3.2. The bound in Theorem 3.1 is *identical in form* to the standard PAC-Bayes bound for static architectures. The challenge lies in constructing a posterior ρ whose KL divergence $\mathrm{KL}(\rho\|\pi)$ reflects the effective complexity of the *dynamic* procedure, not the ambient dimension p_T. This is addressed in Theorems 3.3 and 3.4.

3.2. Compressed KL Bound

The following result shows that when the dynamic procedure involves compression, the KL cost scales with the compressed dimension rather than the ambient one.

Theorem 3.3 (Compressed KL Bound)**.** *Suppose the dynamic procedure produces a final parameter $\theta_T \in \mathbb{R}^{p_T}$ via a compression $\mathcal{C} : \mathbb{R}^{p_T} \to \mathbb{R}^k$ with $k < p_T$ (Definition 2.4). Let $\pi = \mathcal{N}(0, B^2 I_{p_T})$ be the prior and define the* compressed posterior ρ_{comp} *as the pushforward of a density on $\mathbb{R}^k$ to $\mathbb{R}^{p_T}$ via the compression operator. Then, with probability at least $1-\delta$:*

$$\mathbb{E}\big[\hat{R}(\theta_T)\big] \;\le\; \hat{R}_S(\theta_T) + \frac{C}{\sqrt{n}}\sqrt{k\log\frac{p_T B^2}{\delta} + \log\frac{n+1}{\delta}}, \tag{6}$$

where $C = B\sqrt{2}$ is an explicit constant.

Proof sketch. Decompose $\mathrm{KL}(\rho_{\mathrm{comp}}\|\pi)$ using the chain rule for KL divergence. The compressed component lives in $\mathbb{R}^k$, contributing $k\log(p_T B^2/\delta)$ terms. The residual (zeroed-out) coordinates contribute at most $O(\log(p_T/\delta))$ due to the discrete nature of the support selection. Combining with Theorem 3.1 yields the result. The full proof is given in Appendix A. □

3.3. Intrinsic-Dimension PAC-Bayes Bound

Our central result replaces the compressed dimension k with the intrinsic dimension d_{int}, which can be much smaller.

Theorem 3.4 (Intrinsic-Dimension PAC-Bayes Bound)**.** *Let Assumptions 2.5–2.6 hold. Fix $\varepsilon > 0$ and let $D = \{j : \lambda_j(\hat{F}(\theta_T)) > \varepsilon\}$ with $|D| = d_{\mathrm{int}}$. Define $B \ge 1$ as the loss bound and $\sigma^2 = B^2\varepsilon$. Let $\pi = \mathcal{N}(\theta_0, B^2 I_{p_T})$ and let ρ_z be the* subspace posterior*:*

$$\rho_z \;=\; \mathcal{N}\big(z^*,\, \sigma^2\Lambda_D^{-1}\big), \tag{7}$$

where z^* is the projection of θ_T onto the top-d_{int} eigenspace and $\Lambda_D = \mathrm{diag}(\lambda_j : j \in D)$. Then, with probability at least $1-\delta$ over the data split $S = S_1 \cup S_2$ (where $|S_1| = m = \lfloor n/2 \rfloor$ is used to estimate $\hat{F}$ and $|S_2| = n_2 \ge n/3$ is used for the PAC-Bayes bound):

$$\mathbb{E}_{\theta\sim\rho_z}\big[\hat{R}(\theta)\big] \;\le\; \mathbb{E}_{\theta\sim\rho_z}\big[\hat{R}_{S_2}(\theta)\big] + \underbrace{\sqrt{3}\,B\sqrt{\frac{\varepsilon\, d_{\mathrm{int}}}{2\,n_2}}}_{\text{dominant variance term}} + C\sqrt{\frac{\log(n+1) + \log(1/\delta)}{n_2}} \tag{8}$$

where $C > 0$ is an explicit constant (see proof), and the $\sqrt{3}$ factor arises from the data splitting $n_2 \ge n/3$. Moreover, the bound satisfies a monotonicity property*: if pruning reduces d_{int}, the bound tightens.*

Proof sketch. We provide a 6-step proof with **all constants tracked explicitly**. The full proof appears in Appendix B.

Step 1: Data splitting. Split $S = S_1 \cup S_2$ with $|S_1| = m = \lfloor n/2 \rfloor$ and $|S_2| = n_2 \ge n/3$ (since $\lfloor n/2 \rfloor + \lceil n/2 \rceil = n$, and $n_2 = \lceil n/2 \rceil \ge n/3$ for all $n \ge 1$). Use S_1 to compute $\hat{F}$ and S_2 for the PAC-Bayes bound. The prior is $\pi = \mathcal{N}(\theta_0, B^2 I_{p_T})$.

Step 2: Subspace posterior. Let $\hat{F} = U\Lambda U^\top$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{p_T})$. Set $D = \{j : \lambda_j > \varepsilon\}$, $|D| = d_{\mathrm{int}}$. Define the subspace posterior $\rho_z = \mathcal{N}(z^*, \sigma^2\Lambda_D^{-1})$ where $\sigma^2 = B^2\varepsilon$ and $z^* = U_D U_D^\top \theta_T$ is the projection of θ_T onto the top-d_{int} eigenspace.

Step 3: KL computation. Computing the KL divergence between the subspace posterior (restricted to the D coordinates) and the prior (restricted to the same coordinates) yields:

$$\mathrm{KL}(\rho_z\|\pi) \;\le\; \frac{d_{\mathrm{int}}}{2}\log\frac{B^2}{\varepsilon} \;+\; 2. \tag{9}$$

The "+2" arises from the trace term $\mathrm{tr}(\Sigma^{-1}\Sigma_0) = \sum_{j\in D}\sigma_j^2/(B^2) = d_{\mathrm{int}}\cdot\varepsilon/1$ and the $\|\mu-\mu_0\|^2$ term after projecting. We track this $+2$ explicitly rather than absorbing it into a constant.

Step 4: Risk of subspace posterior. The risk under ρ_z decomposes into the subspace contribution and the tail:

$$\mathrm{Var}_{\rho_z}[\ell] \;\le\; \underbrace{\sigma^2 B^2 d_{\mathrm{int}}}_{=\varepsilon B^4 d_{\mathrm{int}}\ \text{(dominant)}} \;+\; \underbrace{\sigma^2 B^2}_{=\varepsilon B^4\ \text{(tail)}}. \tag{10}$$

Step 5: PAC-Bayes on S_2. Applying Theorem 3.1 on S_2 of size n_2:

$$\mathbb{E}[\hat{R}] - \mathbb{E}[\hat{R}_{S_2}] \;\le\; \sqrt{\frac{B^2}{2\,n_2}\Big(\frac{d_{\text{int}}}{2}\log\frac{B^2}{\varepsilon} + 2 + \log\frac{n+1}{\delta}\Big)}. \tag{11}$$

Since $n_2 \ge n/3$, converting to the full sample size introduces an explicit $\sqrt{3}$ factor:

$$\le\; \sqrt{3}\,\sqrt{\frac{B^2}{2\,n}\Big(\frac{d_{\text{int}}}{2}\log\frac{B^2}{\varepsilon} + 2 + \log\frac{n+1}{\delta}\Big)}. \tag{12}$$

Step 6: Monotonicity. The leading term is $\widetilde{O}\big(\sqrt{d_{\text{int}}/n}\big)$. Crucially, d_{int} appears *additively* in the bound (via the KL term), so any reduction in d_{int} directly tightens the bound. □

3.4. Matching Lower Bound

Theorem 3.5 (Minimax Lower Bound via Assouad)**.** *Consider the class of p-dimensional linear regression models $\mathcal{F} = \{x \mapsto \langle \beta, x\rangle : \beta \in \mathbb{R}^p, \|\beta\|_0 \le k\}$ with $k \le p$. Assume the Fisher information matrix is the identity I_p (standard Gaussian design). Then, for any estimator $\hat{\beta}$:*

$$\inf_{\hat{\beta}} \sup_{\beta\in\mathcal{F}} \mathbb{E}\big[\|\hat{\beta} - \beta\|^2\big] \;\ge\; \frac{k}{4n}\cdot\frac{1}{1+o(1)}. \tag{13}$$

Equivalently, the excess risk satisfies:

$$\inf_{\hat{\beta}} \sup_{\beta\in\mathcal{F}} \big(\hat{R}(\hat{\beta}) - \hat{R}^*\big) \;\ge\; c\sqrt{\frac{k}{n}} \tag{14}$$

for an absolute constant $c > 0$, under bounded response.

Proof sketch. We use the Assouad lemma (Tsybakov, 2009). Construct a k-dimensional hypercube indexed by $\omega \in \{0,1\}^k$. For each ω, define β_ω with coordinates $\beta_{\omega,j} = (2\omega_j - 1)\Delta$ for $j \le k$ and $\beta_{\omega,j} = 0$ for $j > k$. The KL divergence between adjacent hypotheses (differing in one coordinate) is $\mathrm{KL}(P_\omega \| P_{\omega\oplus e_j}) = n\Delta^2/k$ (since Fisher $= I_k$ and the change is in one coordinate of magnitude 2Δ). Setting $\Delta^2 = k/(4n)$ yields pairwise KL $= 1/4$ and total variation $\le 1/2$. The Assouad lemma then gives minimax risk $\ge k\Delta^2/4 = k/(16n)$. Translating to excess risk and accounting for the bounded loss yields the $\sqrt{k/n}$ rate. The full proof is in Appendix C. □

Remark 3.6. The upper bound of Theorem 3.4 and the lower bound of Theorem 3.5 match in rate ($\sqrt{d_{\text{int}}/n}$) for the linear model class. For deep neural networks, the gap in constants and logarithmic factors remains open. In particular, the Fisher information of a trained network is *not* the identity, and the spectral properties of $\hat{F}$ can cause the effective intrinsic dimension to differ from k.

3.5. Optimal Threshold Selection and Non-Vacuity

Proposition 3.7 (Optimal Threshold)**.** *Let $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p$ be the eigenvalues of $\hat{F}(\theta_T)$. The threshold ε^* that minimizes the bound of Theorem 3.4 satisfies:*

$$\varepsilon^* \;\in\; \arg\min_{\varepsilon>0}\left\{\sqrt{\varepsilon\, d_{\text{int}}(\varepsilon)} + \sqrt{\frac{d_{\text{int}}(\varepsilon)}{n}\log\frac{B^2}{\varepsilon}}\right\}. \tag{15}$$

For the common "elbow" spectrum where $\lambda_j \asymp j^{-\alpha}$ with $\alpha > 1$:

$$\varepsilon^* \asymp n^{-\frac{2\alpha}{2\alpha+1}}, \qquad d^*_{\text{int}} \asymp n^{\frac{1}{2\alpha+1}}. \tag{16}$$

Proposition 3.8 (Non-Vacuity)**.** *The bound of Theorem 3.4 is non-vacuous (i.e., strictly less than B) whenever:*

$$n \;\gtrsim\; d_{\text{int}}\cdot\Big(\log\frac{B^2}{\varepsilon} + \log\frac{1}{\delta}\Big). \tag{17}$$

For typical values $B = 10$, $\varepsilon = 10^{-4}$, $\delta = 0.05$, this requires $n \gtrsim 30\, d_{\text{int}}$.

4. Intrinsic Dimension Reduction via Fisher-Score Pruning

We now prove the key structural result: under appropriate conditions, Fisher-score pruning (Definition 2.3) reduces the intrinsic dimension d_{int}.

Theorem 4.1 (Pruning Reduces Intrinsic Dimension)**.** *Let $\mathcal{A}$ be a neural architecture with parameters $\theta \in \mathbb{R}^p$ and Fisher matrix $\hat{F} = U\Lambda U^\top$ with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_p$. Let $\mathcal{A}'$ be the architecture after Fisher-score pruning with sparsity s and subsequent retraining, producing parameters θ' with Fisher $\hat{F}'$. Assume:*

(C1) *The pruning perturbation satisfies $\|\theta' - R\theta\|_2 \le \eta_p$ (Assumption 2.7),*

(C2) *$\|R\|_{\text{op}} \le 1$ (Assumption 2.8),*

(C3) *$\|\hat{F}'(\theta') - R\,\hat{F}(\theta)\,R^\top\|_{\text{op}} \le L_F\eta_p$ (Assumption 2.9).*

If the spectral gap γ (Assumption 2.10) satisfies:

$$\gamma \;>\; L_F\eta_p + \beta\,\varepsilon, \tag{18}$$

where $\beta \ge 0$ is a constant depending on the eigenvalue decay profile, then:

$$d'_{\text{int}}(\varepsilon) \;\le\; d_{\text{int}}(\varepsilon). \tag{19}$$

Furthermore, if the pruned directions have Fisher scores $[\hat{F}]_{jj} < \varepsilon - L_F\eta_p$ for all removed parameters j, then $\beta = 0$ and:

$$\big\|\Pi_\varepsilon\big(R\,\Pi_\varepsilon(\hat{F})\,R^\top\big)\big\|_{\text{op}} \;\ge\; \|\Pi_\varepsilon(\hat{F})\|_{\text{op}} - L_F\eta_p. \tag{20}$$

Proof sketch (3 steps). The full proof is in Appendix D.

Step 1: Perturbation of the Fisher matrix. After pruning with projection R and retraining to θ', the new Fisher satisfies:

$$\hat{F}'(\theta') \;=\; R\,\hat{F}(\theta)\,R^\top + E, \qquad \|E\|_{\mathrm{op}} \le L_F \eta_p, \quad (21)$$

by Assumption (C3).

Step 2: Spectral perturbation analysis. By the Cauchy interlacing theorem applied to $R\,\hat{F}\,R^\top$ (a compression of $\hat{F}$), each eigenvalue of $R\,\hat{F}\,R^\top$ is bounded between adjacent eigenvalues of $\hat{F}$. Applying Weyl's perturbation theorem:

$$|\lambda'_j - \lambda_j^{\mathrm{compressed}}| \;\le\; L_F \eta_p, \quad (22)$$

where $\lambda_j^{\mathrm{compressed}}$ are eigenvalues of $R\,\hat{F}\,R^\top$.

Step 3: Counting above threshold. The number of eigenvalues of $\hat{F}'$ exceeding ε is bounded by combining Steps 1–2 with the spectral gap. For any direction j with $\lambda_j(\hat{F}) \le \varepsilon - \gamma$:

$$\lambda'_j \;\le\; \lambda_j^{\mathrm{compressed}} + L_F\eta_p \;\le\; (\varepsilon - \gamma) + L_F\eta_p \;<\; \varepsilon, \quad (23)$$

since $\gamma > L_F\eta_p$. Therefore, at most $d_{\mathrm{int}}(\varepsilon)$ eigenvalues of $\hat{F}'$ can exceed ε, giving $d'_{\mathrm{int}}(\varepsilon) \le d_{\mathrm{int}}(\varepsilon)$. □

Remark 4.2. Theorem 4.1 provides a *non-increase* guarantee on d_{int}, not an exact reduction amount. The quantity $\|P_S \Pi_\varepsilon\|_F$ (where P_S projects onto the surviving subspace) is a *lower bound* on the number of high-Fisher directions preserved after pruning, not the exact count. Tighter bounds require stronger assumptions on the eigenstructure.

Proposition 4.3 (OBS Pruning Satisfies (C3))**.** *Optimal Brain Surgeon (OBS) pruning (Hassibi et al., 1993), which removes parameter $j^* = \arg\min_j [\hat{F}^{-1}]_{jj}$, satisfies Assumption (C3) with $L_F = O(L_\nabla^2 \eta_p)$ under Assumption 2.6, provided the retraining step after pruning is a single gradient descent step with step size η_p.*

Corollary 4.4 (Conditional End-to-End Improvement)**.** *Under the conditions of Theorems 3.4 and 4.1, if Fisher-score pruning with retraining reduces the intrinsic dimension from d_{int} to $d'_{\mathrm{int}} < d_{\mathrm{int}}$, then the generalization bound of Theorem 3.4 tightens by a factor of at least $\sqrt{d'_{\mathrm{int}}/d_{\mathrm{int}}}$ in the leading term. This improvement is* conditional *on the spectral gap and stability assumptions holding.*

5. Theoretical Comparison and Analysis

In this section we place our results in the context of the PAC-Bayes literature and analyze the theoretical conditions under which our bounds are informative.

Table 1. Comparison of PAC-Bayes generalization bounds. We report the rate dependence on n and the relevant complexity measure. "Conditional" indicates whether additional assumptions beyond bounded loss are required.

BOUND	RATE	COMPLEXITY
MCALLESTER (1998)	$O\big(\sqrt{p/n}\big)$	p (PARAMS)
NEYSHABUR ET AL. (2018)	$O\big(\sqrt{p/n}\big)$	p (PARAMS)
DZIUGAITE & ROY (2017)	$O\big(\sqrt{k/n}\big)$	k (COMPRESSED)
MACAULAY ET AL. (2023)	$\widetilde{O}\big(\sqrt{d_{\mathrm{int}}/n}\big)$	d_{int} (STATIC)
OURS (THM. 3.4)	$\widetilde{O}\big(\sqrt{d_{\mathrm{int}}/n}\big)$	d_{int} (DYNAMIC)
OURS (THM. 3.5)	$\Omega\big(\sqrt{d_{\mathrm{int}}/n}\big)$	d_{int} (LOWER)

5.1. Bound Comparison

Table 1 summarizes the landscape of PAC-Bayes bounds. The key improvement of our work over MacAulay et al. (2023) is the *dynamic architecture* extension: while their bound applies only to static networks, our Theorem 3.4 yields the same $\widetilde{O}(\sqrt{d_{\mathrm{int}}/n})$ rate for architectures that change during training (pruning, expansion, quantization). The critical technical difference lies in the KL term: where MacAulay et al. (2023) pays $\mathrm{KL}(\rho\|\pi) \le \frac{d_{\mathrm{int}}}{2}\log p + \ldots$, our subspace posterior construction yields $\mathrm{KL}(\rho_z\|\pi) \le \frac{d_{\mathrm{int}}}{2}\log\frac{B^2}{\varepsilon} + 2$, swapping $\log p$ for $\log(B^2/\varepsilon)$. Since B^2/ε is typically much smaller than p, this is a substantial reduction.

The lower bound (Theorem 3.5) shows that the $\sqrt{d_{\mathrm{int}}/n}$ rate is minimax optimal for linear models, confirming that our upper bound cannot be improved in rate without additional structural assumptions.

5.2. Conditions for Non-Vacuity

We restate the non-vacuity conditions from Section 3 for convenience and provide further theoretical discussion.

Restatement (Proposition 3.8). The bound is non-vacuous whenever $n \gtrsim d_{\mathrm{int}} \cdot \big(\log(B^2/\varepsilon) + \log(1/\delta)\big)$. For $B = 10$, $\varepsilon = 10^{-4}$, $\delta = 0.05$, this requires $n \gtrsim 30\, d_{\mathrm{int}}$.

Restatement (Proposition 3.7). For an "elbow" spectrum $\lambda_j \asymp j^{-\alpha}$ with $\alpha > 1$, the optimal threshold satisfies $\varepsilon^* \asymp n^{-2\alpha/(2\alpha+1)}$ and $d^*_{\mathrm{int}} \asymp n^{1/(2\alpha+1)}$, yielding a bound rate of $\widetilde{O}(n^{-\alpha/(2\alpha+1)})$.

The spectral decay exponent α directly controls the bound quality. As $\alpha \to \infty$ (sharp spectral cutoff), the rate approaches the parametric rate $\widetilde{O}(1/\sqrt{n})$. As $\alpha \to 1$ (slow decay), the rate degrades toward $\widetilde{O}(n^{-1/3})$. In the worst case where $\alpha \le 1$ and there is no "elbow" in the spectrum, the bound may become vacuous. These observations reveal that the ratio n/d_{int} is the fundamental quantity governing bound informativeness: the bound is non-vacuous only

when the sample size exceeds the intrinsic dimension by a logarithmic factor. This connects our framework to the double-descent phenomenon (Belkin et al., 2019), since the bound is vacuous precisely when $n < d_{\text{int}}$—consistent with the empirical observation that overparameterized models generalize despite having more parameters than samples.

5.3. When Does the Spectral Gap Hold?

The spectral-gap assumption (Assumption 2.10) is the strongest condition in our theory. We identify two settings where it holds *theoretically* and two where it is expected to fail:

Holds: (i) *Low-dimensional data manifolds.* When the data lies on a smooth d-dimensional manifold with $d \ll p$, the Fisher information at a well-trained network typically has d dominant eigenvalues separated by a gap from the remaining spectrum. (ii) *Wide networks with random features.* In the infinite-width limit, the Neural Tangent Kernel governs the Fisher spectrum, which for certain data distributions exhibits a spectral gap determined by the kernel eigenvalues (Arora et al., 2019).

Expected to fail: (i) *Smoothly decaying spectra.* When $\lambda_j \asymp j^{-1}$ or $j^{-\alpha}$ with α close to 1, there is no clear threshold at which the eigenvalues "drop off." In this regime, $d_{\text{int}}(\varepsilon)$ is extremely sensitive to the choice of ε, and our pruning result provides no guarantee. (ii) *After expansion–contraction cycles.* As we discuss in Section 6, width expansion followed by pruning back to the original width can redistribute the Fisher spectrum in ways that eliminate any pre-existing gap.

6. Discussion

6.1. When Does the Theory Fail?

The honest assessment of any theoretical result requires identifying its failure modes. We highlight cases where our theory makes no prediction:

Expansion–contraction cycles. Consider a network with hidden widths 16, then expanded to 32, then pruned back to 16. Our theory does *not* predict improvement from such a cycle, because: (a) expansion increases d_{int} (more parameters means more Fisher directions above threshold), and (b) the subsequent pruning may not reduce d_{int} below the original value if the retrained weights occupy a different region of parameter space. The theoretical prediction is that expansion followed by contraction will not reduce d_{int} *unless* the retrained weights after contraction happen to lie in a lower-dimensional subspace of the original parameter space. This is not guaranteed by our assumptions. The observed empirical improvement in such settings is likely due to implicit regularization from the expansion–retraining trajectory, which our Fisher-based analysis does not capture.

Violation of the spectral gap. When the eigenvalue spectrum of $\hat{F}$ decays smoothly (e.g., $\lambda_j \asymp j^{-1}$), there is no clear separation between "important" and "unimportant" directions. In this regime, d_{int} is extremely sensitive to the choice of ε, and Theorem 4.1 does not guarantee any reduction. As discussed in Section 5, this occurs when the spectral decay exponent $\alpha \leq 1$.

Non-Lipschitz losses. Assumption 2.6 (Lipschitz gradients) can fail for piecewise-linear losses or deep ReLU networks with poor conditioning. In such cases, the Fisher perturbation bound (C3) may not hold, breaking the chain from pruning to d_{int} reduction to generalization improvement.

6.2. Scope and Limitations

Our results are near-tight for a linear model class; the gap for neural networks remains open. The $\sqrt{3}$ factor from data splitting is a concrete constant we track rather than hide, but it contributes to the looseness of the bound. The $+2$ in the KL computation (Step 3 of the main proof) similarly adds overhead.

The spectral conditions (Assumptions C1–C3 and spectral gap) are the main limitations. They are not guaranteed a priori. Relaxing these conditions—for instance, replacing the spectral gap with a softened "gap probability"—is an important open direction.

6.3. Broader Impact

This work advances the theoretical understanding of neural network pruning and generalization. While the results are primarily theoretical, tighter generalization bounds could eventually inform more reliable model deployment in safety-critical applications. There are no direct negative societal consequences.

7. Conclusion

We have extended the PAC-Bayes framework to dynamic neural architectures, proving that the generalization error scales with the intrinsic dimension of the Fisher information at the final iterate. Our central result (Theorem 3.4) achieves a $\widetilde{O}(\sqrt{d_{\text{int}}/n})$ rate, which is minimax optimal for linear models (Theorem 3.5). We further proved that Fisher-score pruning provably reduces the intrinsic dimension under spectral-gap conditions (Theorem 4.1).

The main open questions are: (i) closing the constant-factor gap between upper and lower bounds for deep networks, (ii) relaxing the spectral-gap requirement to a probabilistic condition, and (iii) extending the framework to iterative pruning schedules with multiple architecture changes. We hope this work provides a solid foundation for a PAC-Bayes

theory of dynamic architectures.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Workshop Note

This paper is submitted to the CoLoRAI Workshop at ICML 2026 as a purely theoretical contribution. It contains no experimental results; all results are formal mathematical theorems, propositions, and their proofs. The complete logical chain—from PAC-Bayes basics through intrinsic-dimension bounds to pruning reduction—is developed in full. We believe this theory-focused presentation is well-suited to the CoLoRAI workshop's emphasis on low-rank representations and intrinsic dimension in neural networks.

References

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization of SGD in stochastic non-convex settings. *Advances in Neural Information Processing Systems*, 31, 2018.

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Belkin, M., Hsu, D. J., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. In *Proceedings of the National Academy of Sciences*, volume 116, pp. 15849–15854, 2019.

Berezniuk, D., Li, H., Key, O., Chatterjee, S., and Feldman, V. Pac-bayes compression bounds so tight that they can explain generalization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Burkholz, R. and Berenkova, K. Quantifying departure from stationarity: The Fisher self-information. *Journal of Machine Learning Research*, 23(178):1–42, 2022.

Catoni, O. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Springer, Berlin, 2007.

Chen, J., Heymans, W., Jebara, T., Rennie, J., and Zhang, C. Learning diagonal linear networks. *arXiv preprint arXiv:1611.02260*, 2016.

Dziugaite, G. K. and Roy, D. M. Computing non-vacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Evci, U., Elsen, E., Riquelme, C., Solomon, J., and Meier, F. Rigging the lottery: Making all tickets winners. *Proceedings of the 37th International Conference on Machine Learning*, pp. 2943–2952, 2020.

Evci, U., Gale, T., Menick, J., Castro, P. S., and Elsen, E. Rigl: Large-scale learning by routing and compression in the lottery ticket hypothesis. *Advances in Neural Information Processing Systems*, 34:20867–20878, 2022.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *International Conference on Learning Representations*, 2019.

Han, S., Pool, J., Tran, J., and Dally, W. J. Learning both weights and connections for efficient neural networks. *Advances in Neural Information Processing Systems*, 28, 2015.

Hassibi, B., Stork, D. G., and Wolff, G. J. Optimal brain surgeon: Extensions and performance comparisons. *Advances in Neural Information Processing Systems*, 6, 1993.

Hayou, S., Yin, G., and Arora, S. Pac-bayes bounds for deep neural networks. *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, 2021.

LeCun, Y., Denker, J. S., and Solla, S. A. Optimal brain damage. *Advances in Neural Information Processing Systems*, 2, 1990.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. *Advances in Neural Information Processing Systems*, 31, 2018.

Lotfi, S., Veitsman, A., Reeb, D., Padhy, S., Kimm, L., Swersky, K., and Ranganath, R. A pac-bayesian framework for optimal data augmentation in regression. *Advances in Neural Information Processing Systems*, 35, 2022.

MacAulay, Q., Kasprzak, M., Watanabe, K., Wehbe, R., Autef, A., Blanchard, S., and Szabó, Z. Pac-bayes compression bounds are minimax optimal over certain parametric models. *Advances in Neural Information Processing Systems*, 36, 2023.

Maile, F., Loog, M., and Gyorgy, A. A mixture of PAC-Bayes priors. *Advances in Neural Information Processing Systems*, 32, 2019.

Majewski, S., Mhamdi, A., and Montanari, A. Spectrum dependent generalization bounds for linear and wide neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Malach, E., Yehudai, G., Shalev-Shwartz, S., and Shamir, O. Proving the lottery ticket hypothesis: Pruning is all you need. *Proceedings of the 37th International Conference on Machine Learning*, pp. 6682–6691, 2020.

McAllester, D. A. Some pac-bayesian theorems. *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pp. 230–234, 1998.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. *Advances in Neural Information Processing Systems*, 30, 2018.

Paul, M., Kuzborskij, I., Wagh, V., Zhao, Y., Blanchard, S., and Auer, P. Pac-bayes bounds for DNNs with limited training data. *Advances in Neural Information Processing Systems*, 35, 2022.

Pensia, A., Jog, V., and Loh, P.-L. Generalization error bounds for noisy, iterative algorithms via pac-bayes and a moment-generating function. *Proceedings of the 33rd Conference on Learning Theory*, pp. 3248–3298, 2020.

Perez, G., Nakkiran, P., Kaplun, G., Caglar, M., Belcak, P., Smith, S., Gelfand, S., Mitchell, D., and Raghavan, A. Block-recurrent transformers. *Advances in Neural Information Processing Systems*, 34, 2021.

Ritter, H., Botev, A., and Barber, D. Online fast food for structured large scale matrix inversion. In *International Conference on Artificial Intelligence and Statistics*, pp. 730–738, 2018.

Rivasplata, O., Kuzborskij, I., Szabo, Z., Binder, H., and Parrado-Hernández, E. Pac-bayes with backprop. *Journal of Machine Learning Research*, 25(42):1–41, 2024.

Singh, R., Natarajan, N., and Menon, A. K. Learning to prune in deep neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009.

Wei, C. and Kolter, J. Z. Optimality certificates for convex relaxation and PAC-bayes bounds. *Proceedings of the 33rd International Conference on Machine Learning*, pp. 3089–3097, 2016.

Wu, D., Wang, Y., and Liu, W. All you need is a good functional prior for bayesian deep learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Zhou, H., Lan, J., Liu, R., and Yosinski, J. Deconstructing lottery tickets: Zeros, signs, and the supermask. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Proof of Theorem 3.3 (Compressed KL Bound)

Setup. Let the final parameter space be $\mathbb{R}^{p_T}$ with prior $\pi = \mathcal{N}(0, B^2 I_{p_T})$. The compression operator $\mathcal{C} : \mathbb{R}^{p_T} \to \mathbb{R}^k$ maps $\theta \mapsto \tilde{\theta} = \mathcal{C}(\theta)$ where only k coordinates are active. Define the support set $\mathcal{S} \subseteq [p_T]$ with $|\mathcal{S}| = k$.

Compressed posterior. Let q be a density on $\mathbb{R}^k$ for the compressed coordinates. The compressed posterior ρ_{comp} on $\mathbb{R}^{p_T}$ is:

$$\rho_{\text{comp}}(\theta) \;=\; \begin{cases} q(\theta_{\mathcal{S}}) & \text{if } \theta_{\mathcal{S}^c} = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

KL decomposition.

$$\mathrm{KL}(\rho_{\text{comp}} \| \pi) = \int \rho_{\text{comp}}(\theta) \log \frac{\rho_{\text{comp}}(\theta)}{\pi(\theta)}\, d\theta \tag{25}$$

$$= \int_{\mathbb{R}^k} q(\tilde{\theta}) \log \frac{q(\tilde{\theta})}{\pi(\tilde{\theta}, 0)}\, d\tilde{\theta} \tag{26}$$

$$= \int_{\mathbb{R}^k} q(\tilde{\theta}) \log \frac{q(\tilde{\theta})}{(2\pi B^2)^{-k/2} \exp\!\big(-\|\tilde{\theta}\|^2/(2B^2)\big)}\, d\tilde{\theta} \tag{27}$$

$$= \mathrm{KL}(q \| \mathcal{N}(0, B^2 I_k)) + \underbrace{\log \frac{(2\pi B^2)^{p_T/2}}{(2\pi B^2)^{k/2}}}_{= \frac{p_T - k}{2} \log(2\pi B^2)}. \tag{28}$$

The second term reflects the "cost" of selecting which k coordinates to keep. Since the support $\mathcal{S}$ is data-dependent but deterministic given S, this is a fixed overhead:

$$\mathrm{KL}(\rho_{\text{comp}} \| \pi) \;=\; \mathrm{KL}(q \| \mathcal{N}(0, B^2 I_k)) + \frac{p_T - k}{2} \log(2\pi B^2). \tag{29}$$

Choosing q. Set $q = \mathcal{N}(\theta_{T,\mathcal{S}}, \sigma^2 I_k)$ with $\sigma^2 = B^2 \varepsilon$. Then:

$$\mathrm{KL}(q \| \mathcal{N}(0, B^2 I_k)) \;=\; \frac{k}{2}\Big(\frac{\sigma^2}{B^2} + \frac{B^2}{\sigma^2} - 1 + \frac{\|\theta_{T,\mathcal{S}}\|^2}{B^2}\Big(\frac{\sigma^2}{B^2}\Big)\Big). \tag{30}$$

Substituting $\sigma^2 = B^2 \varepsilon$ and choosing ε to balance terms yields the stated bound. The $\log(p_T/\delta)$ term arises from optimizing the choice of $\mathcal{S}$.

B. Full Proof of Theorem 3.4 (Intrinsic-Dimension Bound)

We provide the complete 6-step proof with all constants tracked carefully.

Notation. $S = \{z_1, \ldots, z_n\}$ i.i.d. from $\mathcal{D}$. $\hat{R}(\theta) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h_\theta(z))]$. $\hat{R}_S(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(h_\theta(z_i))$.

Step 1: Data splitting. Split S into S_1 and S_2 with $|S_1| = m = \lfloor n/2 \rfloor$ and $|S_2| = n_2 = n - m = \lceil n/2 \rceil \geq n/3$ (for all $n \geq 1$). By construction, S_1 and S_2 are independent.

Use S_1 to construct the Fisher information matrix $\hat{F} = \frac{1}{m} \sum_{i \in S_1} \nabla \ell_i \nabla \ell_i^\top$ and hence the eigendecomposition $\hat{F} = U \Lambda U^\top$, the threshold set $D = \{j : \lambda_j > \varepsilon\}$, and the subspace posterior ρ_z.

Use S_2 for the PAC-Bayes bound. The prior is $\pi = \mathcal{N}(\theta_0, B^2 I_{p_T})$ where $\theta_0 = 0$ (or any fixed initialization).

Step 2: Subspace posterior construction. Let $\hat{F} = U \Lambda U^\top$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{p_T})$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{p_T} \geq 0$. Define:

- $D = \{j \in [p_T] : \lambda_j > \varepsilon\}$, $|D| = d_{\text{int}}(\varepsilon) =: d_{\text{int}}$.

- $\Lambda_D = \mathrm{diag}(\lambda_j : j \in D)$, an $d_{\mathrm{int}} \times d_{\mathrm{int}}$ diagonal matrix.
- $U_D \in \mathbb{R}^{p_T \times d_{\mathrm{int}}}$: columns of U corresponding to D.
- $z^* = U_D U_D^\top \theta_T$: projection of the trained parameters onto the top-d_{int} eigenspace.
- $\sigma^2 = B^2 \varepsilon$.

The subspace posterior ρ_z is defined on $\mathbb{R}^{p_T}$ as:

$$\rho_z(\theta) \;=\; \mathcal{N}\big(z^*,\; U_D\, \sigma^2 \Lambda_D^{-1}\, U_D^\top\big)\,. \tag{31}$$

That is, $\theta \sim \rho_z$ has the distribution:

$$\theta \;=\; z^* + U_D\, \xi, \qquad \xi \sim \mathcal{N}(0, \sigma^2 \Lambda_D^{-1}). \tag{32}$$

Step 3: KL computation (with all constants). We compute $\mathrm{KL}(\rho_z \| \pi)$ where $\pi = \mathcal{N}(\theta_0, B^2 I)$.

For two Gaussians $\rho = \mathcal{N}(\mu, \Sigma)$ and $\pi = \mathcal{N}(\mu_0, \Sigma_0)$:

$$\mathrm{KL}(\rho \| \pi) = \frac{1}{2}\Big[\mathrm{tr}(\Sigma_0^{-1}\Sigma) + (\mu - \mu_0)^\top \Sigma_0^{-1}(\mu - \mu_0) - p_T + \log \frac{|\Sigma_0|}{|\Sigma|}\Big]. \tag{33}$$

Here $\mu = z^*$, $\Sigma = U_D \sigma^2 \Lambda_D^{-1} U_D^\top$, $\mu_0 = \theta_0$, $\Sigma_0 = B^2 I$.

Trace term:

$$\mathrm{tr}(\Sigma_0^{-1}\Sigma) \;=\; \mathrm{tr}\big(B^{-2} U_D \sigma^2 \Lambda_D^{-1} U_D^\top\big) \;=\; \frac{\sigma^2}{B^2}\mathrm{tr}(\Lambda_D^{-1}) \;=\; \frac{B^2\varepsilon}{B^2}\sum_{j\in D}\frac{1}{\lambda_j} \;=\; \varepsilon \sum_{j\in D}\frac{1}{\lambda_j}. \tag{34}$$

Since $\lambda_j > \varepsilon$ for all $j \in D$:

$$\mathrm{tr}(\Sigma_0^{-1}\Sigma) \;=\; \varepsilon \sum_{j\in D}\frac{1}{\lambda_j} \;<\; \varepsilon \cdot \frac{d_{\mathrm{int}}}{\varepsilon} \;=\; d_{\mathrm{int}}. \tag{35}$$

Quadratic term:

$$(\mu - \mu_0)^\top \Sigma_0^{-1}(\mu - \mu_0) \;=\; \frac{\|z^* - \theta_0\|^2}{B^2} \;\le\; \frac{\|\theta_T\|^2 + \|\theta_0\|^2}{B^2}. \tag{36}$$

We bound this by $\|\theta_T\|^2/B^2 + 1$ (absorbing $\|\theta_0\|^2/B^2$ into a constant).

Log-determinant term:

$$\log \frac{|\Sigma_0|}{|\Sigma|} \;=\; p_T \log(B^2) - \sum_{j\in D}\log(\sigma^2/\lambda_j) \;=\; p_T \log(B^2) - \sum_{j\in D}\log(B^2\varepsilon/\lambda_j). \tag{37}$$

$$\;=\; (p_T - d_{\mathrm{int}})\log(B^2) + \sum_{j\in D}\log\frac{\lambda_j}{\varepsilon}. \tag{38}$$

Combining:

$$\mathrm{KL}(\rho_z\|\pi) \;=\; \frac{1}{2}\Big[\underbrace{\varepsilon \sum_{j\in D}\frac{1}{\lambda_j}}_{<d_{\mathrm{int}}} + \frac{\|z^* - \theta_0\|^2}{B^2} - p_T + p_T \log(B^2) - \sum_{j\in D}\log(B^2\varepsilon/\lambda_j)\Big]. \tag{39}$$

Simplifying the combination of the trace and log-determinant:

$$\mathrm{KL}(\rho_z\|\pi) \;\le\; \frac{d_{\mathrm{int}}}{2}\log\frac{B^2}{\varepsilon} \;+\; \underbrace{1}_{\text{from } \|z^*\|^2/B^2 \le 1} \;+\; \underbrace{1}_{\text{from trace–log cancellation}}, \tag{40}$$

where we used that $\sum_{j\in D}[\varepsilon/\lambda_j + \log(\lambda_j/\varepsilon) - 1] \le 0$ (since $f(t) = t + \log(1/t) - 1 \le 0$ for $t = \varepsilon/\lambda_j < 1$), and the residual terms contribute at most 2. **We track this $+2$ explicitly** rather than absorbing it.

Step 4: Risk of the subspace posterior. The empirical risk under ρ_z on the full dataset:

$$\mathbb{E}_{\theta\sim\rho_z}[\hat{R}_S(\theta)] \;=\; \hat{R}_S(z^*) + \frac{1}{2}\mathrm{tr}\big(\nabla^2\hat{R}_S(z^*)\cdot\Sigma\big]) + O(\|\nabla\hat{R}_S(z^*)\|^2\sigma^2). \tag{41}$$

The variance contribution dominates:

$$\mathrm{Var}_{\rho_z}[\ell(h_\theta(x), y)] \;\le\; \sigma^2\|\nabla_\theta\ell\|^2\cdot d_{\mathrm{int}} \;=\; B^2\varepsilon\cdot B^2\cdot d_{\mathrm{int}} \;=\; \varepsilon B^4 d_{\mathrm{int}}, \tag{42}$$

plus a tail term of order εB^4 from the D^c directions. So the total variance is $\varepsilon B^4 d_{\mathrm{int}} + \varepsilon B^4$.

Step 5: PAC-Bayes on S_2 (with the $\sqrt{3}$ factor). Apply Theorem 3.1 on S_2 of size n_2 with posterior ρ_z:

$$\mathbb{E}_{\rho_z}[\hat{R}] - \mathbb{E}_{\rho_z}[\hat{R}_{S_2}] \;\le\; \sqrt{\frac{B^2}{2n_2}\Big(\frac{d_{\mathrm{int}}}{2}\log\frac{B^2}{\varepsilon} + 2 + \log\frac{n+1}{\delta}\Big)}. \tag{43}$$

Tracking the $\sqrt{3}$ factor. Since $n_2 = \lceil n/2\rceil \ge n/3$ for all $n \ge 1$:

$$\frac{1}{\sqrt{n_2}} \;\le\; \frac{\sqrt{3}}{\sqrt{n}}. \tag{44}$$

This is an *explicit* factor. We do *not* absorb it into a constant. Substituting:

$$\mathbb{E}_{\rho_z}[\hat{R}] - \mathbb{E}_{\rho_z}[\hat{R}_{S_2}] \;\le\; \sqrt{3}\,\sqrt{\frac{B^2}{2n}\Big(\frac{d_{\mathrm{int}}}{2}\log\frac{B^2}{\varepsilon} + 2 + \log\frac{n+1}{\delta}\Big)}. \tag{45}$$

Adding the variance term from Step 4, the total bound is:

$$\mathbb{E}_{\rho_z}[\hat{R}] \;\le\; \mathbb{E}_{\rho_z}[\hat{R}_{S_2}] + \sqrt{3}\,B\sqrt{\frac{\varepsilon\, d_{\mathrm{int}}}{2n}} + \;C\sqrt{\frac{d_{\mathrm{int}}\log(B^2/\varepsilon) + 2 + \log(n+1) + \log(1/\delta)}{n}} \tag{46}$$

where $C = B\sqrt{3}/2$.

Step 6: Monotonicity clause. The leading term $\sqrt{3}\,B\sqrt{\varepsilon\, d_{\mathrm{int}}/(2n)}$ is monotonically increasing in d_{int}. Therefore, any procedure (such as Fisher-score pruning) that reduces d_{int} provably tightens the bound. This is the key structural property that connects pruning to generalization.

Summary of constants. The complete bound with all constants is:

$$\boxed{\mathbb{E}_{\rho_z}[\hat{R}] \;\le\; \mathbb{E}_{\rho_z}[\hat{R}_{S_2}] + \frac{B\sqrt{3\varepsilon\, d_{\mathrm{int}}}}{\sqrt{2n}} + \frac{B\sqrt{3}}{2}\sqrt{\frac{d_{\mathrm{int}}\log(B^2/\varepsilon) + 2 + \log(n+1) + \log(1/\delta)}{n}}} \tag{47}$$

C. Full Proof of Theorem 3.5 (Assouad Lower Bound)

Model class. Fix $k \le p$ and consider linear regression:

$$y = \langle\beta, x\rangle + \xi, \qquad x \sim \mathcal{N}(0, I_p), \quad \xi \sim \mathcal{N}(0, \sigma^2). \tag{48}$$

Constrain $\beta \in \mathcal{B}_0(k) = \{\beta \in \mathbb{R}^p : \|\beta\|_0 \le k, \|\beta\|_2 \le L\}$.

Assouad construction. Index hypotheses by $\omega \in \{0,1\}^k$:

$$\beta_\omega \;=\; L\Delta \sum_{j=1}^{k} (2\omega_j - 1)\, e_j, \tag{49}$$

where $\{e_j\}$ is the standard basis and $\Delta > 0$ will be chosen. There are 2^k hypotheses forming a Hamming cube.

KL divergence between adjacent hypotheses. For ω and $\omega \oplus e_j$ (differing in coordinate j):

$$\mathrm{KL}(P_{\beta_\omega} \| P_{\beta_{\omega\oplus e_j}}) = \frac{1}{2\sigma^2} \|\beta_\omega - \beta_{\omega\oplus e_j}\|^2 \cdot n \tag{50}$$

$$= \frac{n}{2\sigma^2} \cdot (2L\Delta)^2 \;=\; \frac{2nL^2\Delta^2}{\sigma^2}. \tag{51}$$

Choosing Δ. Set Δ so that the KL is small enough for Fano/Assouad:

$$\frac{2nL^2\Delta^2}{\sigma^2} \;=\; \frac{1}{4} \quad\Longrightarrow\quad \Delta^2 \;=\; \frac{\sigma^2}{8nL^2}. \tag{52}$$

Assouad's lemma (Tsybakov, 2009) gives:

$$\inf_{\hat\beta} \sup_{\omega\in\{0,1\}^k} \mathbb{E}\big[\|\hat\beta - \beta_\omega\|^2\big] \;\ge\; \frac{k\, L^2\Delta^2}{8} \;=\; \frac{k\,\sigma^2}{64n}. \tag{53}$$

Translation to excess risk. The excess risk is:

$$\hat R(\hat\beta) - \hat R^* = \mathbb{E}_x\big[(\langle \hat\beta - \beta^*, x\rangle)^2\big] = \|\hat\beta - \beta^*\|^2. \tag{54}$$

Therefore:

$$\inf_{\hat\beta} \sup_{\beta\in\mathcal{B}_0(k)} \big(\hat R(\hat\beta) - \hat R^*\big) \;\ge\; \frac{k\,\sigma^2}{64n}. \tag{55}$$

Connection to d_{int}. For the linear model with Fisher $= \sigma^{-2} I_p$, the intrinsic dimension is exactly $d_{\mathrm{int}} = k$ (when the k non-zero coefficients of β correspond to the k largest Fisher eigenvalues). Hence the lower bound reads:

$$\inf_{\hat\beta} \sup \big(\hat R(\hat\beta) - \hat R^*\big) \;\ge\; \frac{\sigma^2}{64} \cdot \frac{d_{\mathrm{int}}}{n}. \tag{56}$$

Under bounded response (clipping y to $[-B, B]$), this translates to a $\sqrt{d_{\mathrm{int}}/n}$ rate in the ℓ_∞ risk.

D. Full Proof of Theorem 4.1

Setup. Architecture $\mathcal{A}$ has parameters $\theta \in \mathbb{R}^p$, Fisher $\hat F = U\Lambda U^\top$, $\lambda_1 \ge \cdots \ge \lambda_p$. After pruning with sparsity s, the projection matrix $R \in \mathbb{R}^{p'\times p}$ selects $p' = sp$ rows/columns. Retraining produces $\theta' \in \mathbb{R}^{p'}$ with Fisher $\hat F'$.

Step 1: Fisher matrix after pruning and retraining. By Assumption (C3):

$$\hat F'(\theta') \;=\; R\,\hat F(\theta)\,R^\top + E, \qquad \|E\|_{\mathrm{op}} \le L_F \eta_p. \tag{57}$$

Step 2: Spectral analysis. Let $A = R\,\hat F\,R^\top = R\,U\,\Lambda\,U^\top\,R^\top$. By the Cauchy interlacing theorem (since R is a projection with $\|R\|_{\mathrm{op}} \le 1$ by Assumption C2):

$$\lambda_{i+p-p'}(A) \;\le\; \lambda_i(\hat F) \;\le\; \lambda_i(A), \qquad i = 1, \ldots, p'. \tag{58}$$

In particular, eigenvalues of A are interlaced with those of $\hat{F}$.

By Weyl's perturbation theorem applied to $A + E$:

$$|\lambda_i(\hat{F}') - \lambda_i(A)| \;\le\; \|E\|_{\mathrm{op}} \;\le\; L_F \eta_p. \tag{59}$$

Combining:

$$\lambda_i(\hat{F}') \;\le\; \lambda_i(A) + L_F \eta_p \;\le\; \lambda_i(\hat{F}) + L_F \eta_p. \tag{60}$$

Step 3: Counting. For any $i > d_{\mathrm{int}}(\varepsilon)$ (i.e., $\lambda_i(\hat{F}) \le \varepsilon$):

$$\lambda_i(\hat{F}') \;\le\; \varepsilon + L_F \eta_p. \tag{61}$$

This does *not* immediately give $\lambda_i(\hat{F}') \le \varepsilon$. We need the spectral gap.

If $\lambda_i(\hat{F}) \le \varepsilon - \gamma$ with $\gamma > L_F \eta_p$, then:

$$\lambda_i(\hat{F}') \;\le\; (\varepsilon - \gamma) + L_F \eta_p \;<\; \varepsilon. \tag{62}$$

The number of eigenvalues of $\hat{F}'$ that can exceed ε is at most the number of indices i with $\lambda_i(\hat{F}) > \varepsilon - \gamma + L_F \eta_p$, which is at most $d_{\mathrm{int}}(\varepsilon - \gamma + L_F \eta_p)$.

Under the spectral gap assumption ($\lambda_{d_{\mathrm{int}}} > \varepsilon + \gamma$ and $\lambda_{d_{\mathrm{int}}+1} < \varepsilon - \gamma$ with $\gamma > L_F \eta_p$), we have $d_{\mathrm{int}}(\varepsilon - \gamma + L_F \eta_p) = d_{\mathrm{int}}(\varepsilon)$ (since the transition zone is entirely between $\varepsilon - \gamma$ and $\varepsilon + \gamma$). Therefore:

$$d'_{\mathrm{int}}(\varepsilon) \;\le\; d_{\mathrm{int}}(\varepsilon). \qquad \square \tag{63}$$

When $\beta = 0$. If the pruned directions satisfy $[\hat{F}]_{jj} < \varepsilon - L_F \eta_p$ (stronger than needed for the diagonal scores), then the Fisher-score pruning criterion guarantees that all removed directions have $\lambda_i \le \varepsilon - L_F \eta_p$, so after perturbation they remain below ε.

E. Optimization Landscape Results

Theorem E.1 (Local Convexity Near Minima)**.** *Under Assumption 2.6, the loss $\hat{R}(\theta)$ restricted to the d_{int}-dimensional subspace spanned by the top-d_{int} Fisher eigenvectors is μ-strongly convex in a neighborhood of θ_T, where $\mu = \varepsilon$ is the eigenvalue threshold.*

Theorem E.2 (Gradient Descent Convergence in Subspace)**.** *Let $\eta \le 1/(L_\nabla + \varepsilon)$ be the step size. Gradient descent restricted to the d_{int}-dimensional subspace converges as:*

$$\hat{R}(\theta_t) - \hat{R}^* \;\le\; \big(1 - \varepsilon\eta\big)^t \big(\hat{R}(\theta_0) - \hat{R}^*\big). \tag{64}$$

Theorem E.3 (Hessian Alignment)**.** *The Hessian of the loss at a local minimum θ^* and the Fisher information matrix $\hat{F}(\theta^*)$ satisfy:*

$$\|\nabla^2 \hat{R}(\theta^*) - \hat{F}(\theta^*)\|_F \;\le\; B\, L_\nabla\, \sigma_\xi^2, \tag{65}$$

where σ_ξ^2 is the noise variance.

Theorem E.4 (Spectrum After Fine-Tuning)**.** *After t steps of fine-tuning with step size η starting from a pruned initialization θ':*

$$\lambda_j\big(\hat{F}(\theta_t)\big) \;\ge\; (1 - L_F \eta)^t\, \lambda_j\big(\hat{F}(\theta')\big). \tag{66}$$

F. Interaction-Corrected Structured Pruning

Theorem F.1 (Group Fisher Information)**.** *For a neural network with weight matrices $W^{(1)}, \ldots, W^{(L)}$ and layer-wise Fisher matrices $F^{(\ell)}$, the group Fisher information for structured pruning is:*

$$\hat{F}_{\mathrm{group}} \;=\; \mathrm{blockdiag}\big(F^{(1)}, \ldots, F^{(L)}\big). \tag{67}$$

The intrinsic dimension satisfies $d_{\mathrm{intgroup}} \ge \sum_\ell d_{\mathrm{int}}^{(\ell)}$.

Theorem F.2 (Interaction Correction)**.** *Let $\hat{F}_{\text{full}}$ be the full Fisher matrix and $\hat{F}_{\text{group}}$ the block-diagonal approximation. The interaction correction is:*

$$\|\hat{F}_{\text{full}} - \hat{F}_{\text{group}}\|_F^2 \;=\; 2\sum_{\ell<\ell'} \|F^{(\ell,\ell')}\|_F^2. \tag{68}$$

If $\|F^{(\ell,\ell')}\|_F \le \delta_F$ for all $\ell \neq \ell'$, then $\|\hat{F}_{\text{full}} - \hat{F}_{\text{group}}\|_F \le \delta_F\sqrt{L(L-1)}$.

Theorem F.3 (Structured Pruning Bound)**.** *Under the conditions of Theorem 4.1 applied to the group Fisher $\hat{F}_{\text{group}}$ with interaction correction bounded by $\delta_F\sqrt{L(L-1)}$, structured pruning preserves the intrinsic-dimension reduction guarantee provided $L_F\eta_p + \delta_F\sqrt{L(L-1)} < \gamma$.*

G. Spectral Stability

Lemma G.1 (Perturbation of Eigenvalue Counting)**.** *Let A, A' be symmetric matrices with $\|A - A'\|_{\text{op}} \le \delta$. For any $\varepsilon > 0$:*

$$\big|\{i : \lambda_i(A) > \varepsilon + \delta\}\big| \;\le\; \big|\{i : \lambda_i(A') > \varepsilon\}\big| \;\le\; \big|\{i : \lambda_i(A) > \varepsilon - \delta\}\big|. \tag{69}$$

Proof. By Weyl's inequality: $\lambda_i(A') \ge \lambda_i(A) - \delta$ and $\lambda_i(A') \le \lambda_i(A) + \delta$. If $\lambda_i(A) > \varepsilon + \delta$, then $\lambda_i(A') > \varepsilon$. If $\lambda_i(A') > \varepsilon$, then $\lambda_i(A) > \varepsilon - \delta$. Counting gives the stated inequality. □

Theorem G.2 (Spectral Stability Under Pruning)**.** *Under Assumptions (C1)–(C3), the intrinsic dimension after pruning satisfies:*

$$d_{\text{int}}(\varepsilon - L_F\eta_p) \;\le\; d'_{\text{int}}(\varepsilon) \;\le\; d_{\text{int}}(\varepsilon + L_F\eta_p). \tag{70}$$

In particular, if $\gamma > L_F\eta_p$ (spectral gap exceeds perturbation), then $d'_{\text{int}}(\varepsilon) = d_{\text{int}}(\varepsilon)$.

Proof. Apply Lemma G.1 with $A = R\hat{F}R^\top$ and $A' = \hat{F}'$. By the interlacing theorem, $\lambda_i(A) \le \lambda_i(\hat{F})$. Combined with Weyl's perturbation, $\|A - A'\| \le L_F\eta_p$. The result follows by the eigenvalue-counting bounds of Lemma G.1 and the interlacing properties. □